Cairo University
Faculty of Graduate Studies for Statistical Research

The 54th Annual Conference on Statistics, Computer Sciences and Operations Research

COMPUTER SCIENCE

9-11 Dec. 2019

Index

Computer Sciences

A Review on Fruits Recognition through Different Techniques

Mohamed S. Rabie[1] Ahmed H. Asad[2] Hesham A. Hefny[3]

Abstract

Fruits recognition through machines is one of the most technologies that are needed in agriculture sector because the huge quantities of fruit productions that are produced every day need to promising contributions of machines instead of human to control the huge processes that occur on fruits like harvesting, packaging, sorting, grading, counting and etc. In this paper we detailed an overview of different approaches and algorithms that are used in fruits recognition through different studies that are relatedly, previously and currently with discussion of the steps they followed for building classifiers for fruits recognition. Some feature extraction methods are discussed with the common features of fruits like color, size, shape and texture. Concepts, definitions, process, advantages, disadvantages and challenges, occurring in fruits recognition, are discussed in this paper.

Keywords: Fruits recognition; Feature descriptors; Convolutional Neural Networks (CNN).

1. Introduction

Fruits are very important nutritional value for human, and there are huge processes applied on fruits every day, starting from planting to preparing fruits for human in different forms like fruit salad, fruit juice, ice fruits, fruit pieces and etc. To perfect management for fruits production processes, we need to promising systems to control all of these processes like planting, harvesting, packaging, sorting, grading, counting and etc. One of the most technologies that are followed to build these systems is artificial intelligent systems, behind each system different contributions are introduced through different studies. In this paper we focused on fruits recognition studies that introduced many different approaches and algorithms for fruits recognition. We displayed the steps that are followed to building a fruit classifier through different studies as shown in fig 1. In section 2, we started by an overview of challenges facing fruits recognition. In section 3 for the different methods that are followed for dataset collection and preparing for fruits, and we discussed pre-processing methods that are applied on dataset before data processing in the same section. In section 4 we reviewed features extraction methods like Scale Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP) with the common features of fruits like color, size shape and texture. Detailed discussion for the approaches and algorithms that were followed for fruits recognition through different related previous and current studies were presented in section 5 like Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

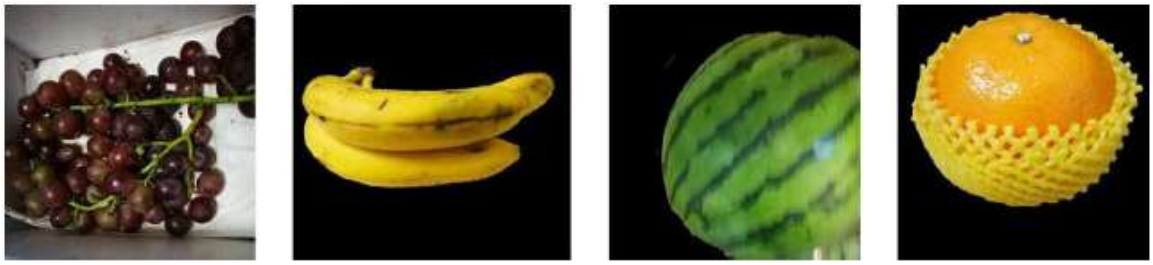
[1]Computer and Information Sciences Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt. Mohamedsayed_Rabie1987@outlook.com
[2]Computer and Information Sciences Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt. ah_assad@hotmail.com
[3]Computer and Information Sciences Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt. Hehefny@ieee.org

In section 6, summary of results were reviewed between machine learning algorithms that were introduced by researchers for fruits classification, and the influences of feature types on results. Finally we concluded this review in section 7.

Fig 1. The steps that are followed to building a fruit classifier.

2. Challenges

Processing on fruits through Artificial Intelligent (AI) has a different challenges, the top of these challenges represents in the big similarity between different kinds of fruits in different features like color, shape, size. Example: orange and grapefruit, apple and tomato, yellow apple and guava, apricot and lemon and etc. Also there are different challenges in fruits on trees and orchard because the imperfect form of fruit shapes that is influenced by different variables in weather and etc. In Yu-Dong Zhang et al. (2017) there are other influences that may occur on fruit image based on the environment of work like (a) background, (b) decay, (c) unfocused and (d) occlusion as shown in fig 2. Also there are little researches that discussed many different problems in fruits, especially in imperfect fruit images that are represented in expired fruits, fruits on trees, fruits inside containers and etc.



(a) background **(b) decay** **(c) Unfocused** **(d) occlusion**

Fig 2. Example of imperfect fruit images Yu-Dong Zhang et al. (2017)

3. Materials

The first work to create classifier in artificial intelligence is dataset collection. Dataset must be collected according to the environment that the system will interact. For example if the system will be installed in supermarket to know the different categories of a fruit element to determine their prices, and the fruits will be putted on white table all times, in

this case we don't need to interest by background during dataset collection, because there is no challenges that may be occurred during fruit classification because the system will process fruits under white background all times. But if we classify fruits on trees, in this case we will face some challenges like background and imperfect fruit images because different conditions that may be occurred on trees, here we can't train our classifier by collecting dataset for fruits on containers or inside boxes or in fruit labs etc. but we must collect dataset for fruits on trees, because this is the environment that the system will interact, and this doesn't prevent us to collect a set of perfect or direct images of fruit object to determine its characteristics clearly to classifier and after that we must collect another set of images of fruit object with the challenges that may be occurred on trees. After dataset collection, many researchers go to pre-processing for the collected dataset by applying different techniques on dataset object for enhancement the performance of classifier.

3.1 Dataset Collection

There are different methods that are followed by researchers for dataset collection through different studies. Some researchers prefer using a camera for dataset collection. According to Horea and Mihai (2018), the researches obtained fruits dataset by filming the fruits while they are rotated by a motor and then extracting frames. Behind the fruits they placed a white sheet of paper as background. Other researchers prefer internet by using specialist websites for dataset collection. In Morteza Akbari et al. (2016), the researchers downloaded fruits dataset from ImageNet through internet. ImageNet is a large visual dataset, and it consists of more than 14 million images and 20,000 categories. In ImageNet the images are real world images generated by ordinary users of the internet. Other researches use more than one method for dataset collection. In Yu-Dong Zhang et al. (2017), the researchers proposed a system for fruit category classification, and they obtained fruits dataset through three parts: (i) Via digital camera (ii) download from http://images.google.com; (iii) download from http://images.baidu.com. They obtained a 3600-image dataset with 200 images for each fruit type. Other researchers prefer to use dataset that were created previously by another researchers especially in the researches that are interested to creating implementations like Shadman sakib et al. (2019), where all the dataset images were selected from the fruits-360 dataset that is obtained in Horea and Mihai (2018). In Table 1 we referred to some of the mostly used fruit datasets that were used in different studies for fruit recognition.

3.2 Pre-Processing

Pre-processing procedures are the most important contributions to enhancement the performance of classifier and reduce training and classification time. After dataset collection, the images are influenced by background clutter, illumination variations and sensor capturing artifacts, specular reflections, shading and shadows. Also, the image may be needs to resize to reduce training and classification time. Sometimes, the researchers need to change the color mode, view point and rotation of image to gain faster classifier. All of these processes are applied in pre-processing step. In Yu-Dong Zhang et al. (2017), the researchers proposed a system for fruit category classification.

They used a four step preprocessing techniques. First, moving the fruit object into the center of the image.

Dataset Name	Author(s)	# of Images	# of Categories	Website
Fruits-360.	Horea Muresan, Mihai Oltean.	82213	120 Fruits and Vegetables	On GitHub: https://github.com/Horea94/Fruit-Images-Dataset On Kaggle: https://www.kaggle.com/moltean/fruits
ImageNet	Research team	188000	309 Fruits	http://www.image-net.org/
Fruit Recognition Dataset	Israr Hussain, Qianhua He, Zhuliang Chen, Wei Xie	44406	-	https://zenodo.org/record/1310165#.XcIsH-YzbIU
FIDS30	Marco skrjanec	971	30 Fruits	https://www.vicos.si/Downloads/FIDS30
ACFR Orchard Fruit Dataset	Agriculture team at the Australian Centre for Field Robotics	Apple: 1120 Mangoes: 1964 Almonds: 620	-	https://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/

Table 1. Some of the mostly used fruit datasets

Second, they cropped and resized image into 256x256 matrix. Third, they removed the background of image by split-and merge algorithm Dai-Ton H et al. (2016). Fourth, they labeled each image manually to one of the 18 fruit types, but there is one challenge may be occurred during resize image dataset that may lose some image characteristics especially if the image is resized into smaller size. In Horea and Mihai (2018), the researchers proposed a system for fruit recognition, and they wrote a dedicated algorithm which extracts the fruits from background and the fruits were scaled to fit 100x100 pixels image. In Morteza Akbari et al. (2016), the researchers proposed a system for counting fruits and vegetables calories. They resized all of the images in the database to the width and height of 256 pixels without preserving the aspect ratio. Then the resized images were cropped to the size of 227 × 227 with random offsets. According to Arivazhagan et al. (2010), the researchers extracted the region of interest on image by converting image

into HSV (Hue, Saturation and Value) color space. In Hossam M. Zawbaa et al. (2014), the researchers proposed a system for automatic fruits recognition for classifying and identifying fruit types. They proposed a model for resizing images to 90 x 90 pixels; in order to reduce their color index, and the acquired image is in RGB format which is a real color format for an image. But the main challenge in pre-processing step is the possibility of losing some crucial features of object so we must follow carefully the recommendations and instructions that are putted for algorithm which used for pre-processing.

4. Feature Extraction

Each object has a set of crucial features that determine its characteristics, and the all other features that are belong to the same object, and are not crucial shouldn't be taken, but only some of them can be taken to support classifier. According to Sapan Naik and Bankim Patel (2017), in computer vision, there are many different feature descriptors like Scale Invariant Feature Transform (SIFT) Tony Lindeberg (2012), Speeded Up Robust Features (SURF) Bay H et al. (2006), Histogram of Oriented Gradient (HOG) Patwary et al. (2015), Local Binary Pattern (LBP) di Huang et al. (2011). All of these descriptors are followed according to the type of features that will be decided by researchers in the study like color, shape and texture features, all of these features are decided according to the object characteristics, purpose of classification and environment that the classifier will interact. Sometimes the object is similar to other objects in color, shape and etc. and one type of features is not sufficient to classify object so we must go to additional feature type like color, shape, texture to support the classifier to recognize object.

4.1 Feature Descriptors

A feature descriptor is an algorithm which takes an image and outputs feature vectors where feature vector is just a vector that contains information describing an object's important characteristics and after that the descriptor encodes interesting information into a series of numbers. In Hossam M. Zawbaa et al. (2014), SIFT was used for fruit recognition based on shape and color features, where the image content is transformed into local feature coordinates that are invariant to translation, rotation, scale and other imaging parameters Sergio Cubero et al. (2011); Sapan Naik and Bankim Patel. (2014). SIFT works in four stages: Scale-space extreme detection, keypoint localization, orientation assignment and keypoint description. The first stage searches for stable features over multiple scales using a continuous function (Gaussian function) for scale. The second stage fits a model to determine location and scale, and select key points based on a measure of stability. The third stage computes the best orientation(s) for each keypoint region. The fourth stage uses local image gradient at selected scale and rotation to describe each keypoint region. Also, SURF is specific algorithm used for feature detector and feature descriptor. This algorithm has three main parts Alwanin (2014), feature detection, feature description and feature matching where feature detection is a set of methods that are employed at finding image features by making decisions at every point in the image based on approximate hessian matrix, and feature descriptor is an n-dimensional vector that represents an object based on some measurements on image

features, and feature matching is a set of methods based on some distance calculations. The HOG is defined in Sapan Naik and Bankim Patel (2017) as a feature descriptor used for object detection in computer vision, which counts frequencies of oriented gradients in localized part of an image. Local Binary Pattern (LBP) is referred in Sapan Naik and Bankim Patel (2017) which used for classification problem and it is a visual/texture descriptor, which works on texture/pattern in an image. Also, there are different researches that followed other methods for Describing features. In Arivazhagan et al. (2010), the researchers introduced a system for fruits recognition by using color and texture features. They obtained one channel containing the luminance information and two other channels containing chrominance information. They used HSV representation to compute the texture features from luminance channel 'v' and color features from chrominance channels 'H' and 'S' Drimbarean and Whelan (2001). In Morteza Akbari et al. (2016), the researchers proposed a system for counting the calorie of fruits and vegetables, and they used color, size, shape and texture features for fruits. They proposed a method uses state of the art deep-learning techniques for feature extraction and classification.

4.2 Color Features

The cluster of fruits appears in their different colors. So, the color features in fruits can contribute significantly in classification process. There are many different color descriptors like Color Histogram (CH) Jain and Vailaya (1996), Color Moments (CM) Flickner et al. (1995) and etc. Color Histogram is a representation of the distribution of colors in an image. In Chakravarti and Meng (2009), the researchers introduced a project that implements and tests a simple color histogram based search and retrieve algorithm for images. According to Dong ping Tian (2013), color features are defined subject to a particular color space or model. A number of color spaces have been used in literature, such as Red-Green-Blue (RGB), Hue-Saturation-Value (HSV) and Hue-Min-Max-Difference (HMMD). Once the color space is specified, color feature can be extracted from images or regions. Among the above mentioned color descriptors, the CM is one of the simplest yet very effective features. The common moments are mean, standard deviation and skewness. The corresponding calculations (1), (2), (3) can be defined as follows:

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij} \tag{1}$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}\left(f_{ij-\,\mu_i}\right)^2\right)^{\frac{1}{2}} \tag{2}$$

$$\gamma_i = \left(\frac{1}{N}\sum_{j=1}^{N}\left(f_{ij-\,\mu_i}\right)^3\right)^{\frac{1}{3}} \tag{3}$$

Where f_{ij} is the color value of the i-th color component of the j-th image pixel and N is the total number of pixels in the image. μ_i, σ_i, γ_i (i =1,2,3) denote the mean, standard deviation and skewness respectively of each channel of an image respectively. In Hossam M. Zawbaa et al. (2014), the researchers proposed a system for fruit recognition based on color and shape features. They described the images by using CM, so they used color mean, color variance, color skewness and color kurtosis. Also there are several models of colors like Hue-Intensity-Saturation (HIS), Hue-Saturation-Value (HSV), RGB and etc. These models are described in Sergio Cubero et al. (2011); Sapan Naik and Bankim Patel. (2014). In Arivazhagan et al. (2010), the researchers proposed a system for fruit recognition by using color and texture features. They extracted the region of interest on image by converting image into HSV colour space and the statistical features such as mean, standard deviation, skewness and kurtosis are derived from H and S components.

4.3 Size Features

One of the features that can be studied in fruits is fruit size feature which can be used for classify different characteristics in fruits like its kind, quality, expiration. In Baohua Zhang et al. (2014), the researchers introduced explanation for applications of size inspection that are one of the important parameter of fruit's quality measurement. In Sergio Cubero et al. (2011), the researchers proposed a system for automatic examination and quality assessment of fruits and vegetables. They described size and volume estimation methods. According to Sapan Naik and Bankim Patel (2017), there are different size measures like area, perimeter, weight, height, length and width are used in fruit size recognition.

4.4 Shape Features

Shape feature is one of the important features especially in fruits because some kinds of fruits have the same color of other kinds like strawberry and tomato, banana and lemon, orange and mango, yellow apple and guava, and etc. Here the color feature only is not sufficient to classify these kinds of fruits, and we must go to other additional features to make the object in unique form. In Hossam M. Zawbaa et al. (2014), the researchers used color and shape features to recognize fruits. They used Centroid, Eccentricity, and Euler Number features to describe the shape. The shape centroid (center of gravity) determines the image centroid position which is fixed in relation to the shape. Eccentricity measures the aspect ratio of the length of major axis to the length of minor axis. It is computed by minimum bounding rectangle method or principal axes method. The relation between the number of connecting parts and the number of holes on image shape is described by Euler number image describes. Euler number is computed by subtract the number of holes on a shape from the number of contiguous parts Mingqiang et al. (2008). According to Sapan Naik and Bankim Patel (2017), the objective of the shape description is characterize the shape in such a way that the values are very similar objects in the same form class or category, and quite different for objects in different categories.

4.5 Texture Features

According to Sapan Naik and Bankim Patel (2017), image texture is a set of two-dimensional array calculated to quantify visual texture of an image. It provides details about the spatial organization of color or intensities in an image (region of an image). Texture features achieved good results 95% accuracy in Fernandez et al. (2005), for classification of grade of apples after dehydration. According to Arivazhagan et al. (2010), texture features are useful information for quality evaluation of fruit and vegetables. The researchers proposed a system for fruit recognition by using color and texture features. They used HSV representation to compute the texture features from luminance channel 'v' and color features from chrominance channels 'H' and 'S' Drimbarean and Whelan (2001). LBP or HOG descriptors can be used for texture features extraction.

5. Machine Learning Algorithms

Machine Learning algorithms are developed as mathematical models that can be applied on sample of data. This data is divided into two parts, one for training classifiers and another for testing classifiers. The usage of machine learning algorithms is to build artificial intelligent systems that can classify objects and make decisions. In this section, we will review some machine learning algorithms that were applied for fruits classification through different studies like Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

5.1 Artificial Neural Networks (ANN)

The neural network itself is not an algorithm, but it is a framework for many different machine learning algorithms to work together and process complex data inputs. ANNs are trained computer programs which work similar to our brains. As shown in fig. 3, the layers of ANNs consist of one input layer, one or more hidden layer and one output layer Zhang 100-CH03 .

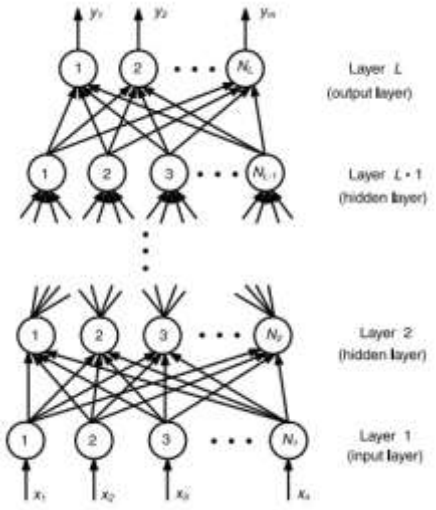


Fig 3. Artificial Neural Networks (ANN) layers Zhang 100-CH03 .

Nodes in ANNs are known as neurons. Each layer consists of set of neurons. Activation function is in hidden layer and output layer. Some of the activation functions are rectifier, sigmoid, hyperbolic tangent. The links contain weights connect nodes with each other. There are different studies classified fruits by using ANNs. In Giacomo Capizzi et al. (2015), Fruit defects are classified by using a new approach which uses Radial Basis Probabilistic Neural Networks (RBPNN) where the texture and gray of defect area are extracted by computing a gray level co-occurrence matrix and then defect areas are classified by the applied RBPNN solution. In this study Giacomo Capizzi et al. (2015), RBPNN consists of an input layer which represents the feature vectors and this layer is fully interconnected with the hidden layer, which consists of the training set (example vectors), and one other important element of the RBPNN is the output layer and the determination of the class for which the input layer fits. The equations that are used in this study Giacomo Capizzi et al. (2015) appear below:

$$x_j^{(1)} \propto exp(\frac{\|W^{(0)}.u\|}{2\sigma^2}) \tag{4}$$

$$x_j^{(1)} \triangleq \rho(\frac{\|u - W^0\|}{\beta}) \tag{5}$$

$$x_K^{(2)} = \sum_J W_{jk} x_j^{(1)} \tag{6}$$

In (4) u is the input pattern vector, $W^{(0)}$ is the weight vector and each pattern unit performs a nonlinear operation on the result, this nonlinear operation gives output $x^{(1)}$, and σ represents the statistical distribution spread. In (5) ρ is radial basis function and β is the distribution shape control parameter. In (6) k is the summation unit and W_{jk} represents the weight matrix.

In this study Giacomo Capizzi et al. (2015), the researchers used little dataset that is contained 400 images only for orange fruit surface include stab wounds, bruise, abrasion, sunburn, injury, hail damage. They resulted 389 out of 400 fruits are accurately classified. In José de Jesús Rubio. (2016), the researcher proposed a method for fruits and vegetables classification by using neural network with arduino microcontroller and some electronic sensors. This method was proven with the adaline, multilayer, and radial basis function neural networks. The three neural networks gave a system efficiency of 100%, but the adaline and multilayer neural networks were more convenient for the proposed classification strategy in the case of fruits or vegetables because they were more compact and more accurate. But in this study José de Jesús Rubio. (2016), the researcher didn't relay on ANNs alone but he went to another technology that is represented in arduino microcontroller and some electronic sensors, so we can't back this result for ANNs only. In Chowdhury et al. (2013), a proposed automatic recognition system (vegetable vision) to facilitate the process of a supermarket or grocery stores for the blind. This system consists of an integrated measurement and imaging technology with a user friendly

interface. The classification based methods for image segmentation; size, shape, color, and texture features for object measurement, statistical and neural network methods for classification. The results that were produced from this study were 96.55% accuracy. But, in this study Chowdhury et al. (2013), the researchers used color digital camera to gain more accurate details for object on image and this may influence on classifier speed. In Aibinu et al. (2011), a system for automatic identifying and sorting fruit was proposed. The researchers used ANNs, Fourier Descriptors (FD) and Spatial Domain Analysis (SDA) for the development of automatic fruits identification and sorting system, this approached is evaluated by using apple, banana and mango images. Fruits images were captured using digital camera inclined at different angles to the horizontal. The stand of fruit image acquisition system is shown in Fig. 4. The highest accuracy of this work achieved 99.1%. In ANNs, one of the different libraries that can be used as open source neural network library is Fast Artificial Neural Network (FANN) which is a free open source neural network library, which implements multilayer artificial neural networks in C with support for both fully connected and sparsely connected networks.

Fig 4. The stand of fruit image acquisition system Aibinu et al. (2011).

5.2 Convolutional Neural Networks (CNN)

A convolutional neural networks (CNNs) are one of the most widely used type of deep artificial neural networks that are used in various fields such as image and video recognition, speech processing as well as natural language processing Tasneem Gorach (2018). According to Morteza Akbari et al. (2016), Convolution Neural Network (CNN) is one of the best machine learning algorithms that produced promising results in image and video recognition, voice recognition, signal processing and natural language processing. According to Sapan Naik and Bankim Patel (2017), CNN mainly contain four steps namely Convolution, Pooling, Flattening and Full Connection as shown in fig. 5. In Convolution step, feature detector (Kernel/Filter) is applied on input image and Feature Map is created, next is ReLU layer comes in which our all Feature Maps are passed through Rectifier function. Rectefier function is an activation function defined as the positive part of its argument:

$$f(x) = x^{+} = \max(0, x) \tag{7}$$

Where x is the input to a neuron In Pooling step, all feature maps are given as input and pooled feature map is received as output. Pooling function can be any like min, max or median, next flattening step, we simply convert pooled feature map into single vector because it will next given as input to artificial neural network. Last step is full connection, which is nothing but implementing artificial neural network where all nodes

of input, hidden and output layers are connected with each other. In Horea and Mihai (2018), the researchers proposed a system for fruit recognition by creating four convolution layers and they specified filters for each convolution layer to scan image.

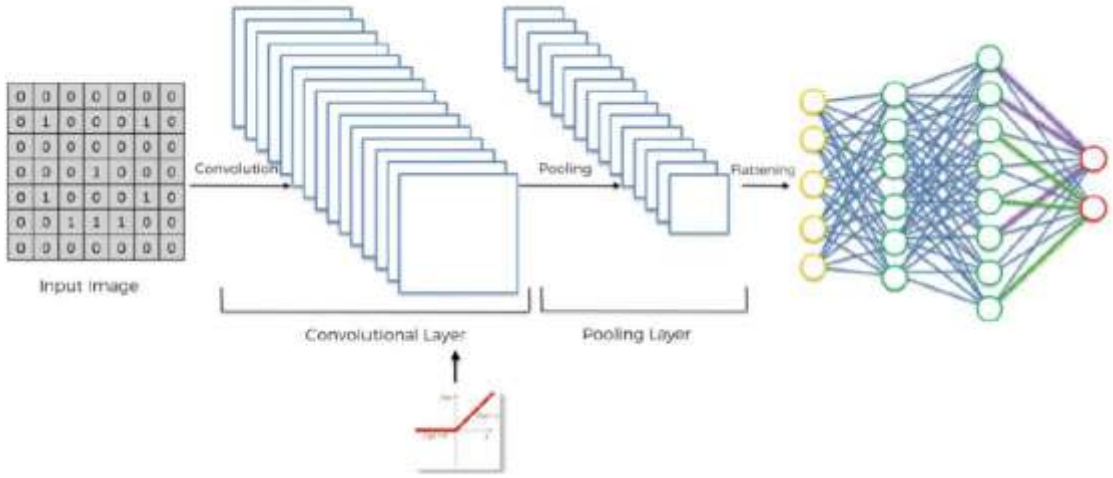


Fig 5. Convolutional Neural Network intuition Sapan Naik and Bankim Patel (2017).

Each filter was sized by 5 x 5. Next, a feature maps were created for each convolution layer, and all of these feature maps were passed through rectifier function. Each convolution layer had set of feature maps that were created by filter detector that was applied on each specified region on image. So, in each convolution layer, the image was split into set of regions, and each region was filtered by filter detector. After that, one feature map was created for each region, and rectifier function was applied for each feature map. The output result that was produced from rectifier function for each feature map was pooled by taking the maximum value or the average value for it. According to Yu-Dong Zhang et al. (2017), the max pooling gives better slight performance than average pooling. In Horea and Mihai (2018), after pooled all results, one pooled feature map was created, and this pooled feature map was simplified into one vector in flattening step, and the next convolution layer used these output that was produced from the previous layer as input for it and so on. Last step is full connection, where all nodes of input, hidden and output layers are connected with each other. In deep learning, there are different open source frameworks that can build Convolution Neural Network (CNN) in complete form like TensorFlow, Keras. In Horea and Mihai (2018), the researchers used TensorFlow for building four convolution layers, and they produced accuracy 96.3%. But in this study Horea and Mihai (2018), a perfect dataset was processed for fruits which make classifier can't classify fruits under more constraints and influences. TensorFlow is open source software that is used for different purposes. Some of these purposes are related to machine learning applications such as neural networks. TensorFlow was developed by Google brain team for internal Google use. In Morteza Akbari et al. (2016), Convolution Neural Network was created by three convolution layers, three pooling

layers and two fully connected layers, for fruits and vegetables calorie counter, with average accuracy 75%. But this result is not sufficient to relay on this classifier in different tasks that are related to calculate calories of fruits and vegetables. In Yu-Dong Zhang et al. (2017), convolutional neural network was created by four convolution layers, four pooling layers and two fully connected layers for fruit classification system. The researchers initialized set of filters for each convolution layer to scan image. In convolution layer 1, each filter was sized by 7x7, and 40 filters were created in this layer. In convolution layer 2, each filter was sized by 5x5, and 80 filters were created in this layer. In convolution layers 3, 4 the filters were sized by 3x3, and 120 filters were created for convolution layer 3, and 80 filters were created for convolution layer 4. In this research Yu-Dong Zhang et al. (2017), the researchers yielded an overall accuracy 94.94%. The equations that were used in this study Yu-Dong Zhang et al. (2017) were:

$$p(r|x) = \frac{p(x|r)p(r)}{\sum_{k=1}^{R} p(x|k)p(k)} \tag{8}$$

$$A_r = \ln(p(x,r)p(r)) \tag{9}$$

$$p(r|x) = \frac{\exp(A_r(x))}{\sum_{k=1}^{R} \exp(A_k(x))} \tag{10}$$

In (8), $p(r)$ is an assumption for the class prior probability when the softmax layer used the softmax function, also known as the multiclass generalization of logistic regression, and $p(x|r)$ is the conditional probability of sample given class r, and R is the total number of classes when $A_r = \ln(p(x,r)p(r))$. In Shadman sakib et al. (2019), a proposed fruits recognition system was introduced to recognize fruits by using CNN, the researchers used Fruits- 360 dataset that is obtained in Horea and Mihai (2018), in this study Shadman sakib et al. (2019) the researchers created two convolutional layers, each of them followed by pooling layer, and they created two fully connected layers as shown in fig.6

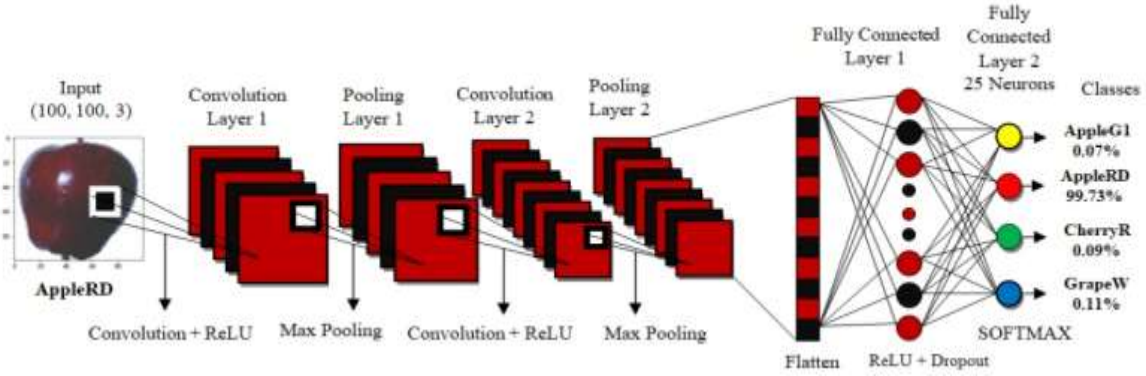


Fig 6. Schematic of the architecture of a convolutional neural network Shadman sakib et al. (2019).

They calculated error rate to compute model performance by using categorical cross-entropy cost function that is known loss function. According to Yash Srivastava et al. (2019), the loss function is used on the top of CNN to judge the goodness of any network, and it plays an important role in CNN training. Basically, it generates a huge loss, if the network does not perform well using the current parameter setting. In Shadman sakib et al. (2019), the researchers used loss function as below:

$$L_i = -\sum_j t_{i,j} \log(p_{i,j}) \tag{11}$$

$$C(w,b) = \frac{1}{2n}\sum_x [y(x) - a^2]^2 \tag{12}$$

Where w is the cumulation of weights in the network, b is all the biases, n is the total number of training inputs and a is the actual output. $C(w,b)$ is non-negative as all the terms in the sum are non-negative. In this study Shadman sakib et al. (2019), they resulted at 5 cases, and each one case achieved sufficient results that were more than 99%, but in this study Shadman sakib et al. (2019) the researchers went to the same dataset that is obtained in Horea and Mihai (2018).

5.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the powerful classification algorithms, which tries to find an optimal separating hyper plane which effectively separates between classes for solving the classification problem Hossam M. Zawbaa et al. (2014). As shown in fig 7, SVM algorithm tries to draw a hyper plane between to classes such that the distance between support vectors can be maximized. According to Whamalai (SVM), SVM performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories.

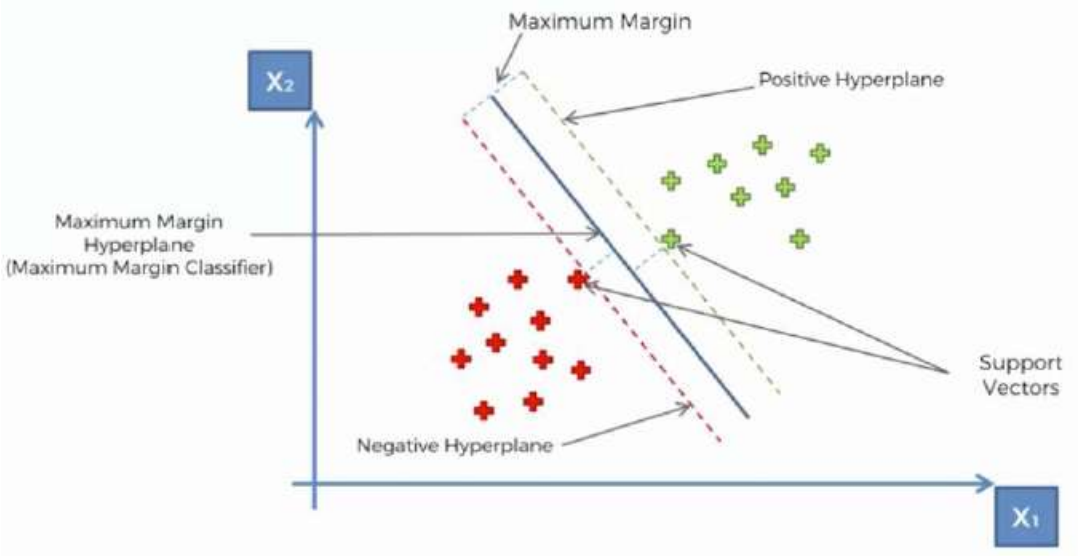


Fig. 7. SVM intuition Sapan Naik and Bankim Patel (2017)

According to Hossam M. Zawbaa et al. (2014), The SVM finds an optimal hyper plane with the maximum margin between two classes by solving the optimization problem, as shown in the equations (13) and (14).

$$maximize \sum_{i=1}^{n} \alpha i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j . K(x_i, x_j) \tag{13}$$

$$Subject - to: \sum_{i=1}^{n} \alpha_i y_i \, , 0 \leq \alpha_i \leq C \tag{14}$$

Where, αi is assigned weight to the training sample x_i when αi > o, the x_i is a support vector. A regulation parameter *C* is used to trade-off the training accuracy and the model complexity to achieve a superior generalization capability. K is a kernel function, which is measure the similarity between two samples. In Hossam M. Zawbaa et al. (2014), the researchers applied K-Nearest Neighborhood (K-NN), and support vector machine (SVM) algorithms to classify different kinds of fruits. When they used SVM algorithm, they achieved classification accuracy 90.91% for apple and 78.89% for orange where it is not sufficient result for orange and there is 9.09% out for apple. In Hong Zheng and Hongfei Lu (2012), the researchers introduced a system to detect the degree of browning on mango fruits by using a least -squares support vector machine (LS-SVM). They produced classification accuracy up to 100%. In Winda Astuti et al. (2018), a proposed approach for classifying fruits based on their shape was introduced under two different techniques, first is SVM algorithm. Second is ANNs. In the training stage, SVM-based fruit classification system gives results of 100% accuracy. Also, the system gives 100% of classification rate in SVM. But in this study Winda Astuti et al. (2018), the researcher

biased to SVM rather ANNs. Also Support vector machine (SVM) was used to classify tomato images into 2 classes (infected/uninfected) in Semary et al. (2015). The proposed system achieved 92% classification accuracy. In Suresha M et al. (2012), a system for apples grading was proposed by using SVM. Classification accuracy achieved 100%.

5.4 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (K-NN) algorithm is one of the widely used machine learning algorithms in classification problems. It can be used in classification, regression and predictive problems. However, it is more widely used in classification problems. According to Adi Bronshtein (2017), K-NN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. K-NN focuses on similarity (proximity) of samples measured by a distance metric and are a statistical classifier. Its job is to assign data to the most represented category within its closest k neighbors Sapan Naik and Bankim Patel (2017). Fruits recognition system was proposed in Woo Chaw Seng and Mirisaee (2009), the researchers used K-Nearest Neighbor (K-NN) algorithm. The Fruit Recognition System using the K-NN algorithm as a classifier to classify fruit based on mean color values, shape roundness value, area and perimeter values of the fruit. The recognition results achieved accuracy up to 90%. In Hossam M. Zawbaa et al. (2014), the researchers proposed model that applies K-Nearest Neighborhood (K-NN) algorithm classification, and support vector machine (SVM) algorithm of different kinds of fruits. When using SIFT features descriptor and K-NN classifier, the classification accuracy was 97.37% for apple and 92.31% for strawberry.

6. Summarization

In this section we summarized some contributions for different algorithms that were used in fruits and vegetables classification through different studies.

Artificial Neural Networks (ANNs) contributions in fruits and vegetables classification through different studies

Feature				Classification Purpose	Fruit Kinds	Accuracy (%)	Reference	Publications date
Color	Size	Shape	Texture					
✓	-	✓	-	Fruit sorting and identification	Apple, Mango, Banana	99.10%	Aibinu et al. (2011)	2011

Some physical sensors were used beside system				Classify fruits and vegetables	Apple, Orange, Papaya, Banana, Guava, , Onion, Lemon, Garlic	Up to 100%	José de Jesús Rubio. (2016)	2017
✓	-	-	✓	Detect fruit defects	Orange	97.25	Giacomo Capizzi et al. (2015)	2015
✓	✓	-	-	Characterize and recognize olive	Olive	96.55%	Gabriel Gatica et al. (2013)	2013

Table 2. Some contributions of ANNs in fruits and vegetables classification.

Convolutional Neural Networks (CNN) contributions in fruits and vegetables classification through different studies								
Feature				**Classification Purpose**	**Fruit Kinds**	**Accuracy (%)**	**Reference**	**Publications date**
Color	**Size**	**Shape**	**Texture**					
✓	✓	✓	✓	Count the calories of fruits and vegetables	Apple, Strawberry, Blueberry, Peach, Pear	75.00%	Morteza Akbari et al. (2016)	2016
A dedicated algorithm was wrote to extract the fruit from the background				Fruit recognition	More different kinds of fruits	96.30%	Horea and Mihai (2018)	2017
-	-	-	-	Fruit category classification	More different kinds of fruits	94.94%	Yu-Dong Zhang et al. (2017)	2017
-	-	-	-	Fruit recognition	More different kinds of fruits	> 99%	Shadman sakib et al. (2019)	2019

Table 3. Some contributions of CNNs in fruits and vegetables classification.

Support Vector Machine (SVM) contributions in fruits and vegetables classification through different studies								
Feature				Classification Purpose	Fruit Kinds	Accuracy (%)	Reference	Publications date
Color	Size	Shape	Texture					
✓	-	✓	-	Fruit recognition	Apple, Orange	(90.91% Apple, 78.89% Orange)	Hossam M. Zawbaa et al. (2014)	2014
Using Scale Invariant Feature Transform (SIFT) descriptor				Fruit recognition	Apple	96.97%	Hossam M. Zawbaa et al. (2014)	2014
-	-	✓	-	Fruit classification	-	Up to 100%	Winda Astuti et al. (2018)	2018
✓	-	-	✓	Classify the infection of tomato	Tomato	92%	Semary et al. (2015)	2015
✓	-	-	-	Apples grading	Apple	100%	Suresha M et al. (2012)	2012

Table 4. Some contributions of SVM in fruits and vegetables classification.

K-Nearest Neighbor (KNN) contributions in fruits and vegetables classification through different studies								
Feature				Classification Purpose	Fruit Kinds	Accuracy (%)	Reference	Publications date
Color	Size	Shape	Texture					
Using Scale Invariant Feature Transform (SIFT) descriptor				Fruit recognition	Apple	97.30%	Hossam M. Zawbaa et al. (2014)	2014
✓	-	✓	-	Fruit recognition	Orange	92.30%	Hossam M. Zawbaa et al. (2014)	2014
Using Scale Invariant Feature Transform (SIFT) descriptor				Fruit recognition	Orange	100%	Hossam M. Zawbaa et al. (2014)	2014

✓	-	✓	-	Fruit recognition	Straw berry	97.10%	Hossam M. Zawbaa et al. (2014)	2014
Using Scale Invariant Feature Transform (SIFT) descriptor				Fruit recognition	Straw berry	100%	Hossam M. Zawbaa et al. (2014)	2014
✓	-	✓	-	Fruit recognition	More different kinds of fruits	Up to 90%	Woo Chaw Seng and Mirisaee (2009)	2009

Table 5. Some contributions of KNN in fruits classification

7. Conclusion

Fruit object requires preprocessing and using more than one feature like color, size, shape and texture because the similarity in colors, shape between fruits like apple and strawberry are similar in color, also apple and orange are similar in shape and so on. The dataset must be collected according to the environment that the classifier will interact. Some of researches achieved good accuracy for fruit classification, but the classification time is not sufficient because some challenges in image size, quality and etc. which can be solved in preprocessing step. Some of researches used open source software for building their classifiers, and they achieved good results. Most researches that achieved good results in fruit classification used perfect images for fruits. So, in the future we can work on imperfect images for fruits. This paper reviewed the basic steps and some techniques that were used for building different classifiers for different tasks in fruit classification. Some of preprocessing methods that were used in previous works were discussed. Feature extraction methods for color, size, shape and texture were explained with SIFT, SURF, HOG and LBP feature descriptors. Finally some machine learning algorithms like ANN and CNN, SVM and K-NN were discussed. In the future, we can work on fruits recognition by using imperfect dataset for fruits. Also, we can work on three types of fruit features that are represented in color, shape and texture to gain good accuracy and enhanced method for preprocessing to gain fast classification time.

Acknowledgment

Special mention goes to my enthusiastic supervisor, Ahmed H. Asad, not only for his tremendous academic support, but also for giving me so many wonderful opportunities.

Similar, profound gratitude goes to Hesham A. Hefny, who has been a truly dedicated mentor. I am particularly indebted to Hesham A. Hefny for his constant faith in my inventions works, and for his support when so generously hosting me in his office. I have very fond memories of my time there.

References

Adi Bronshtein (2017). A Quick Introduction to K-Nearest Neighbors Algorithm. From: https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7.

Aibinu, A.M., Salami, M.J.E., Shafie, A.A., Hazali, N., Termidzi, N. (2011). Automatic fruits identification system using hybrid technique. In: Sixth IEEE International Symposium on Electronic Design, Test and Application (DELTA), Queenstown, pp. 217–221.

A. K. Jain., & A. Vailaya. (1996). Image retrieval using colour and shape. Pattern Recognition, vol. 29, no. 8, pp. 1233-1244.

Alwanin, Rawabi. (2014). An Approach to Image Classification Based on SURF Descriptors and Colour Histograms. Diss. University of Manchester.

Arivazhagan S., Shebiah N., Nidhyanandhan S., & Ganesan L. (2010). Fruit recognition using color and texture features. Journal of Emerging Trends in Computing and Information Sciences, vol. 1, no. 2, pp. 90-94.

Baohua Zhang et al. (2014). Principles, developments and applications of computer vision for external quality inspection of fruits and vegetables: A review. Food Research International, 62, 326–343.

Bay H., Tuytelaars T., & Van Gool L. (2006) SURF: Speeded Up Robust Features. In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg.

Dai-Ton H., Duc-Dung N., & Duc-Hieu L. (2016). An adaptive over-split and merge algorithm for pagesegmentation. Pattern Recogn Lett 80:137–143.

di Huang et al. (2011). Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), DOI: 10.1109/TSMCC.2011.2118750.

Dong ping Tian. (2013). A Review on Image Feature Extraction and Representation Techniques. International Journal of Multimedia and Ubiquitous Engineering, vol. 8, no. 4, pp. 385-396.

Drimbarean, A., & Whelan, P. F. (2001). Experiments in colour texture analysis. Pattern Recognition Letters, Vol.22, No.10, pp.1161-1167.

Fernández, L., Castillero, C., & Aguilera, J. M. (2005). An application of image analysis to dehydration of apple discs. Journal of Food Engineering, vol.67, pp.185-193.

Gabriel Gatica et al. (2013). Olive Fruits Recognition Using Neural Networks. Information Technology and Quantitative Management (ITQM2013).

Giacomo Capizzi et al. (2015). Automatic Classification of Fruit Defects based on Co-Occurrence Matrix and Neural Networks. Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 861–867, DOI: 10.15439/2015F258, ACSIS, Vol. 5, 978-83-60810-66-8, IEEE.

G. Pass., & R. Zabith. (1996). Histogram refinement for content-based image retrieval. In Proc. Workshop on Applications of Computer Vision, pp. 96-102.

Hong Zheng., & Hongfei Lu. (2012). A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning

degree on mango (Mangifera indica L.). Computers and Electronics in Agriculture, vol. 83, pp. 47-51.

Horea Muresan., & Mihai Oltean. (2018). Fruit recognition from images using deep learning. Acta Univ. Sapientiae, Informatica, vol. 10, no. 1, pp. 26-42.

Hossam M. Zawbaa., Mona Abbass., Maryam Hazman., & Aboul Ella Hassenian. (2014). Automatic fruit image recognition system based on shape and color features. International Conference on Advanced Machine Learning Technologies and Applications (AMLTA), vol. 488, pp. 278-290.

J. Huang et al. (1997). Image indexing using colour correlogram. In Proc. CVPR, pp. 762-765.

José de Jesús Rubio. (2016). A method with neural networks for the classification of fruits and vegetables. Soft Computing, pp. 1-14.

M. Flickner et al. (1995). Query by image and video content: the QBIC system. IEEE Computer, vol. 28, no. 9, pp. 23-32.

Mingqiang, Y., Kpalma, K., & Ronsin, J. (2008). A Survey of shape feature extraction techniques. In book: Pattern Recognition, Publisher: IN-TECH, Editors: Peng-Yeng Yin, pp.43-90.

Morteza Akbari., Hamed Haddadi., & Alireza Tavakoli. (2016). Fruits and vegetables calorie counter using convolutional neural networks. International ACM Conference on Digital Health, pp. 121-122.

M. T. Chowdhury., M. S. Alam., M. A. Hasan., & M. I. Khan. (2013). Vegetables detection from the glossary shop for the blind. IOSR J. Electr. Electron. Eng. 8, 43–53.

Muhammed J. A. Patwary., Shahnaj Parvin., & Subrina Akter. (2015). Significant HOG-Histogram of Oriented Gradient Feature Selection for Human Detection**.** International Journal of Computer Applications, Volume 132 – No.17.

N.A. Semary et al. (2015). Fruit-Based Tomato Grading System Using Features Fusion and Support Vector Machine. Advances in Intelligent Systems and Computing 323, DOI: 10.1007/978-3-319-11310-4_35.

Ramin Zabih et al. (1996). Comparing Images Using Color Coherence Vectors. Proceeding MULTIMEDIA '96 Proceedings of the fourth ACM international conference on multimedia, pp. 1-14.

Rishav Chakravarti., & Xiannong Meng. (2009). A Study of Color Histogram Based Image Retrieval. Sixth International Conference on Information Technology: New Generations, pp. 1323-1328.

Sapan Naik., & Bankim Patel. (2014). CIELab based color feature extraction for maturity level grading of Mango (Mangifera Indica L.). National journal of system and information technology (0974-3308), VOLUME 7, NO. 1.

Sapan Naik., & Bankim Patel. (2017). Machine vision based fruit classification and grading - a review. International Journal of Computer Applications, vol. 170, no.9, pp. 22-34.

Sergio Cubero et al. (2011). Advances in Machine Vision Applications for Automatic Inspection and Quality Evaluation of Fruits and Vegetables. Food Bioprocess Technol 4:487–504, doi 10.1007/s11947-010-0411-8.

Shadman sakib et al. (2019). Implementation of Fruits Recognition Classifier using Convolutional Neural Network Algorithm for Observation of Accuracies for Various Hidden Layers. https://www.researchgate.net/publication/332258470, vol. 01.

Suresha M et al., Shilpa N.A., & Soumaya B. (2012). Apples Grading based on SVM Classifier. International Journal of Computer Applications (0975-8878), pp. 27-30.

Tasneem Gorach. (2018). DEEP CONVOLUTIONAL NEURAL NETWORKS- A REVIEW. International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 07.

Tony Lindeberg. (2012). Scale Invariant Feature Transform. https://www.researchgate.net /publication/235355151, DOI: 10.4249.

Whamalai (SVM). Support Vector Machine. From: cs.joensuu.fi/pages/whamalai/ expert/svm.pdf.

Winda Astuti et al. (2018). Automatic fruit classification using support vector machines: a comparison with artificial neural network. the 2nd International Conference on Eco Engineering Development, doi:10.1088/1755-1315/195/1/012047.

Woo Chaw Seng., & Seyed Hadi Mirisaee. (2009). A New Method for Fruits Recognition System. Electrical Engineering and Informatics, vol. 01, pp. 130-134.

Yash Srivastava et al. (2019). A Performance Comparison of Loss Functions for Deep Face Recognition. From: https://arxiv.org/abs/1901.05903.

Yu-Dong Zhang et al. (2017). Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. Multimedia Tools Appl, pp. 1-20.

Zhang 100-CH03. NeuralNetworkStructure. From https://www.ieee.cz/knihovna /Zhang/Zhang100-ch03.pdf.

A soft-computing approach for Increasing Relational Databases Flexibility

Ahmed I. Wagih [1] ahmed.ib.wagih@gmail.com
Ahmed M. Gadallah [2] ahmgad10@yahoo.com
Hesham Ahmed Hefny [3] hehefny@hotmail.com
[1,2,3] **Faculty of Graduate Studies for Statistical Research , Cairo University ,Egypt**

ABSTRACT

Commonly, traditional database management systems are unable to satisfy human needs for data selection based on Linguistic expressions and degrees of truth but it is still a Powerful tool. The target of this paper is to make the database management system more flexible to enable the human-like expressions and make them suitable for database definition and manipulation. Also it attempts to make the database management system have the ability of computing with words based on fuzzy set theory as a soft-computing approach. It adds some features to relational databases to enable users to benefit from the flexibility of fuzzy set theory. Such features are applied to both data manipulation and definition languages. It include Fuzzy Aggregations, Fuzzy Relationships , Fuzzy data type, Fuzzy attribute, fuzzy Insert, Fuzzy update, Fuzzy Delete, Fuzzy query and Fuzzy join. An illustrative case study shows the benefits of the proposed approach against some recent related ones.

Keywords

Fuzzy Query, Fuzzy Logic, Information Retrieval, Fuzzy SQL. Fuzzy DB, FSQL, FSQL Server

1. INTRODUCTION

Database management system (DBMS) is a computer program designed to manage a database; a large set of structured data, and run operations on the data requested by numerous users. It manipulates and retrieves data which is crisp and precise by nature. In contrary, it is unable to do human-like processes in data which are uncertain, imprecise and vague in nature like inserting salary is good or updating salary from good to very good or selecting where salary is good. It is unable to store fuzzy data .The researcher looks forward to create a new approach that allows this fuzzy features in the database. This approach enables us to store fuzzy data, retrieve fuzzy Linguistic expressions and update fuzzy data based on fuzzy Linguistic expressions. So, the classical database system is unable to satisfy needs for data selection based on Linguistic expressions and degrees of truth. also This new approach will be able to be very sensitive with nature of data for example database has players table and its attributes are (name , age , height), players ages between 10 to 30 years , the manager want to choose the best players in playing basketball by comparing the height of the players . This example has a problem when putting

fuzzy function on attribute ‘height ’ , fuzzy function on 'height ' with player that has 10 years old different from player that has 30 years old. The new approach of fuzzy database can deal with this problem.

2. PREVIOUS WORKS

This section presents several approaches that have been developed for allowing fuzzy databases. Unfortunately, most of previously developed fuzzy database have set of Defects. In [Dwibedy, Sahoo and Dutta (2013)] This paper redefines the fuzzy class in an efficient manner and proposes the structure of the fuzzy class using more effective generalized techniques to develop a new object based fuzzy data model in order to manipulate imprecise information and exposed to wider range of applicability Also, The researcher defines a formal framework for generalized fuzzy constraints which can be applied effectively to fuzzy specialized classes in fuzzy class hierarchy. Yet, this approach works on Entity-Relationship (ER) Model level only including FUZZY CLASS, FUZZY SUPER CLASS- SUBCLASS and CONSTRAINTS TO FUZZY SUPER CLASS - SUB CLASS. The authors in [Kacprzyk (2015)] give a comprehensive, brief presentation of the history and some original solutions of one of the most relevant and interesting areas related to the use of fuzzy logic in database management systems. Not ably in its querying component, and – to some extent – in a broader issue of data and information management. The researcher briefly summarizes the roots of those new applications of fuzzy logic more relevant proposals and development in the context of fuzzification of the basic relational database model, and then some of its features in other generalizations. The researcher particularly focuses on fuzzy querying as a human consistent and friendly way of retrieving information due to real human intentions and preferences expressed in natural language represented via fuzzy logic and possibility theory. The researcher mentions some extensions, notably fuzzy queries with linguistic quantifiers, and points their close relation to linguistic summaries. As for newer, prospective developments, The researcher mainly focuses on bipolar queries that can accommodate the users’ intentions and preferences involving some sort of a required and desired, mandatory and optional, etc. conditions. The researcher shows various ways of handling such queries. The researcher concludes with some brief position statements of researcher’s view on relevant and promising directions, and challenges. Some approaches about fuzzy ER/EER model have been published recently. Few of these works studies how to relax constraints expressed in the data model in [Galindo (2003)]. In this article the researcher’s aim is to relax some constraints which have not been studied in previous works and to extend the EER model with fuzzy capabilities. The researcher studies new constraints that are not considered in classic EER models. The researcher uses the fuzzy quantifiers which have been widely studied in the context of fuzzy sets and fuzzy query systems for databases. The researcher also examines the representation of these new features in an EER model and their practical repercussions. The studied extensions are: fuzzy disjointed or overlapping constraints on specializations, fuzzy attribute defined specializations, fuzzy constraints in union types or categories and fuzzy constraints in shared subclasses (or intersection types). All these fuzzy extensions have a novel meaning and offer greater expressiveness in conceptual design and these constraints may be

used in Data Mining activities. Generally speaking, it is impossible to capture all of the semantics of real world information and to model it in a perfect way. Uncertain information being lacking information, individual interpretation of knowledge is taken into account. Most widely used relational model faced much difficulty to handle complex data requirements. The size of databases and their changing composition such as graphics, video and sound, as well as numbers and text invited a reorganization of existing information systems. Object oriented database draw their strength from graphical user interfaces, advanced data management capabilities and powerful modeling techniques. So in order to deal with inexactness fuzzy techniques have been extensively integrated with more advanced object oriented database system which is capable to represent and manipulate the complex objects as well as complicated and uncertain relationship exiting among them.

On the other hand, uncertainty handling using Fuzzy Object Oriented Database and Fuzzy Relational Database is illustrated in [Sonia (2013)]. In data mining applications, sharing of huge volume of detailed personal data is proved to be beneficial. Different types of data include criminal records, medical history, shopping habits, credit records, etc. These types of data are very much significant for any company and governments for decision making. When analyzing the data, privacy policy may avoid data owners from sharing information. In privacy preserving presented in [Saxena and Pushkar (2009)], data owner must provide a solution for achieving the dual goal of privacy preservation as well as accurate clustering result. Also, the research done in [Kumar (2016)] is achieved for designing and representing a fuzzy object-oriented database through UML, for that purpose, a domain of hospital has been taken for diagnosis the patient for various diseases with their uncertain input values (disease sign, symptoms, and pathology test result values). This will enhance the functioning of the developed software design on the present approach. A UML class & Sequence diagram are also designed for representing the complete process of registration and diagnosis of patients and design the fuzzy object-oriented database i.e. FOOD through the SQL server 2008. On the other hand, the Researchers in [Hoque, Ali and Aktaruzzaman (2008)] designed a sample fuzzy database and applied both classical and fuzzy query on this fuzzy database. This work shows that, the time cost of fuzzy query on classical database (DB) is the same as the classical query on classical DB. But the time cost of the fuzzy query on sample fuzzy database has been reduced. Commonly, With the development of cloud computing and big data, data privacy protection has become an urgent problem to solve. Data encryption is the most effective way to protect privacy; however, it will change the data format and result in: (1) Database structure and application software will be changed; and (2) structured query language (SQL) operations cannot work properly, especially in SQL-based fuzzy query. As a result, a need arise for SQL-based fuzzy query mechanism over encrypted databases, including traditional databases and cloud outsourced databases. The work presented in[Liu (2014)] establishes a secure database system using format-preserving encryption (FPE) as the underlying primitive to protect the data privacy while not change the database structure. It further proposes a new SQL-based fuzzy query mechanism supporting directly query over encrypted data, which is constructed by FPE and universal hash function (UHF). The security of

the proposed mechanism is analyzed as well. In the end, it makes extensive experiments on the system to demonstrate its practical performance.

3. Limitations of Previous Works

The goal of work presented in [Hudec (2009)] is to capture linguistic expressions and make them suitable for queries. For this purpose the fuzzy generalized logical condition for the WHERE part of SQL was developed. In this way, queries based on linguistic expressions are supported and are accessing relational databases in the same way as with the SQL. Fuzzy query is not only a querying tool. it improves the meaning of a query and extracts additional valuable information. the lack of [Hudec (2009)] approach is as follows, it does not take into account the data sensitivity, It Forces the user to add stored procedure for each linguistic variable, it has some limitations on the fuzzy query. For example, fuzzy aggregation does not take into account fuzzy operators or fuzzy hedges and it is has one linguistic variable for all attributes in the database.

Fuzzy conceptual data modeling is concentrated in [Kumar, Jena and Devi (2013)]. Based on dealing with imprecise and uncertainty super class-subclass relationships, a conceptual data model ER is enhanced. Different levels of fuzziness are introduced and the corresponding super class-subclass relationships representations are given. ER data model is thus enhanced to fuzzy ER data model denoted EER. Where that could store, retrieve, and manipulate the imprecise and uncertain complex data is stored in the form of objects with fuzzy techniques applied on it to encapsulate the impreciseness and uncertainty. In this research, the attention is pain on the fuzzification of attributes, and objects, especially on that of super class-subclass relationships.

Paper in [Sabour, Gadallah and Hefny (2014)] presents a flexible fuzzy-based approach for querying relational databases. Although, many fuzzy query approaches have been proposed, there is a need for a more flexible. The lack of this approach is as follows, it does not take into account the data sensitivity, it Forces the user to add stored procedure for each linguistic variable for each user, it has some limitations on the fuzzy query, for example, fuzzy aggregation, fuzzy join ,it does not take into account fuzzy operators or fuzzy hedges, it has One linguistic variable for all attributes in the database and The volume of Fuzzy Meta-knowledge Base is very big compared to any other approach.

GEFRED model in [Grissa and Ben Hassine (2009)]. The lack of this approach is as follows, it cannot deal with all data types in a database, does not take into account the sensitivity of the data, does not take into account fuzzy operators or fuzzy hedges and depends on personal appreciation in insert and update.

In [Galindo (2008)] ,The researcher presents the Soft-SQL project whose goal is to define a rich extension of SQL aimed at effectively exploiting flexibility offered by fuzzy sets theory to solve practical issues when querying classic relational databases. The Soft-SQL language is based on previous approach that introduced soft conditions on tuples in the classical relational database

model. The researcher retains the main features of these approaches and focus on the need to provide tools allowing users to directly specify the context dependent semantics of soft conditions. the SELECT command in this approach is extended in order to support soft predicates based on the user defined term sets, the semantics of grouping and aggregation can be modified, and finally, the clauses in the SELECT command can be combined effectively. The lack of this approach is as follows, it cannot deal with all data types in a database ,it does not take into account fuzzy operators or fuzzy hedges, it does not solve the problem of the semantics of some linguistic values might rely on the context for example; the notion of cheap flat in Milan is not the same in Paris or New York Further, users cannot explicitly define the semantics of the linguistic values.

4. THE PROPOSED FUZZY DATABASE APPROACH

The following subsections explain a set of used notation and definitions that are used in the proposed approach.

4.1 Fuzzy number notation

A Fuzzy number is generalization of a regular, real number in the sense that it does not refer to one single value but rather to a connected set of possible values, where each possible value has its own weight between 0 and 1, this weight called the membership. There are many fuzzy number types, this paper will take care of triangle and trapezius as an example of the fuzzy number. The notation of triangle as fuzzy number in database management system is ~tri-x-y-z where x , y and z are real numbers as shown in fig.1.

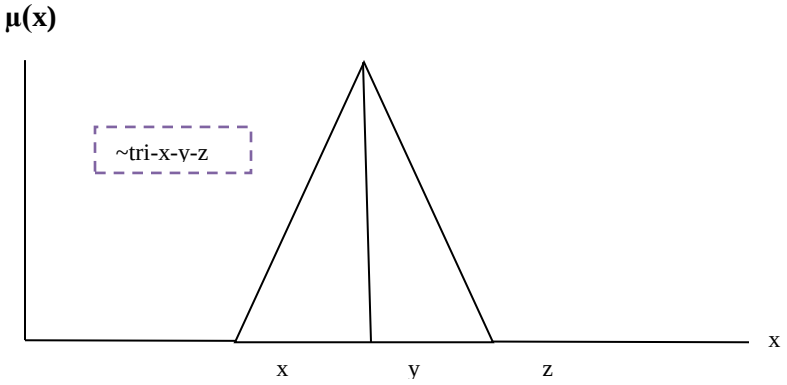


Fig. 1 . Triangle Fuzzy Number

Also, the notation of trapezius as fuzzy number in database management system is ~tra-w-x-y-z where w , x , y and z are real numbers as shown in fig.2.

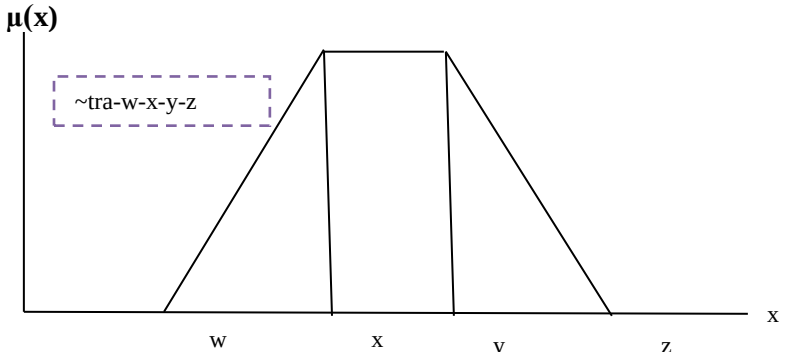


Fig. 2 . Trapezoid Fuzzy Number

4.2 Fuzzy number operations

The database system will be able to understand different calculations for all the fuzzy numbers. Fuzzy number operations based on the respective real arithmetic operation in [Hefny (2014)] are defined as follows:

i. ***Fuzzy Addition:***

~tri-x-y-z + ~tri-a-b-c = ~tri-r1-r2-r3

where r1= x+a , r2= y+b , r3 = z+c .

~tra-w-x-y-z + ~tra-a-b-c-d = ~tra-r1-r2-r3-r4

where r1= w+a , r2= x+b , r3 = y+c , r4 = z+d .

ii. ***Fuzzy Subtraction:***

~tri-x-y-z - ~tri-a-b-c = ~tri-r1-r2-r3

Where r1= x-a , r2= y-b , r3 = z-c .

~tra-w-x-y-z - ~tra-a-b-c-d = ~tra-r1-r2-r3-r4

Where r1= w-a , r2= x-b , r3 = y-c , r4 = z-d .

iii. ***Fuzzy Multiplication:***

~tri-x-y-z * ~tri-a-b-c = ~tri-r1-r2-r3

Where r1= x*a , r2= y*b , r3 = z*c .

~tra-w-x-y-z * ~tra-a-b-c-d = ~tra-r1-r2-r3-r4

Where r1= w*a , r2= x*b , r3 = y*c , r4 = z*d .

iv. ***Fuzzy Division:***

~tri-x-y-z / ~tri-a-b-c = ~tri-r1-r2-r3

Where r1= x/a , r2= y/b , r3 = z/c .

~tra-w-x-y-z / ~tra-a-b-c-d = ~tra-r1-r2-r3-r4

Where r1= w/a , r2= x/b , r3 = y/c , r4 = z/d .

4.3 Fuzzy data domain

Data domain refers to all the values which a data element may contain; therefore fuzzy data domain refers to set off the fuzzy numbers which a data element may contain for example fuzzy data domain shown in fig.3.

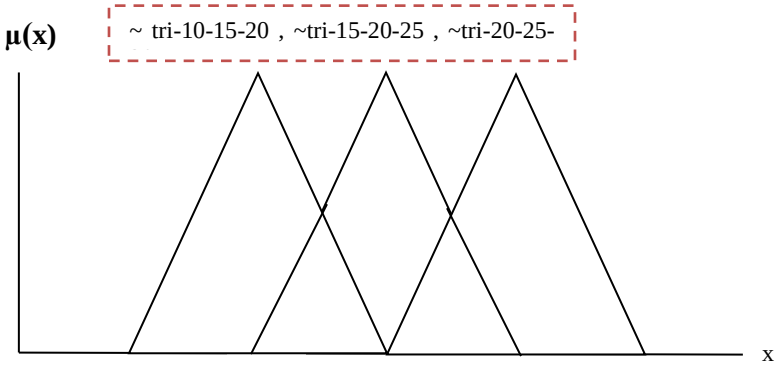


Fig. 3 . Fuzzy Data Domain

4.4 Fuzzy data type

A data type or simply type is a classification of data which tells the compiler or interpreter how the programmer intends to use the data, this data type defines the operations that can be done on the data. Therefore fuzzy data type tells the compiler intends to use the fuzzy data. Any data type has a name, domain and Predefined operations , in this approach users can define the fuzzy data type by declaring the name and fuzzy data domain and the compiler can understand the fuzzy data operations.
Example to create fuzzy data type

Create fuzzy data type ~size as Small (~tri-10-20-30), Medium(~tri-20-30-40) , Large (~tri-30-40-50)

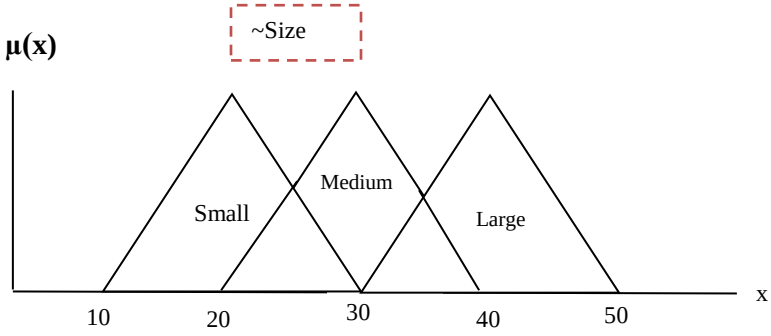


Fig. 4 . Fuzzy Data Type

4.5 Built-in Fuzzy data functions

Built in Fuzzy data functions are built into database management system and it can be accessed by end-users as follows:

I. get_memberShip

This function gets the maximum membership for any fuzzy type and a linguistic value. For example consider operation Op1.

Op1 : Decalre @size as ~size ;
Set @ size = 'Medium:٧٥';
get_memberShip(Size,'Medium')

In Op1, size is a data variable of fuzzy data type ~size , this function returns the maximum membership of medium in fuzzy data type ~size Fuzzy type. Figure 4 shows size as a fuzzy data type, so the returned value of this function is ٧٥.

II. get_linguistic

This function gets the linguistic value for any fuzzy type variable for example operation

Op2:Decalre @size as ~size
Set @ size = 'Medium:٧٥'
get_linguistic(Size)

In Op2 , size is a data variable of fuzzy data type ~size , this function returns the linguistic value of variable Size , the Result of this function is 'Medium' .

4.6 Fuzzy data type declaration

Commonly, users can declare and use this fuzzy data type as any other data type in database management system for example

Op3: Declare @x as ~size(10)
Set @x = 'Medium:75'

As depicted in Op3 , the user declares variable x as a fuzzy data type ~size with scaling value 10, the meaning of this scaling value is scaling the fuzzy data type values by 20 as shown in fig.5. The user set the fuzzy data variable x by linguistic value Medium and membership 50 means that the value of x is 275 or 325, this approach deals with minimum value 275.

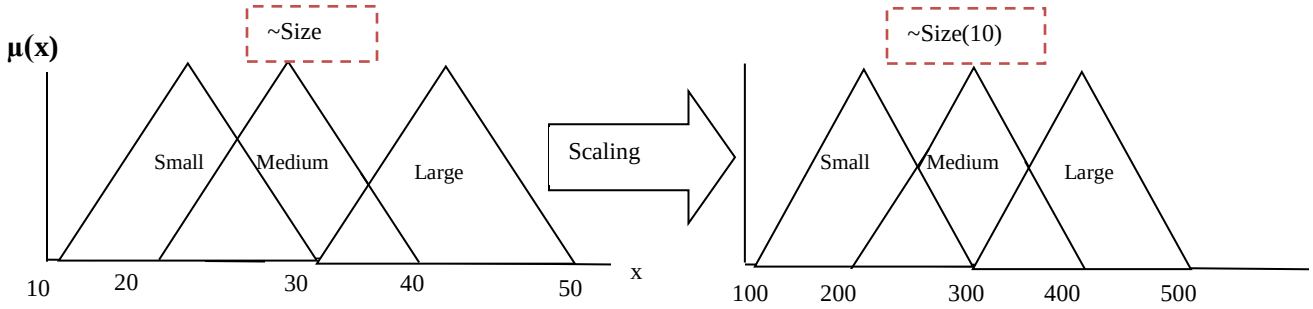


Fig. 5 . Scaling of fuzzy values in fuzzy data type

4.7 Fuzzy data type operations

The user can perform basic calculations on the same fuzzy data type for example

Op4: Declare @x as ~size (٥);

Declare @y as ~ size (١٠);

Declare @r1 as ~ size (10);

Declare @r2 as ~ size (2);

Declare @r3 as ~ size (500);

Declare @r4 as ~ size (.1) ;

Set @x = ~Medium:50 ; Set @y = Small:50 ;

Set @r1= @x+@y ; Set @r2= @y-@x ;

Set @r3= @x*@y ; Set @r4= @y/@x;

When executing Pp4 with @x =125 , @y =150, the values assigned to r1, r2, r3 and r4 will be as follows:

r1 will be 125+150= 275 in **~size (10)** means **'Medium:75',**

r2 will be 150-125= 25 in **~size (2)** means **'Small:25'**,

r3 will be 150*125= 18750 in **~size (500)** means **'large:75' , and**

r4 will be 150/125= 1.2 in **~size (.1)** means **'Small:20'**.

4.8 Fuzzy data type conversion

In this approach the user can convert between different fuzzy data types for example

Op5 : Declare @x as ~Size (2)

Set @x = 'Medium:50'

Declare @x2 as ~Size (1.5)

Set @x2 =cast @x as ~Size (1.5)

Also, in Op5 , the result of *x = 50* is 'Medium:50' while with *x2 =50* is 'Medium:66.6' .

4.9 Fuzzy attributes

The proposed approach allows users to use fuzzy data type as fuzzy data attribute for example.

Op6 : CREATE TABLE employees
(EmpID int, nam varchar(255),
Position varchar(255), Salary ~salary(5000))

In Op6 , the user creates the employees table with 4 attributes : Empid , name, position and fuzzy attribute ~salary with scaling value 5000. Also the user can make a variable scaling value depending on the position attribute in Op7 as follows.

Op7 : CREATE TABLE employees
(empid int, name varchar(255),
position varchar(255), salary ~salary (case when Position ='manager' then 10000 else 5000 end))

So this variable scaling value reflects the sensitivity of data.

4.10 Fuzzy Operators in database expressions

Fuzzy operator is determined by a threshold with certain value, the Expression that includes fuzzy operator is true when the membership of fuzzy attribute is equal or greater than the value of this threshold, a threshold is alpha cuts for this operators as shown in op8.

Op8 : select * from customers where ~salary almost good

Almost is a fuzzy operator and its threshold store in the fuzzy database system for example 60% That means that statement condition is retrieve all attribute from Customers where salary is good by membership 60% or more . op9 shows how to create fuzzy operator in database system with Data definition language.

Op9 : Create fuzzy operator almost with 60

4.11 Linguistic hedges

Fuzzy Every Linguistic variable can have multiple of linguistic values each Linguistic value is called "atomic term".Such atomic terms represent the values allowed for the linguistic variable inwards over its universe of Discourse. This is quite suitable for use in natural language. For example, the linguistic variable "temperature" may have the following linguistic values , or atomic terms : (cold , cool , warm and hot). In natural language, atomic terms are often modified with the adjectives(nouns) , or adverbs (verbs) like :Very , low , slightly , more or less , fairly , ……. etc.These adjectives are the hedges as shown in op10.

Op10 :select * from customer where ~salary almost very slightly good

Almost is a fuzzy operator, very and slightly are hedges. Fuzzy hedge is determined by a value, op11 shows how to create fuzzy hedges in database system with Data definition language.

Op11 :Create fuzzy hedges very with 2
Create fuzzy hedges slightly with .5

to calculate the membership of atomic term with all hedges calculate the membership power the first hedge value then power the second hedge value to the end of hedges and The fuzzy operator deal with the final value as a membership .

4.12 Fuzzy database operations

The proposed fuzzy database approach aims to support human-like queries and fuzzy database operations, which almost contain linguistic terms and fuzzy connectors. Such linguistic terms include linguistic variables, linguistic values and fuzzy hedges and deal with fuzzy data type .

I. Fuzzy Query

In Op12, The user want to retrieve data from employees where salary can achieve this Linguistic term "almost not very mins low "

Op12:select * from employees where salary almost not very mins low

in natural database system the database user cannot achieve this query. Assuming that The salary column is of the fuzzy data type . "almost" is a Predefined fuzzy operator . 'not','very','mins' and 'low' is a Predefined fuzzy hedges . "salary" is a Predefined fuzzy data type.

II. Fuzzy Insert

Op13 shows The salary column is of the fuzzy data type so its input is fuzzy data "good:60,bad:40". This expression expresses the ability of the fuzzy database user in this approach to enter the fuzzy data. in natural database system the database user cannot achieve this expression.

Op13: insert into employees (empid,name, position,salary)

values ('123','mohamed', 'manager','good:60,bad:40')

III. Fuzzy Delete

Op14 shows The same query idea, the database user can delete any data he wants by writing the appropriate expression which includes fuzzy hedges , fuzzy operators and fuzzy columns. In this sentence , "almost" is a Predefined fuzzy operator . 'not','very','mins' and 'low' is a Predefined fuzzy hedges . "salary" is a Predefined fuzzy data type .

Op14:"delete from employees where salary almost not very mins low "

IV. Fuzzy Update

Op15 shows The salary column is of the fuzzy data type so its input is fuzzy data "very_good:40". The same query and insert idea, the database user can update any data he wants by writing the appropriate expression which includes fuzzy hedges , fuzzy operators and fuzzy columns and setting fuzzy column by fuzzy data .

Op15 : update employees Set salary = 'very_good:40' where salary almost very low

V. Fuzzy Join

By adding this fuzzy data type to database, the database user can perform many flexible joins that take into account data sensitivity.

Op16 : CREATE TABLE employees (emp_id int, name varchar(255), salary ~salary (2));

CREATE TABLE customers (cust_id int, name varchar(255), salary ~salary (5));

In Op16 ,the scaling value for employee salary is 2 by the scaling value for customer is 5 . fuzzy data type ~salary(2) and ~salary(5) shown in fig.6, Table .1 presents employees data and Table .2 presents customers data.

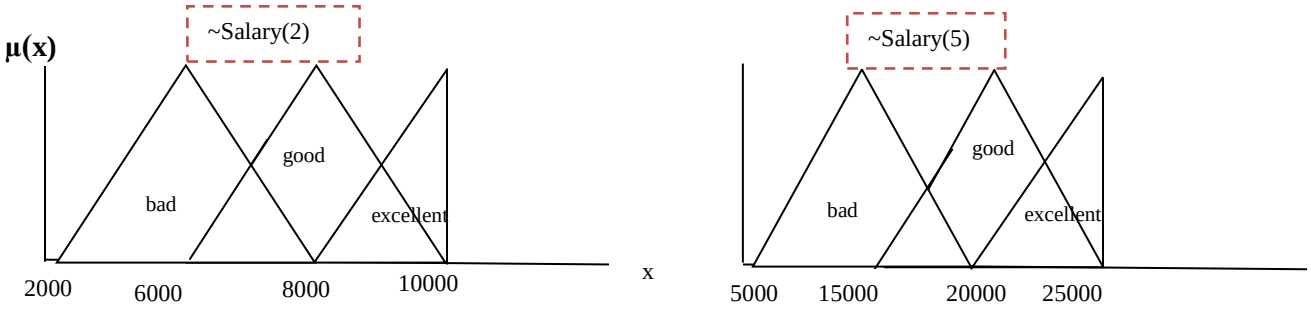


Fig. 6 . Different Type Of Fuzzy Data Type

table . 1 employees

Id	salary val	salary fuzzy val
1	7500	good:75
2	6500	good:25
3	8500	excellent:25
4	9500	excellent:75

table . 2 customers

Id	salary val	salary fuzzy val
10	18750	good:75
20	16250	good:25
30	21250	excellent:25
40	23750	excellent:75

On the other hand, the proposed approach allows comparing the fuzzy data type ~salary(1000) in employees and the fuzzy data type ~salary(15000) in customers. In this approach, the result of Op17 query is shown in fig.7

Op17: select * from employees e inner join customers c on e.salary = c.salary

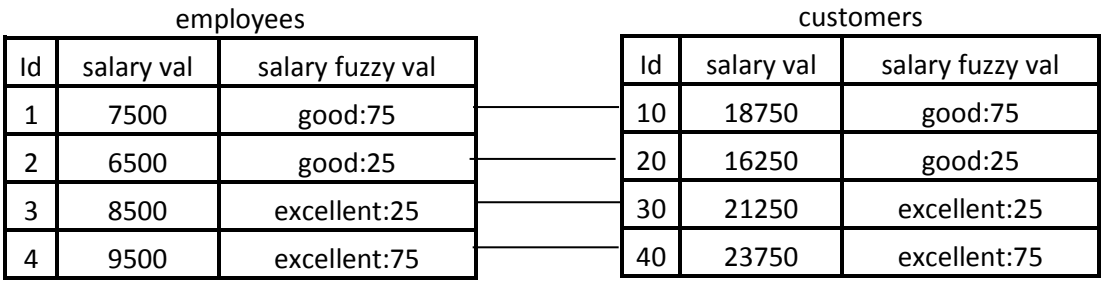


Fig . 7 . Join between Employees and Customers

The database user can use the fuzzy join in another way as shown in Op17 and Fig . 8.

Op17: SELECT * FROM employees e INNER JOIN customers c

ON (get_linguistic(e.salary) =

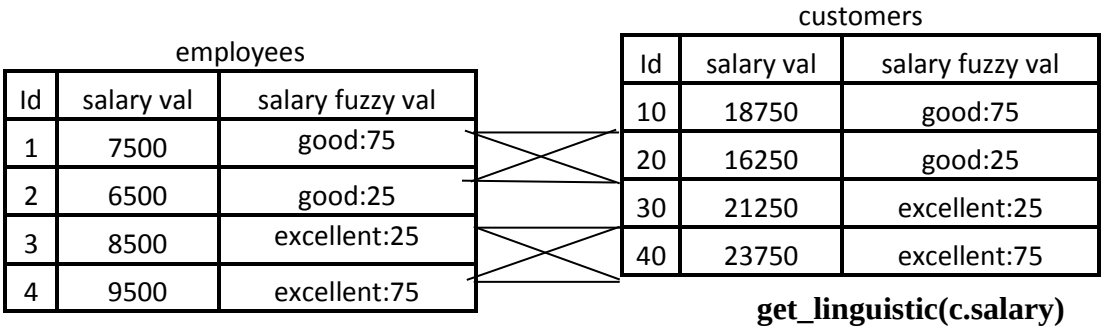


Fig . 8 . Join between Employees and Customers 2

VI. Fuzzy Aggregation

Assuming that the salary is a fuzzy attribute. the database user want to retrieve the count for each level of salary from employees table . natural database system cannot achieve this fuzzy aggregation , op18 shows the query that achieves this criteria.

Op 18 :select count(*) , get_languistic(salary) from employees e group by get_languistic(salary)

5. Conclusion

This paper presents a method proposed for fuzzy database system. this system will be useful in dealing with fuzzy data. The researcher has applied this technique in practice under the SQL

server 2012 as database engine and programming language C # . Of course after applying the new approach to create fuzzy database system , The system has become very clever, traditional database does not have the following functionality ,it can not achieve fuzzy query like retrieving data from employees where the salary attribute can achieve this Linguistic term "almost not very low ", the user cannot achieve fuzzy insert, it cannot achieve fuzzy aggregation like retrieving the count of each level of salaries namely, high, medium and low salaries from employees table and the user cannot achieve fuzzy join like retrieving employees taking same salaries. In contrary, the user can achieve fuzzy join , fuzzy aggregation using the proposed approach , and the proposed approach allows users to insert human-like uncertain data likes salary is good with a membership 80% . the proposed approach does not need to add stored procedures or functions for each linguistic variable. the volume of Fuzzy Meta-knowledge Base in proposed approach is very small compared with any of the previous approaches. The proposed approach can do all of fuzzy database operations (fuzzy insert, fuzzy select, fuzzy update, fuzzy delete, fuzzy aggregation and fuzzy join). Also, it deals with data sensitivity and works very efficiently on database engines.

6. REFERENCES

1.Galindo, J. (2008): Handbook of Research on Fuzzy Information Processing in Databases. Chapter XI FSQL and SQLf: Towards a Standard in Fuzzy Databases.

2.Grissa, A., Ben Hassine, M.(2009): New Architecture of Fuzzy Database Management Systems. The International Arab Journal of Information Technology" The International Arab Journal of Information Technology, Vol. 6, No. 3.

3.Elmasri, R., Navathe, S.B. (2011): Fundamental of Database Systems .

4. Sabour, Adel.A., Gadallah, Ahmed.M., Hefny, Hesham.A. (2014): Flexible Querying of Relational Databases fuzzy set Approach Institute of Statistical Studies and Research , International Conference on Advanced Machine Learning Technologies and Applications, pp 446-455, Cairo University .

5.Dwibedy, Debasis., Sahoo, Laxman., Dutta, Sujoy. (2013): A New Approach to Object Based Fuzzy Database Modeling, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-1.

6.Hudec, Miroslav. (2009): An Approach to Fuzzy Database Querying, Analysis and Realization, INFOSTAT, UDC 004.4'2, DOI: 10.2298/csis0902127H.

7.Kumar, Nitesh., Jena, Abhishek., Devi, Thiyam.Romila. (2013): An Enhanced ER Conceptual Data Modelling for Fuzzy Object-Oriented Database Models (FOODBMs), International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 10.

8.Peng, Chong. (2015): cutting database system based on machining features and TOPSIS, Robotics and Computer-Integrated Manufacturing, LA, 70803, USA.

9. Kacprzyk, Janusz.(2015): Fuzziness in database management systems: Half a century of developments and future prospects, m3SC+; v1. 207; Prn; 8:58] P.1 (1-9).
10. Galindo (2003): Fuzzy Extensions to EER Specializations, International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD'03), pp. 218
11. Sonia, Subita.Kumari. (2013): Fuzzy Object Oriented Database versus FRDB for Uncertainty Management, International Journal of Computer Applications (0975 – 8887) Volume 74– No.17.
12. Saxena, V.K., Pushkar, Shashank. (2009): Fuzzy-Based Privacy Preserving Approach in Centralized Database Environment, Springer Science Business Media Singapore, Advances in Computational Intelligence, Advances in Intelligent Systems and Computing 509, DOI 10.1007/978-981-10-2525.
13. Kumar, Santosh. (2016): Object-Oriented Fuzzy Database Representation through UML, International Journal of Computer Applications (0975 – 8887) Volume 147 – No.6, August 2016.
14. Hoque, A.H.M.Sajedul., Ali, Sadek., Aktaruzzaman (2008): Performance Comparison of Fuzzy Queries on Fuzzy Database and Classical Database, 5th International Conference on Electrical and Computer Engineering ICECE 2008, 20-22.
15. Kumar, Praveen. (2011): A Survey of Fuzzy Techniques in Object Oriented Databases", International Journal of Scientific & Engineering Research, Volume 2, Issue 11.
16. Liu, Zheli. (2014): SQL-Based Fuzzy Query Mechanism Over Encrypted Database" International Journal of Data Warehousing and Mining, 10 (4), 71-87.
17. Hefny, Hesham. A. (2014): Introduction to Fuzzy Sets and Fuzzy Logic Book.

Soft-Computing Technique for Predicting Crops Diseases Respecting the Ongoing Changes in Climate

Ahmed M. Yousef[1], Ahmed M. Gadallah[2], Maryam Hazman[3], Hesham A. Hefny[4]

Abstract

Plant diseases cause unprecedented crop loss across the world. The effect on crop depends on the disease involved, the type of crops grown, the management practices used, and the various environmental factors. The integration of plant disease management supports the use of multiple control measures including where possible. Long term predictions of plant disease provide the decision makers with the warnings that one or more diseases may occur at a particular location in a particular time in the future. This Prediction of that diseases which will be occurred, ensures that control measures, especially the treatments are used more effectively to avoid these disease by not planting it in that area or take the actions to be ready to face it. This paper provides a fuzzy long term prediction technique of agriculture diseases with climate changes in a given place. The proposed technique consists of three phases; one for crop's diseases determination for each crop and the suitable environmental temperature and humidity for that disease, the second one for defining the disease cause's membership functions and the third for fuzzy-based Disease detection phase. The proposed technique is mainly based on the spatial agro-climatic database. It proved that some of crops have much risk and may be lost or are not suitable at that area because of the risks of diseases infectious if we plant it at that traditional plantation time like past years.

Keywords: Fuzzy, plant disease, prediction, spatial database, weather-driven, disease epidemics, warning system, climatology

1-INTRODUCTION

Plant diseases cause significant loss of valuable food crops throughout the world. Diseases cause at least 10% of crop losses globally and are in part responsible for the lack of adequate food [Strange and Scott, 2005]. Plant diseases may reduce crop yields depending on the diseases involved, the crop species grown, the management practices followed, and various environmental factors.

[1]Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.
[2]Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.
[3] Researcher at Agricultural central and research center ARC At Egypt
[4]Dept. of computer sciences Institute of statistical studies and researches ,Cairo University.

The infection of diseases to a plant community is determined by the host, the pathogen and the environment that can be depicted in the form of a disease triangle (Fig. 1) [Agrios, 2005]. Environmental factors have traditionally been considered to have the major impact on disease development [Agrios, 2005]. Even if a susceptible host and a virulent pathogen are present in a certain locality, a common situation when the farmer has no choice to plant the particular host, serious disease will not occur unless the environment favors its development

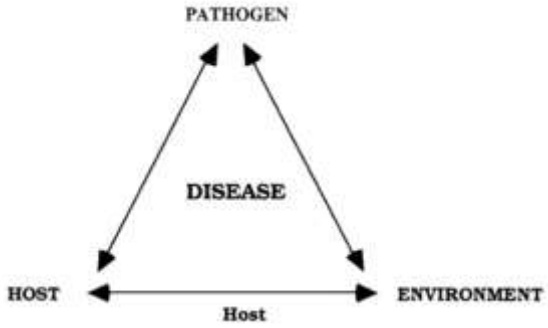


Fig.1 disease triangle

Commonly, soft computing represents a collection of methodologies that aim to exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost. Its principal constituents are fuzzy logic, neurocomputing, and probabilistic reasoning. Soft computing is likely to play an increasingly important role in many application areas, which include Fuzzy theory. The role model for soft computing is the human mind. Fuzzy set theory was initiated by Zadeh [Zadeh, 1978]. Since then, many researches and applications have been published. Fuzzy queries have appeared in the last 30 years to handle the necessity to soften the Boolean logic in database queries. A fuzzy query system is an interface to users to get information from database using natural language sentences.

In a real predict technique must deal with real spatial database. A spatial database is a collection of data concerning objects located in some reference space, which attempts to model some enterprise in the real world [Maliene et al., 2010]. The real world abounds in uncertainty, and any attempt to model aspects of the world should include some mechanism for incorporating uncertainty [Zhang and Goodchild,2002]. There may be uncertainty in the understanding of the enterprise or in the quality or meaning of the data.So that there has been a strong demand to provide approaches that deal with inaccuracy and uncertainty in geographical information systems (GIS) and their underlying spatial databases. Spatial agro-climatic database is a spatial database that contains the data of the climate variables for some places during a specific time period.

This paper introduces a proposed fuzzy prediction technique to predict the possible plant diseases which can occur in a given place as a result for climate changes. The proposed technique consists of three phases: (1) determining the diseases that can harm a given crop with its suitable environmental conditions namely temperature and humidity; (2) Defining the diseases causes membership functions. (3) Designing fuzzy-based prediction algorithm to detect if a given plantation period has a suitable environment conditions (humidity and temperature) that enables disease to appear at a specific place.

The proposed technique is mainly based on the spatial agro-climatic database. It proves that some of crops have much risk and may be lost if it plant in a critical period. This critical period is suitable for infecting by diseases in a specific area. Also, it was a suitable plantation period in the past years which not affected by climate change.

The rest of this paper is divided into four sections. The second section presents the related works. The proposed technique is addressed in the third section. The fourth section illustrates a case study. The conclusion is presented in the last section.

2-Related work

Recently, prediction of agriculture diseases becomes an important research field:
A Fuzzy Cognitive Map is a based tool for prediction of infectious diseases was proposed by E.I. Papa Georgiou [Papageorgiou et al., 2009]. The main topic of the presented effort is the representation of the cause effect relationships within medical data by the application of the soft computing technique of fuzzy cognitive maps. On the other hand, Kevin Fong-Rey Liu et al [Liu et. al, 2009] integrated the case-based and fuzzy reasoning to qualitatively predict risk in an environmental impact assessment. They use artificial intelligence and management science to assist developers: case-based reasoning (CBR) for retrieval of similar cases, fuzzy reasoning (FR) for qualitative risk forecast, and importance–performance analysis (IPA) for risk management. But not work with spatial data.

Also, the work presented in [Fernandes et al.,2011] represents a Generic Web-based Plant Disease Forecasting System SISALERT-A. It uses generic and modular simulation models for predicting diseases establishment based on weather data. They implemented a disease warning system using simulation models and the near real-time weather data acquisition system plus local specific weather forecast. On the other hand, Modular structure of web-based decision support systems for integrated pest management [Damos and Petros,2015] has been developed. SAS Pro, presents the development of a platform and mobile application that integrates disease infection models, on-field spray reports, and fungicide resistance management guidelines [Perondi and Daniel,2017].

Unfortunately, all of the above models don't represent long term prediction models. They are limited just for a small number of weeks or days, and depend on daily feed data from networks weather stations. So, they're like a diagnosis models for disease that have already existed not for prediction. So, it can't be used for making the agriculture planting maps for next year. Also, it cannot be used for modifying the planting dates for a crop or some crops after the currently changing in climate every year.

3- THE PROPOSED technique

The proposed fuzzy prediction technique for crop disease prediction considers the ongoing changes of climate data. It consists of three phases:

1. Determination of Disease for crops and preparing environmental data:
   In this phase, we determine the disease for each crop, then the causes or the suitable climate factor which is suitable or required for the infectious for that crop, the plantation period or the traditional plantation period is given, also the expected climate data of the next year is given.
2. Define a set of fuzzy membership functions phase:
   In this phase, a set of fuzzy membership functions is defined for describing the suitable climatic variables values for the disease. These functions are used to evaluate the suitability of the infectious of that disease for the crop of interest plantation.
3. Fuzzy disease detection phase:
   At this phase, we detect by fuzzy test from the climate data (plant age intervals) based on the required climatic data of the crop to check if that period is suitable disease infection or not as shown in Fig. 2.

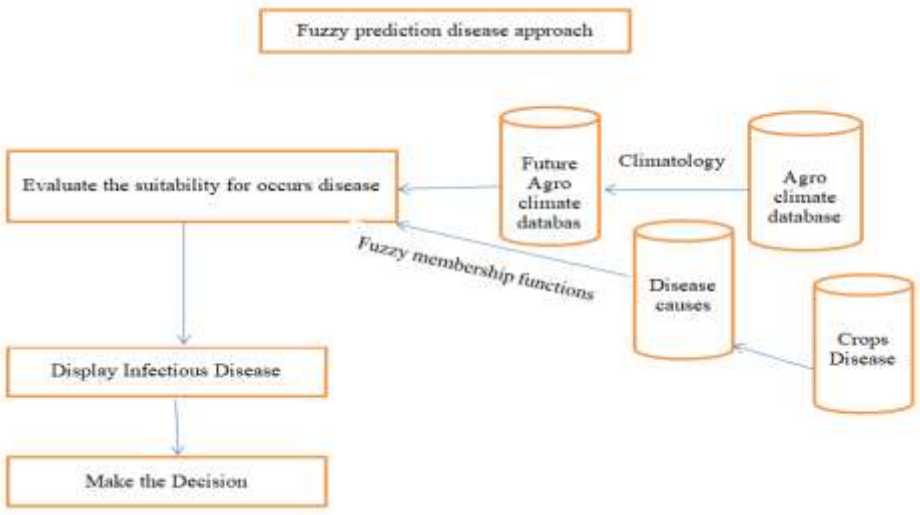


Fig.2 Fuzzy disease prediction phase

3.1-Predicting climatic data phase:

Each crop has one or more disease, which is affected by environmental data like temperature and humidity [Strange and Scott, 2005]. These diseases are determined. An important step which affects our technique accuracy is for the long term prediction how to determine the next year environmental data so that phase concentrates on it. So, this phase uses the climate data of Egypt from January 2009 to May 2018 from C to predict the next year climate data. Commonly, this phase includes the following two steps:

1. The first step is preprocessing step: it checks for missing or extreme values. If any missed value exists for any climate variable the proposed technique predicts it. It gets the values of that climate variable at the same day in a set of previous and next years. Consequently, a weighted average of such values is computed representing the missed value.

2. The second step predicts the climate data of the incoming year based on a climate prediction method called climatology which was presented in Atmos (2013).

In literature, climatology method is a simple method for making a forecast. This method involves the average weather statistics collected over many historical years to make the forecast. For example, equation (1) is used to predict a value of temperature in a specific day in the incoming year.

$$PT_d = \alpha_1 * AT_{d-1YEAR} + \alpha_2 * AT_{d-2YEAR} + \ldots\ldots\ldots + \alpha_n * AT_{d-nYEAR} \qquad \text{equation (1)}$$

Where PT_d represents the predicted value of temperature at the day of date *d;* $ATd-nyears$ denotes the actual temperature degree at the same day before *n* years; and $\alpha 1$, $\alpha 2$,..*and* αn are weighting factors such that $\alpha 1 + \alpha 2 + \cdots + \alpha n = 1$.

3.2 Defining Diseases Fuzzy Membership Functions Phase

Determines the fuzzy membership functions which define the suitable causes for the disease suitable environmental variables which are temperature or/and humidity as follow:

- humidity degrees in [b, c] are the most suitable with full matching degree of 1
- humidity degrees in [a, b[and [c, d[are partially suitable with a matching degree in [0,1[,
- And humidity degrees greater than d or less than a are not suitable at all with 0 matching degree.

Consequently, the above description of the suitability of humidity degrees for a specific disease can be easily defined as the trapezoidal fuzzy membership function depicted in equation (2).

Also the Temperature climate variable is defined as the following:

- Temperature degrees in [b, c] are the most suitable with full matching degree of 1
- Temperature degrees in [a, b[and [c, d[are partially suitable with a matching degree in [0,1[,
- And temperature degrees greater than d or less than a are not suitable at all with 0 matching degree.

Consequently, the above description of the suitability of temperature degrees for a specific disease can be easily defined as the trapezoidal fuzzy membership function which depicted in equation (2) and fig. 3.

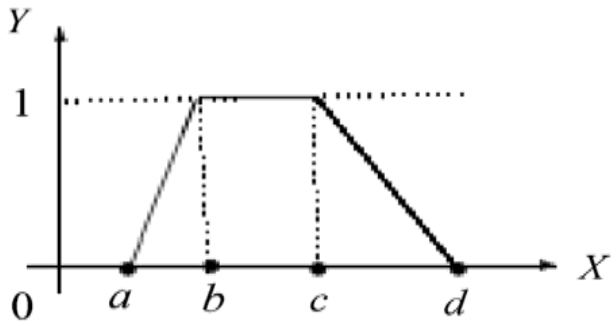


Fig.3 Trapezoidal fuzzy membership function

$$\mu_{CV\_Suitability}(C,x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x<b\\ 1 & b\le x<c\\ \frac{d-x}{d-c} & c\le x<d\\ & 0x\ge d\end{cases} \quad \text{Equation (2)}$$

Where $\mu_{CV_Suitability}$represents the suitability of climatic variable named CV with value x to harm the crop C.

3.3 Fuzzy disease prediction phase:

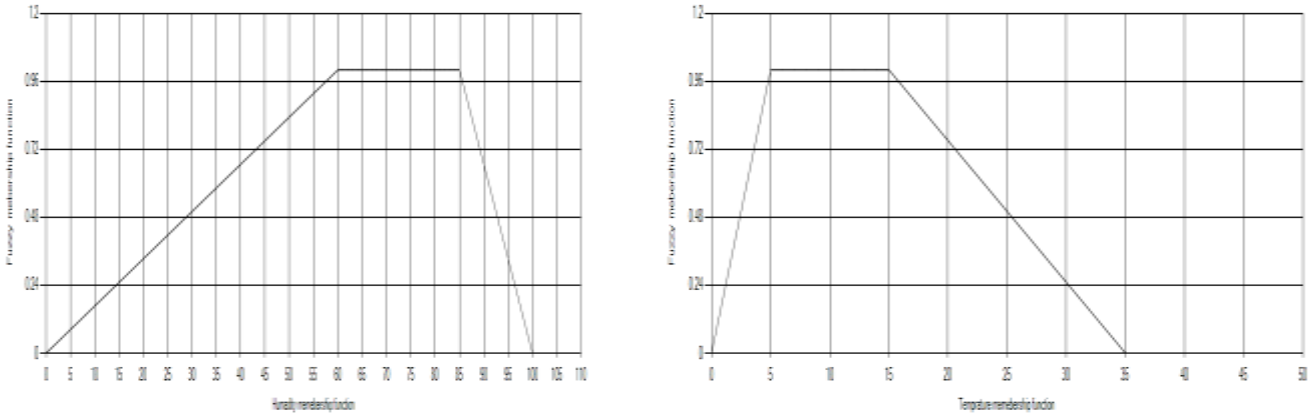
This step responsibility is to search for the days which have a suitable climate data for the infectious. So, it's like a disease detection step by the fuzzy membership functions with the climate data that have been predicted from the climatology phase one.

Detection of the possible infected disease is done using fuzzy test. Dependent on the prediction climate data (from first phase), plantation date, and plant age the proposed technique checks if this period is suitable for infecting by disease or not .

4-AN ILLUSTRATIVE CASE STUDY:

The proposed technique is illustrated to predict the disease of the Yellow rust for wheat crop in Alexandria as a governorate in Egypt. In Alexandria, wheat crop is cultivated at the first of November to the end of April so it needs 180 days before harvesting. The knowledge for the important infectious diseases for wheat crop was extracted from [Rossi, 2001].The following is the suitable climate variables of the disease of the Yellow rust

- Most suitable temperature from 5-15
- Most suitable humidity 60%-85%
- Most suitable host(wheat) age for infectious between (90-150 days)



(a) Suitable Humidity membership function *(b) Suitable Temperature membership function*

Fig. 4 Yellow rust suitable climatic requirements

In phase one, the climate preparation next year data which extract the predicted environment data for the next years **2017 and 2018** from **2009 to 2016** is used to get Then we are going to compare the result of our technique with the real data which has been taking from wheat disease department at the Plant Pathology Research Institute which belongs to the Agriculture Research Center in Egypt for the last year 2017-2018 which obtain the Yellow rust infected area at Alexandria. Fig. 6 shows the fuzzy membership function for predicted Yellow rust infected.

Results:

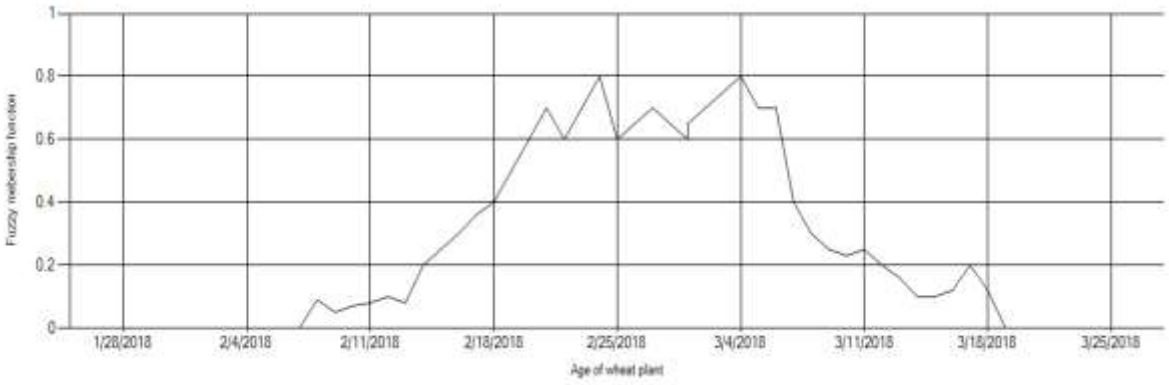


Fig.6 the predicted Yellow rust Infection degrees within the specified period

Table 1. shows Yellow rust infected area and its dates at Alexandria for the last year 2017-2018, and fig 7 shows its fuzzy membership function.

Day February	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Infected cases	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	**1**	**4**
Day February	17	18	19	20	21	22	23	24	25	26	27	28	29			
Infected cases	**4**	**5**	**5**	**6**	**6**	**6**	**6**	**7**	**6**	**5**	**5**	**5**	**5**			
Day March	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Infected cases	**5**	**6**	**6**	**6**	**6**	**6**	**4**	**3**	**2**	**1**	**1**	**1**	**1**	0	0	0
Day March	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
Infected cases	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Table.1: Real Yellow Rust infectious of wheat in Alexandria at 2017-2018

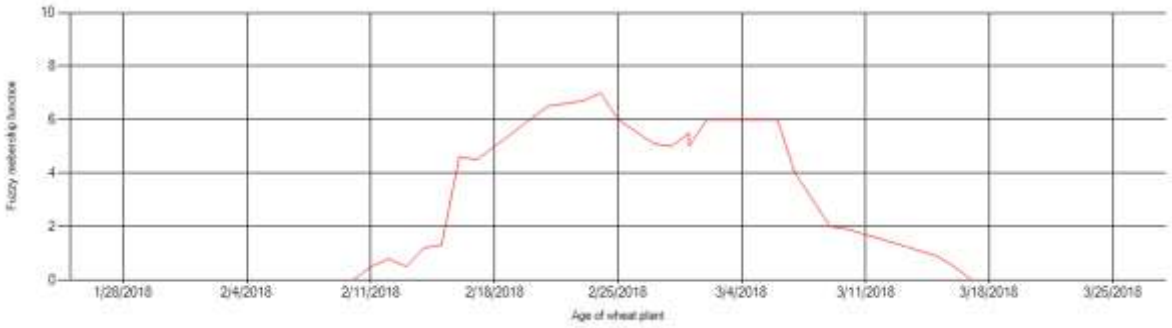


Fig7: predicted Yellow rust infectious using the proposed technique at Alexandria in 2017-2018

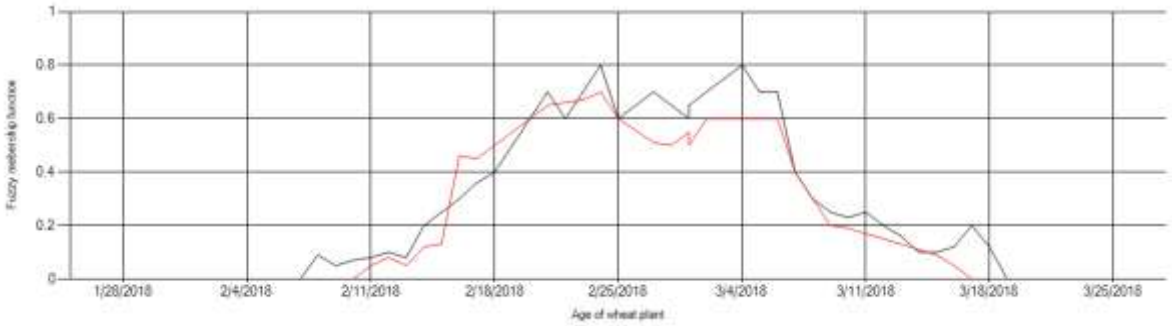


Fig8: predicted technique vs the real yellow rust infections at Alexandria in 2017-2018

As shown in Fig. 8 Yellow Rust maximum appear at first of the February then decrease to end of March end, so at that time we should take care and be ready to test every day for disease occurs and use chemical protection if it's possible or other was for protection, As we see how the technique works and display for the decision maker to predict at which periods can be infectious occurs for a crop and that illustrated at table.2 below:

Table.2: Recorded infectious cases places to Alexandria

	Start Rust yellow infectious	End Rust yellow infectious
Real infected period at Alexandria 2017-2018	10th of February 2018	17th of March 2018
Proposed model	9th of February 2018	19th March 2018
Comment	*The domain of the proposed model is wider , and it's positive to include the whole infected area at real*	

***TABLE3. Comparative analysis for the proposed** technique **against other agriculture prediction disease** techniques*

technique	Inputs	Outputs
SISALERT[8]	hourly weather station data	Short term prediction technique which can predict one disease in three days at maximum.
SAS Pro[11] Mobile tool for just strawberry disease	Plant images	This technique near to be as a diagnosis system not for a prediction technique it developed as just mobile tool for just the strawberry's disease
The proposed technique	Climate data –disease suitable variables	Display the future possibility disease infectious periods for the next year of a particular crop in a particular area, which enable the decision makers to select which time to plant that crop or be ready for treatment.

5-CONCLUSIONS AND FUTURE WORK

This paper presented a proposed fuzzy prediction technique for predicting the crops disease with climate changes, based on the available historical spatial agro-climatic database and climate requirements like temperature degrees and humidity, applying the proposed technique leads to discover when wheat rust diseases can be occurred with a suitability degree in Alexandria. Also, it discovers that some traditional periods become not suitable for planting wheat in Alexandria at all. Accordingly, the proposed technique guides and helps agriculture investors in a flexible manner to be disease avoided or early manage and control its spread. In consequent, such technique greatly helps decision makers in making agricultural maps and the distribution of the crops planting along the year. On the other hand, it directly increases the profit and decreases the cost of any crop plantation.

Consequently, in the future some improvements will be added to the proposed technique like adding other climate variables like rainfall, wind speed and dew point while calculating the suitability degree of each period with infectious disease.

REFERENCES

Agrios,G. N.(2005).Plant Pathology(5th edition).Elsevier-Academic Press.San Diego, CA.

Andreis and José Henrique Debastiani. **(**2018**).** "SAS Pro: an integrated mobile tool for strawberry disease control and multifungicide resistance strategies." .

Damos and Petros. **(**2015**).** "Modular structure of web-based decision support systems for integrated pest management. A review." Agronomy for sustainable development 35.4 : 1347-1372.

Fernandes, José Maurício Cunha, Willingthon Pavan, and Rosa Maria Sanhueza. (2011) "SISALERT-A Generic Web-based Plant Disease Forecasting System." HAICTA.

Liu, K.F.-R., Yu, C.-W., (2009). Integrating case-based and fuzzy reasoning to qualitatively predict risk in an environmental impact assessment review. Environmental Modelling & Software 24 (10), 1241e1251.

Maliene V, Grigonis V, Palevičius V and Griffiths S .(2010)."Geographic information system: Old principles with new capabilities". Urban Design International 16 (1). pp. 1–6.doi:10.1057/udi.2010.25.

Papageorgiou, N.I. Papandrianos, G. Karagianni, G. Kyriazopoulos and D. Sfyras, (2009) "A fuzzy cognitive map based tool for prediction of infectious diseases", Proceeding of FUZZ-IEEE 2009, World Congress, 24-27 August 2009b, Korea, pp. 2094-2099.

Perondi and Daniel (2017). "AgroClimate Smart Crop Season: uma abordagem na simplificação de modelos de simulação para sistemas de auxílio à tomada de decisão." .

Rossi John (2001)."Influence of temperature and humidity on the infection of wheat spikes by some fungi causing Fusarium head blight." Journal of Plant Pathology:189-198.

Strange, R. N. and Scott, P. R. (2005). Plant Disease: A Threat to Global Food Security. Ann. Rev. Phytopathol. 43:83-116

Lotfi Zadeh , (1965) ."Fuzzy sets", Information and Control 8, 338-353

Zhang, J, and Goodchild,M,(2002). "Uncertainty in Geographical Information", Taylor & Francis, London .

A Proposed Approach for IoT Energy Consumption using Fuzzy Logic

Noha Elqeblawy[1] Imane Fahmy[2] Ammar Mohammed[3] Hesham Hefny[4]

[1,2,3,4] Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Abstract

Internet of Things (IoT) is a new arising technology that aims to connect billions of devices everywhere and anytime around the world over the internet. These devices collect and share data among each other to complete tasks automatically without human intervention. However, one of the primary challenges facing IoT is energy resources reservation and efficient consumption, especially in those types of applications that require nonstop monitoring or suffer from lacking ongoing energy resources. This paper proposes a fuzzy logic based approach for temperature and humidity monitoring systems to reduce energy consumption of packets exchange between these systems and the control room's servers in IoT environment. The experimental results on Melbourne dataset shows that the proposed fuzzy-based approach successfully reduces the energy consumption compared to the traditional approach of temperature and humidity monitoring systems.

Keywords: IoT , Fuzzy Logic, Temperature and humidity monitoring systems, Energy reservation, Fuzzy based monitoring, environmental monitoring.

1. Introduction

The scarcity of energy resources is a worldwide problem that raises day after day. Therefore, energy reservation and efficient energy consumption are increasingly required in many applications and technologies as highlited by Suganthi et al. (2015).

One of the emerging technologies nowadays is the internet of Things (IoT) that enables the distribution of smart objects everywhere around us. These smart objects are able to communicate and share information with each other and with their corresponding servers over the internet to carry out useful tasks without any human involvement referred by Whitmore et al. (2015).

Energy reservation could be considered as a highly significant challenge facing IoT not only for the global energy problem but also due to the lack of energy resources in several IoT applications as

mentioned by Shah et al. (2016) and especially for those applications that are applied in critical nonstop monitoring and control tasks, or applications in difficult reachable environment as described by Gazis et al. (2015). Such environments may suffer of lack of ongoing energy resources that should meet applications requirements. Therefore, any interruption of service for recharging or changing its energy resources can obstruct its main function and leads to major problems.

Take for example IoT health monitoring system that monitors vital signs of a patient up to date and alerts doctors in the case of abnormal conditions to take proper procedures. hence, any interruption of this application due to energy resources running out may leads to side effects on the patient's health. Additionally, an IoT temperature monitoring system located inside a huge machine in a factory, if interrupted due to energy resources running out, that may result in stopping the whole machine for maintenance works. Consequentially, it can affect the production process dramatically. Moreover, if that machine was meant to function for a vital constantly running process, e.g. electricity generation or water purification. Thus, any unplanned interruption for these machines may cause damage, service failure, significant money loss, or even disasters. Another example of a critical monitoring system of a natural disaster like an earthquake or a volcano. if a disaster occurs, then all recorded data before and during the disaster, located in the surrounding area, must be transmitted properly on time in order to take actions as alert, evacuation or even search and rescue operations as referred by Khatibi & Khatibi (2018). Therefore, these critical data transmission should not be affected by any lack of energy resources of IoT systems and the consumption of energy have to be handled properly during transmission of the data.

In fact, data and control packets periodical and frequent exchange between monitoring systems and their correspondent servers in IoT application could be considered as one of the most energy consuming tasks for these IoT applications. The most commonly used type of monitoring systems are simply designed to keep transmitting the recorded data captured by their sensors to the controlling servers periodically every predefined time interval. When these captured data reach the server, they could be analyzed and used for prediction/forecast, research or statistics. Take for example the temperature and humidity monitoring systems. Usually this IoT monitoring system contains sensors that sense the temperature and humidity degrees of the system's environment periodically. These sensors are connected to a microcontroller that may transmit the recorded temperature and humidity degrees directly over the internet to their controlling server as appeared

by Karim et al. (2018) or may interpret the recorded degrees locally and acts according to the interpretation result, e.g., speed up the local fan or turn off the air condition as shown by Purwanto et al. (2018). However, the problem of the traditional IoT temperature and humidity monitoring systems, is that systems transmit each captured temperature or humidity degree directly to their controlling servers whether that degrees are critical to be sent and stored or not , moreover they may transmit the same uncritical degrees periodically multiple times even if it remains the same for the indicated time interval. Hence, this traditional monitoring system's data transmission approach results in fast and useless consumption of the system's energy resources. Moreover, the frequent and redundant data received by the controlling server will consume a large amount of the server's storage resources.

To this end, the main contribution of this paper is to propose an approach that decreases energy consumption in IoT temperature and humidity monitoring systems using fuzzy logic. The fuzzy based inference system is used to decide if the captured temperature and humidity degrees should be transmitted to the server or not instead of transmitting all recorded degrees periodically by the traditional monitoring systems approach. To evaluate the performance of the proposed fuzzy-based approach, the experimental result is carried out on City of Melbourne (2017).

This paper is organized as follows: Section 2 gives a brief background about Internet of Things (IoT) and fuzzy logic. Section3 discusses related work. Section 4 presents the proposed fuzzy-based approach. Section 5 discusses the experiment and its results. Finally, section 6 conclude the paper.

2. Background

This section provides a brief background about the state-of-art of Internet of Things (IoT) and a brief background about fuzzy logic.

2.1. Internet of Things (IoT)

Kevin Ashton (ex-cofounder and executive director of the Auto-ID Center) released his article by Ashton (2009) about the new topic at that time the internet of things which became one of the main trends in the computer science these days. It is the idea of distributing smart objects everywhere around us. The core concept of IoT is everyday objects should be smart by equipping them with identifying, sensing, networking and processing capabilities that let them communicate

with each other and with services over the internet to do useful tasks without any human involvement as highlighted by Whitmore et al. (2015). Furthermore the huge quantity of data that could be collected from these objects and be processed, can help to make better decisions for human life as described by Gubbi et al. (2013).

Although there are around 30 billion IoT connected devices and the number of connected things will rise to 50 billion by 2020 referred by Shahrestani (2017), Still there's no global definition for IoT. The main reason for that confusion, is the fact that IoT combines two terms: Internet which indicates network oriented vision of IoT and things which refers to generic objects to be integrated the IoT environment as discussed by Bandyopadhyay & Sen (2011).

Obviously the number of IoT applications increases year after year in all fields and, in the future, people will be surrounded by smart applications everywhere that will make their life much easier from smarter homes and offices to smarter transportation systems, hospitals, factories and cities. For example, the smart cities will contain smart grids that collect data about the energy consumption and make these data available on the internet to help generating reports about the pattern of use and the recommendations for reducing it as explained by [Fu et al. (2011) , Whitmore et al. (2015)]. Also in the healthcare field, the patient will be covered by sensors for monitoring his vital functions, heart beat count and blood pressure. Then, the collected data will be available online for doctors , patient's family and other people involved in treatment provision as demonstrated by Dohr et al. (2010). Moreover, for the transportation systems, the supply chain and logistics efficiency will be improved by IoT using technologies such as sensors and Radio-frequency Identification (RFID) that can collect data about the merchandises life cycle. Then, putting the collected data online, up to date for the customers notification and tracking. Furthermore, in social applications, using IoT devices to collect data about people activities and locations can save people's time and inform them about social events or activities that may be interested for them as pointed by Guo et al. (2011).

Actually, like any new idea or concept IoT faces a lot of challenges. These challenges could be classified to four main types: interoperability challenges, data challenges, hardware and network challenges and security challenges.

Firstly, The interoperability challenges include technological interoperability and semantic interoperability. The technological interoperability seeks to find proper ways for different devices

interactivity with each other and with the people seamlessly. On the other hand, semantic interoperability focuses on how shared data could be understood and interpreted by different devices properly as described by Gazis et al. (2015).

Secondly, data challenges comprises the data management challenge and the data mining challenge. The data management challenge involves how to store and process the huge collected data from different IoT domains that have different types and how to improve the data centers to be suitable for this mission, while data mining challenge deals with how to process and analyze huge stream of data from various digital sensors by improving data mining tools to get useful information on time as highlighted by Lee & Lee (2015).

Thirdly, the hardware and network challenges include the smart things and the architecture challenge. The smart things aims to find ultra-low power circuits and components of IoT systems that need to be developed properly to work in rough environments that have a difficult or impossible reach-ability for system's maintenance duties or batteries recharge regularly. The architecture challenge is mainly concerned with the Wireless Sensors Networks (WSN) perspective. In fact, the European Union projects of SENSEI and Internet of Things Architecture (IoT-A) have addressed the challenges particularly from the WSN perspective, as mentioned by Gubbi et al. (2013). Additionally, they discussed and how to improve WSN to be convenient to all IoT applications.

Finally, The security challenge such as the authorization/ access control challenge and the privacy preservation challenge. Additionally, the authorization and access control challenge are concerned with development and management of authorization and access rights mechanisms to different IoT resources. Also, they attempt to guarantee a safe access to these resources according to access verification that is considered as another challenge in heterogeneous IoT networks. Whereas, the privacy challenge is more concerned with keeping the individual's personal data such as: location data or health data private for her/his safety as proposed by Conti et al. (2018). Furthermore, data ownership, legal and liability issues have to be addressed accordingly as described by Gazis et al. (2015).

Indeed, beside considering the development of IoT implementation and applications, there are also some technology issues and challenges to be improved alongside such as RFID, WSN and the power supply techniques. Essentially, the RFID is mainly used to identify things automatically

and capture their data using radio waves, a tag and reader. While WSNs are initially used to sense, capture and monitor physical and environmental conditions around the clock and send the captured data to the controller or the data centers as described by Lee & Lee (2015).

Obviously, the power supply technology is a critical factor for IoT development. Most of the IoT devices employ limited lifetime weight and size batteries for powering. These devices can be in hash or remote environments such as: under oceans water or inside a building concrete which make it impossible to change their batteries frequently as mentioned earlier in the smart things challenges.

2.2. Fuzzy Logic

Fuzzy logic is a contemporary approach used to deal with common sense problems that can't be solved by classical logic. Common sense can be defined as the ability of human to take quickly decisions without computations even if these decisions are not the ideal every time.

Fuzzy logic introduced by Lotfy zadah , a professor at the University of California at Berkley, in 1965 as appeared by Zadeh (1965). Afterwards, many researches concerned applying fuzzy logic in different fields. Basically, fuzzy logic is used to solve problems with imprecise facts that could be partially true and partially false in the same time as explained by Ross (2005). The imprecise facts in fuzzy logic are represented by the values within the interval [0,1]. In effect, problems that could be solved by classical logic deal with precise facts either totally false or totally true; either 0 or 1. However, using imprecise facts in rule-based fuzzy logic systems, help in increasing the level of abstraction in problems representation. Consequently these imprecise problems could be solve by less and more powerful rules in shorter time.

According to fuzzy logic, the fuzzy set is defined as a set with no well-defined boundaries where transition between membership and nonmember ship is gradual rather than sudden transition as clarified by Dubois (1980). In other words fuzzy set a special type of sets that gives each input a gradual degree of membership. Moreover, the fuzzy set is defined by a membership function. Various fuzzy sets are representing linguistic concepts/values such as: cold, cool or worm and so on. Those linguistic concepts/linguistic values are used to define the states of a variable.

Such a variable is usually called a fuzzy variable/linguistic variable such as the temperature as referred by Dubois (1980). While the universe of discourse is defined as the domain of the fuzzy variable.

Generally, the architecture of fuzzy logic systems consist of four main parts as follow:

• A fuzzifier.

• Fuzzy Rules.

• Fuzzy Inference System.

• Defuzzifier.

Primarily, the fuzzifier is responsible for the fuzzification process. The fuzzification is the process of converting the input crisp values to fuzzy linguistic values with membership degrees using a membership function. According to fuzzy logic concepts, the input can belong to more than one fuzzy set in the same time with different membership degrees. Then the outputs of the fuzzification process, the fuzzified inputs, are inferred by the fuzzy inference engine according to a set of predefined if-then rules and produce the output fuzzy variables. Finally, the defuzzifier is responsible for the defuzzification process. The defuzzification is the process of converting the fuzzy output variable to quantifiable crisp values . These crisp values are used to trigger events or actions according to the system specifications as described by[Bhunia et al. (2014), Singhala et al. (2014)].

3. Related Work

Lots of recent researchers worked on integrating fuzzy logic with IoT monitoring applications in many fields to improve their performance and efficiency. For instance in healthcare monitoring field, Santamaria et al. (2016) presented a two stage fuzzy logic approach. That approach depends on collecting data about user then send it to the cloud. Thereafter, this data is used by the fuzzy logic to determine the current activity of the user and recognizing the abnormal cases. Also Kaur & Kaur (2017) used fuzzy logic in industry field to evaluate and make a decision about the employees performance based on the data that recorded by IoT monitoring devices. Similar to the research of Chan et al. (2018) within the fashion industry field that introduced a fuzzy logic approach to predict purchasing intentions of the customer by analyzing their actions and behaviors in stores. Moreover, in the environmental monitoring field like fire detection, Maksimovic et al. (2015) proposed two fuzzy logic approaches to reduce and optimize the number of rules that should be checked for making the correct decisions in monitoring and early detecting fire systems. And Listyorini & Rahim (2018) used a fuzzy logic for minimize the wrong warnings

from IoT fire detected devices. Hanse, in security field Mahalle et al. (2013) applied fuzzy logic for dynamic access control in IoT for trust calculations deals with the linguistic information of devices . Also Chen et al. (2011) introduced a trust and reputation model; TRM-IoT, based on fuzzy logic that helps in developing the WSNs security and performance. In despite of the using fuzzy logic to improve all works of [Chan et al. (2018), Chen et al. (2011) , Kaur & Kaur (2017), Listyorini & Rahim (2018), Mahalle et al. (2013), Maksimovic et al. (2015), Santamaria et al. (2016)] They all did not concerned about energy reservation and saving energy.

Similar to this work, several works used fuzzy logic for saving energy proposes. For example in smart vehicles domain, Cueva-Fernandez et al. (2015) , tried to save resources of the smart vehicle such as power and communication resources using fuzzy logic to choose the best moment for transmitting the data to control room's server depending on some parameters such as vehicle position. Also in healthcare monitoring domain , Bhunia et al. (2014) introduced iHealth: A Fuzzy approach for provisioning Intelligent Health-care system in a smart city. This system used the fuzzy logic to take the decision of alerting doctors if there is a patient with fatal conditions based on some of his basic health parameters which recorded by several IoT enabled sensors that attached to his body. This approach could to preserve a lot of energy resources. Furthermore, in smart houses domain, Collotta & Pau (2015), employed fuzzy logic to reduce the energy consumption in Bluetooth Low Energy (BLE) used in IoT applications by determining the sleeping time of field devices in smart homes. However, [Bhunia et al. (2014), Collotta & Pau (2015), Cueva-Fernandez et al. (2015)] were concerned by consuming energy using fuzzy logic, they applied in different domain than this work domain. On the other hand, Meana-Llorián et al. (2017) introduced IoFClime system, in this work domain, that is a temperature management system of a specific area that automatically gets the outdoor apparent temperature and humidity from external IoT platforms and senses the indoor temperature Then, they use the fuzzy logic to decide the best moment to heat up or to cool down the atmosphere based on that collected data in order to maintain a convenient temperature. Therefore, this temperature management system helps to save about 40% of energy and thus, saves money and electricity. On the contrary of the proposed fuzzy-based approach in this work, that temperature monitoring system imports the outdoor temperature and humidity degrees from outside the IoT platform, while the new proposed approach exports the indoor temperature and humidity degrees to outside server or platform.

Many techniques and approaches have been presented to enhance the energy reservation and saving in IoT applications. For instance Khatibi & Khatibi (2018) presented an approach to ameliorate the energy consumption of processing queries in IoT using hierarchical clustering index tree. The provided approach assumed there is already a huge amount of nodes in the system. However, that approach will not be efficient if it is applied for a small system. Furthermore Baccelli et al. (2013) presented a new operating system called: "RIOT OS". That RIOT works on facing IoT challenges such as limited power supply by avoiding the redundant development. But applying this new OS to the existing IoT applications may be relatively expensive.

Also the approximate computing is one of the techniques that used for power-/energy-efficient systems. this technique does not depend on the idea of precise computing between the specification and implementation at different layers of the hardware-software to enable energy efficiency. Thus, by utilizing energy resources efficiently, the system's performance efficiency increases but some output quality level could be noticeably affected as described by Shafique et al. (2016). Although, some applications can handle that quality reduction with tolerance like multimedia, wireless communications, artificial intelligence and data mining as mentioned by Gao et al. (2017). But still there are some other applications can't handle or tolerate some error occurrence especially in healthcare monitoring systems.

Moreover , Jayakumar et al. (2016) mentioned another approach for saving the energy consumption in IoT using the recent MicroControllers (MCUs). That approach depends on saving idle energy consumption by supplying two types of sleeping modes: shallow sleep mode that produces fast wake up but saves the state by keeping in on-chip memory some elements that consume significant power and deep sleep mode that provides longer time for wakening-up and does not keep the state that results near zero power consumption but needs the (MCUs) to checkpoint the current state of the system on the on-chip non-volatile memory (typically flash memory). Consequently, this results in significant power consumption for barely write / erase operations. Although using sleep modes, power consumption could be decreased relative to active mode, the monitoring systems that use these sleep modes may still encounter problems and issues of unreliable power supply and sudden power interruption due to drained rechargeable batteries. Hence , some researchers suggested the idea of using low power wireless techniques/ protocols for sensors networks in IoT applications. Some of these protocols can be used for short scope of

connectivity between sensor nodes and others for long connectivity. Mahmoud & Mohamed (2016) presented a comparative study between different protocols and their impact on power consumption and sensor networks lifetime. However, the short scope of connectivity disadvantages appear in case of development the application to cover larger scope of connectivity that necessary to rebuild the system's infrastructure. While the long scope of connectivity disadvantage is that it can be used for transferring data with low rate only.

Furthermore, John et al. (2017) introduced Cluster Head (CH) selection schemes for using energy efficiently and extend the sensor nodes lifetime based on the idea of selecting the node with the higher energy to be the cluster's head. Although, CH tends to extend the network's lifetime, this method is more suitable for the large WSNs with large number of nodes only. Moreover, Kamalinejad et al. (2015) tried to use harvesting energy technique for e recharging the batteries and hence the whole network's lifetime in WSN IoT systems. Thus, they attempted to harvest the energy from surrounding environment such as: wireless Radio-Frequency (RF). However, utilizing energy harvesting is not granted since it depends on the existence of energy source

4. Proposed Approach Using Fuzzy Logic

Essentially, the proposed fuzzy-based approach aims to reduce the consuming of energy in IoT temperature and humidity monitoring system. The key idea of the energy reduction is reducing the energy consumed by packets exchange between the temperature and humidity monitoring systems and the server control rooms using fuzzy logic. Actually, fuzzy logic is used as a controller to decide if the captured degrees should be transmitted to the server or not based on predefined threshold.

4.1. Proposed approach

The proposed approach uses mamdani fuzzy logic inference system, that is extensively used to extract the experts knowledge. It helps us to clarify the expertise in more natural way, more human-like manner as mentioned by Kaur & Kaur (2012).

As appears in figure 1, The key point of the proposed approach using the temperature and humidity degrees that captured by sensors as inputs for the fuzzy logic system and producing the

defuzzification output, to determine if the temperature and humidity reads should be send to their corresponding server or not based on predefined output threshold.

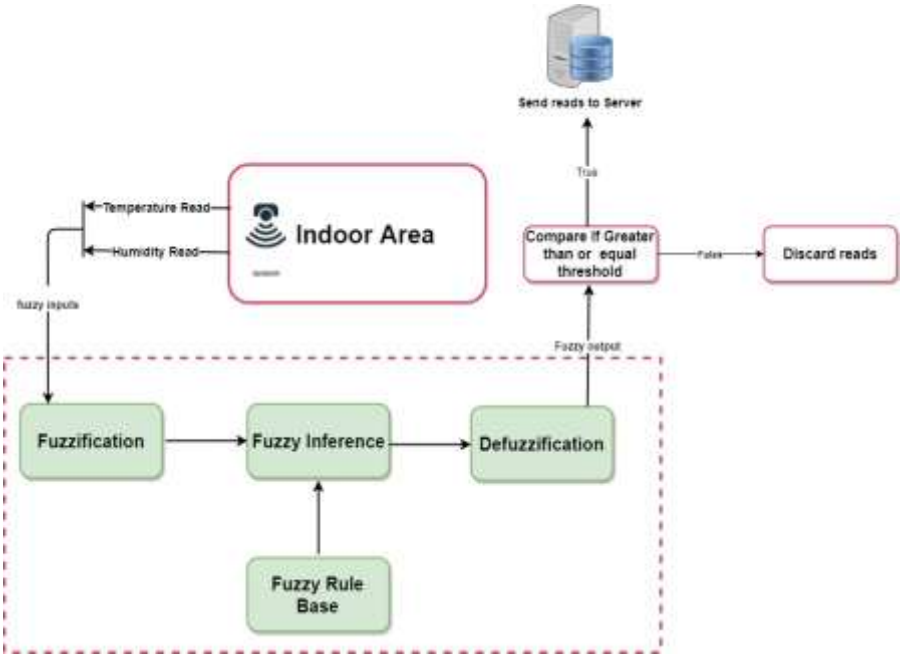


Figure 1: Proposed approach illustration

Firstly The crisp values that are received from temperature sensor are fuzzified to fuzzy values low, medium, high and v.high. Also the crisp values that are received humidity sensor are fuzzified to fuzzy values dry, comfortable, humid and stiki. Next, the fuzzy (if-then) rules that are shown in table 1, are applied to the temperature and humidity fuzzifed values to get the fuzzy output values cool, medium, warm and hot. Then the defuzzification process is applied to produce the output crisp value. Finally the output crisp value is compared to the predefined threshold value, if the output crisp value equal or greater than that threshold value the temperature and humidity crisp values\sensors reads are sent to the corresponding server else that crisp values are discard.

Table 1: Fuzzy (if-then) rules

Rule	Temperature	Humidity	Output
1	Low	Dry	Cool
2	Low	Comfortable	Cool
3	Low	Humid	Cool
4	Low	Stiki	Medium
5	Medium	Dry	Cool
6	Medium	Comfortable	Cool
7	Medium	Humid	Medium
8	Medium	Stiki	Warm
9	High	Dry	Medium
10	High	Comfortable	Warm

11	High	Humid	Hot
12	High	Stiki	Hot
13	V.High	Dry	Medium
14	V.High	Comfortable	Hot
15	V.High	Humid	Hot
16	V.High	Stiki	Hot

5. Experiment & Results

In this section the proposed fuzzy-based approach is evaluated against the traditional approach in IoT temperature and humidity monitoring systems by applying the two approaches to a dataset that contains temperature and humidity sensors reads then compare the amount of energy that is supposed to be consumed due to transmit the temperature and humidity reads using both of approaches.

5.1. Dataset

The dataset used for the experiment implementation, was provided by the City of Melbourne, Australia, May 10, 2017 and contains sensor readings, with temperature, light, humidity every 10 minutes at 8 locations and downloaded from City of Melbourne (2017). Temp_avg and humidity_avg data of sensor 510 were chosen for applying the experiment.

5.2. Energy Consumption Equations:

For calculating the amount of energy that supposed to be consumed during the transmission of data packets process , equations 1 and 2 as used by Fahmy et al. (2014). Equation 1 calculates E(p) as the energy required to transmit a packet p.

$$E(P) = i * v * t_p \quad (1)$$

Where i represents the current consumption equal 280mA in transmit mode. v is the used voltage. IEEE 802.11g network interface card (NICs) assumed to be used for the experiment which uses 5 volt for voltage as highlighted by(Feeney and Nilsson 2001). Then, t_p is the required time in seconds to transmit a packet given by equation 2.

$$t_{p=}\left(\frac{p_h}{6*10^6}+\frac{p_d}{54*10^6}\right) \quad (2)$$

Where p_h is the packet header size in bits. And p_d the packet data size in bits.

5.3. Fuzzy Sets and Threshold

For applying the proposed fuzzy-based approach, the universe of discourse of the temperature degrees is clustered into four fuzzy sets named low, medium ,high and v.high. Trapezoid membership functions are defined for low and v.high fuzzy sets and triangle membership function for medium and high fuzzy sets. The membership functions that define the temperature fuzzy have been designed as highlighted by figure (2).

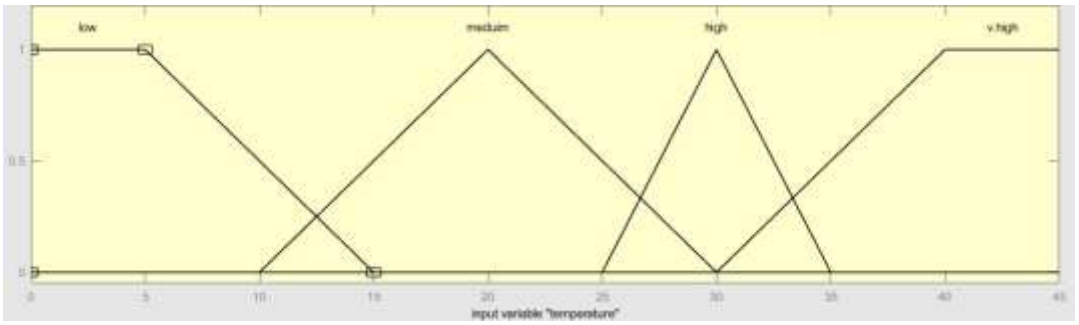


Figure 2: Membership functions Sets for Temperature Degrees

The universe of discourse of the humidity degrees is clustered into four fuzzy sets named dry, comfortable, humid and stiki. Trapezoid membership functions are defined for dry and stiki fuzzy sets and triangle membership function for medium and high fuzzy sets. The membership functions that define the humidity fuzzy sets have been designed as appeared in figure (3).

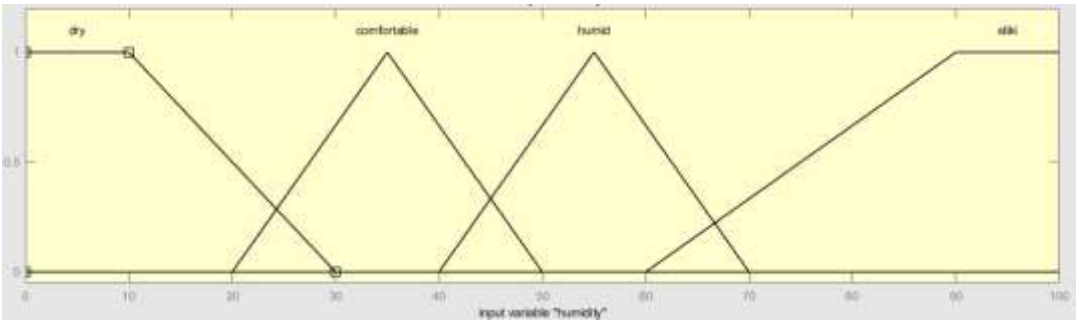


Figure3: Membership functions Sets for Humidity Degrees

The universe of discourse of the output is clustered into four fuzzy sets named cool, medium, warm and hot . Trapezoid membership functions are defined for cool and hot fuzzy sets and triangle membership function for medium and warm fuzzy sets The membership functions that define the output fuzzy sets have been designed as shown in figure (4).

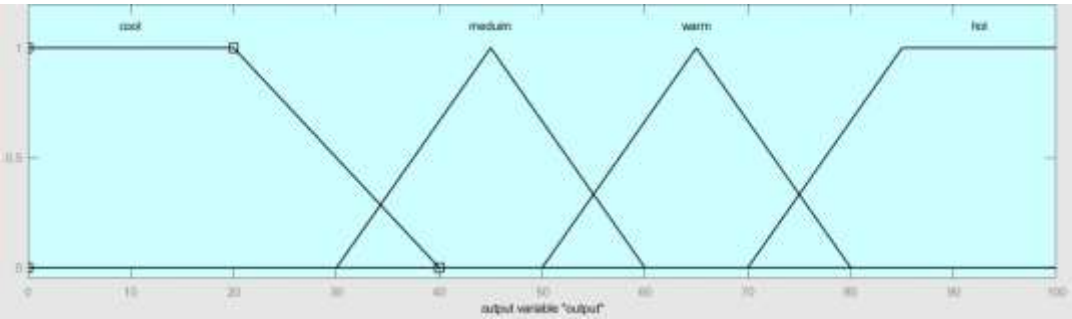
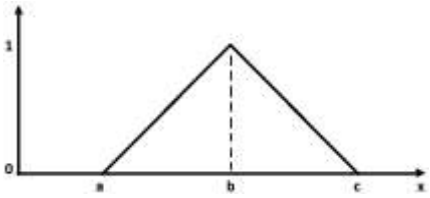


Figure 4 Membership functions sets for output Degrees

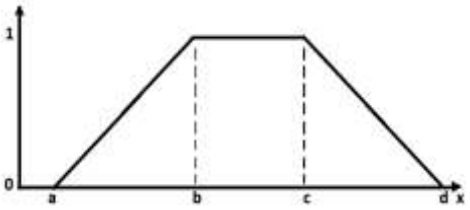
The triangle membership functions could be represented as shown by equation (3).

$$Triangler(x: a,b,c) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ \frac{c-x}{c-b} & b \le x \le c \\ 0 & x > c \end{cases} = \max(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0) \quad (3)$$



While the trapezoid membership functions could be represented as illustrated by equation(4).

$$Trapzoid\ (x: a,b,c,d) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & b \le x \le c \\ \frac{d-x}{d-c} & c \le x \le d \\ 0 & x > d \end{cases} = \max(\min\left(\frac{x-a}{b-a}, \frac{d-x}{d-c}\right), 0) \quad (4)$$



Actually, trapezoid and triangular shapes of membership functions are chosen because they are suitable for realtime operation (low communication complexity joined with enough accuracy) as discussed by[Maksimovic et al. (2015), Kaur & Kaur (2017)].

6. Results

By repeating the experiment mutable times using Melbourne dataset , the proposed fuzzy logic system and the evaluation equalization with different threshold degrees. We found out that, increasing the threshold value leads to decrease the energy consumption in the same time may leads to loss important information and reduce the quality of data. In other words there is tradeoff between the energy reduction and the quality of data.

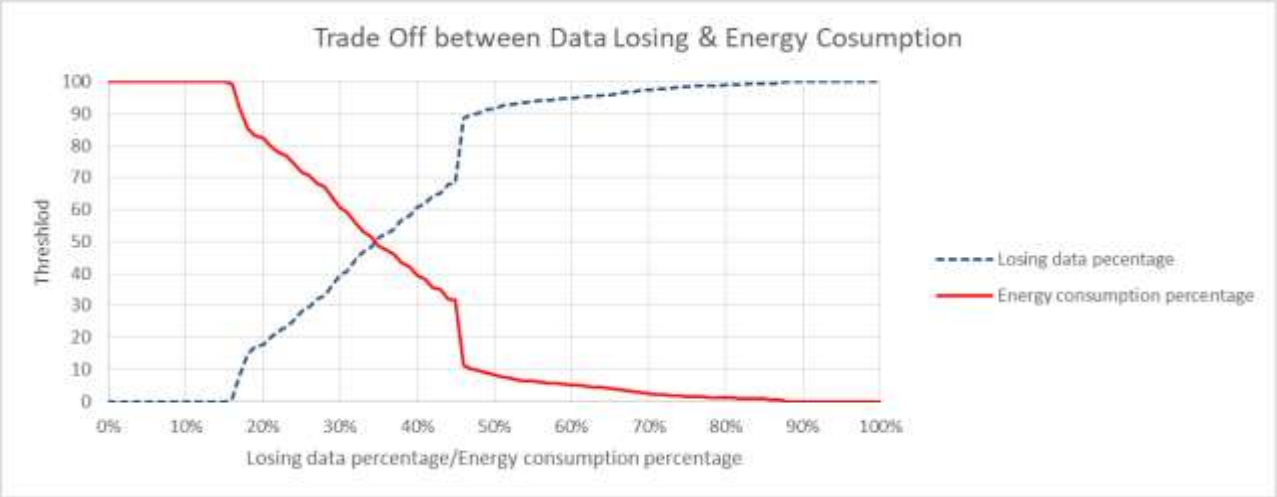


Figure 5: Trade off between Data losing and energy consumption

Figure5 shows the tradeoff between the percentage of data losing and the percentage of energy consumption at various threshold values. The dashed carve represents the percentage of data losing using the proposed approach to using the traditional approach that calculated by equation (5). The solid carve represents the percentage of energy consumption using the proposed approach to using the traditional approach that calculated by equation (6).

$$Losing\ data\ percentage =$$

$$100 - \left(\frac{total\ number\ of\ \ tranmitted\ records\ using\ the\ proposed\ approach\ at\ y\ threshold}{total\ number\ of\ tranmitted\ records\ using\ the\ traditional\ approach} \times 100\right) (5)$$

$$Energy\ consumption\ percentage =$$

$$= 100 - \left(\frac{amount\ of\ energy\ consumption\ using\ the\ proposed\ approach\ at\ y\ threshold}{amount\ of\ energy\ consumption\ using\ using\ the\ traditional\ approach} \times 100\right) (6)$$

So choosing the threshold must be fair enough for saving the tradeoff between losing the quality of data and the energy reduction.

By choosing 30 degree as an output threshold for the experiment. The temperature and humidity reads are sent to the server if their output crisp value equals or exceed 30 degree. Because we concern about the temperature and humidity these belong to medium, warm or hot output fuzzy sets. The comparison shows that, the percentage of reduction of the energy that is consumed during recorded temperature and humidity degrees exchange between the monitoring system and the server using the proposed fuzzy-based approach at threshold 30 degree to the energy that is consumed using the traditional approach was nearly 39.4%. The following table (2) represents a detailed comparison between the total number of data packets transmissions and the amount of energy consumption using the traditional IoT approach and the proposed fuzzy-based IoT approach at threshold 30 degree.

Table 2: Proposed fuzzy-based approach to traditional approach comparison table

	Traditional approach	**Proposed approach**
The total number of transmissions	12036 transmission	7279 transmission
Total energy consumption of total transmissions in joules	584.244 joule	354.148 joule

Figure 6 represents the total energy in joules that is consumed for total transmissions over the whole period from 12/17/2014 to 6/5/2015. The solid column represents energy consumption using the traditional approach that was about 584.244 joule. While the total energy consumption using proposed approach at threshold 30 degree appeared in dashed column that was about 354.148 joule .

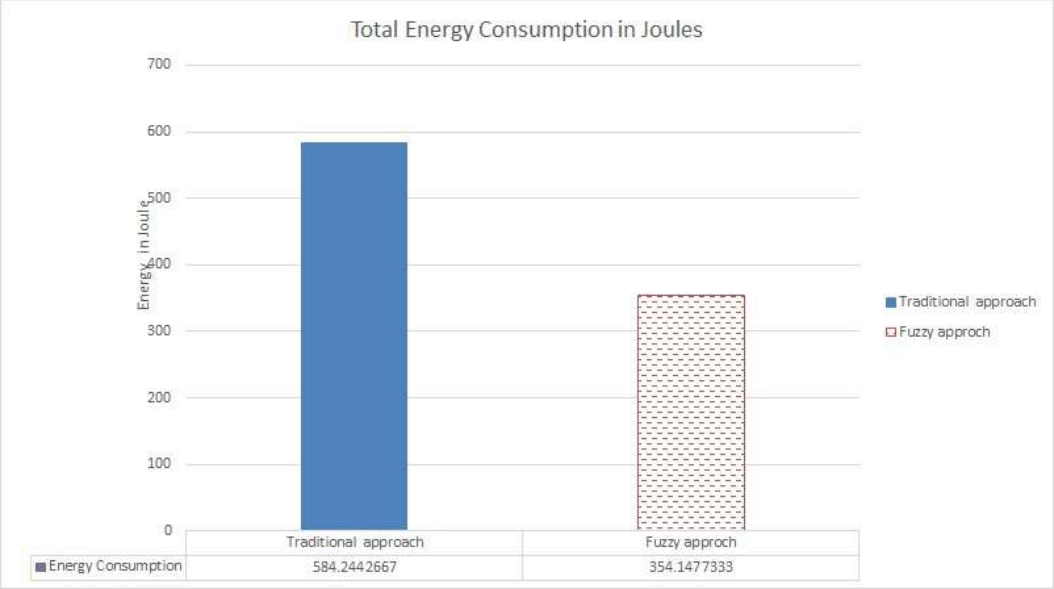


Figure 6: The overall consumed energy in joules

Figure 7 the energy is consumed for total transmissions to the server in the experiment over the whole period from 17/12/2014 to 5/6/2015 in a cumulative form solid curve for using traditional approach and dashed carve using proposed approach at threshold 30 degree.

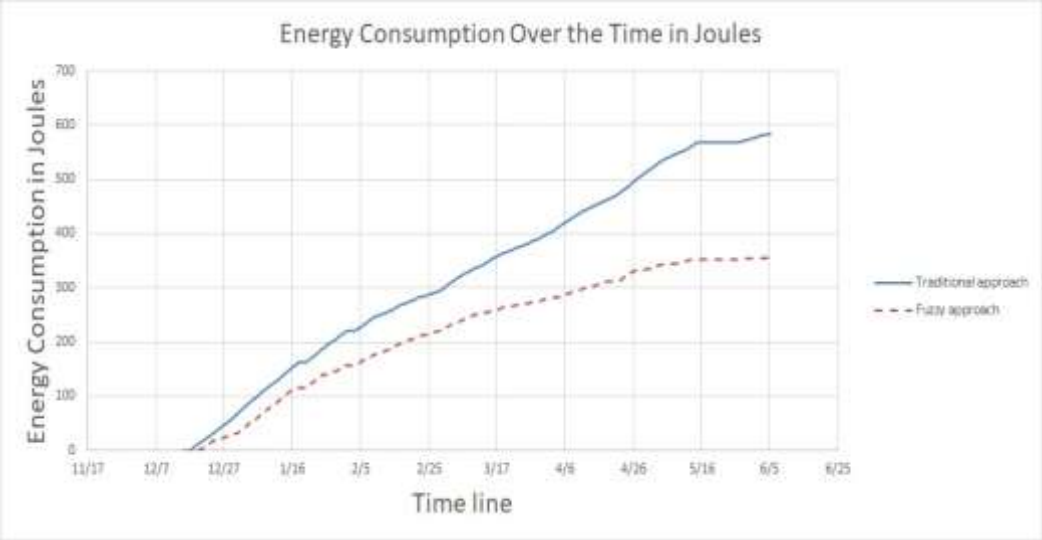


Figure 7: Energy consumption for total transmissions in joules

7. CONCLUSION

The increasing demands of reducing energy consumption especially in IoT environments asks for methods and techniques to reduce energy consumption. For this reason, this paper proposed a fuzzy-based approach for temperature and humidity monitoring systems that reduces the energy consumed during packets exchange. Fuzzy logic was used to determine whether the captured

temperature degree should be transmitted to the control room's server or not. The experimental results on a Melbourne dataset showed that there is a tradeoff between the energy reduction and the quality of data so it's important to be careful when choosing the threshold to keep the balance between them Also at threshold 30 the energy consumed has been reduced by 39.4% when compared to the traditional monitoring systems approach.

Referances

Ashton, K., et al., 2009. That internet of things thing. RFID journal 22, 97–114.

Baccelli, E., Hahm, O., Gunes, M., Wahlisch, M., Schmidt, T.C., 2013. Riot os: Towards an os for the internet of things, in: 2013 IEEE conference on computer communications workshops (INFOCOM WKSHPS), IEEE. pp. 79–80.

Bandyopadhyay, D., Sen, J., 2011. Internet of things: Applications and challenges in technology and standardization. Wireless Personal Communications 58, 49–69.

Bhunia, S.S., Dhar, S.K., Mukherjee, N., 2014. ihealth: A fuzzy approach for provisioning intelligent health-care system in smart city, in: 2014 IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), IEEE. pp. 187–193.

Chan, C.O., Lau, H.C., Fan, Y., 2018. Iot data acquisition in fashion retail application: Fuzzy logic approach, in: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE. pp. 52–56.

Chen, D., Chang, G., Sun, D., Li, J., Jia, J., Wang, X., 2011. Trm-iot: A trust management model based on fuzzy reputation for internet of things. Comput. Sci. Inf. Syst. 8, 1207–1228.

Collotta, M., Pau, G., 2015. Bluetooth for internet of things: A fuzzy approach to improve power management in smart homes. Computers & Electrical Engineering 44, 137–152.

Conti, M., Dehghantanha, A., Franke, K., Watson, S., 2018. Internet of things security and forensics: Challenges and opportunities.

Cueva-Fernandez, G., Espada, J.P., García-Díaz, V., Gonzalez-Crespo, R., 2015. Fuzzy decision method to improve the information exchange in a vehicle sensor tracking system. Applied Soft Computing 35, 708–716.

Dohr, A., Modre-Opsrian, R., Drobics, M., Hayn, D., Schreier, G., 2010. The internet of things for ambient assisted living, in: 2010 seventh international conference on information technology: new generations, Ieee. pp. 804–809.

Dubois, D.J., 1980. Fuzzy sets and systems: theory and applications. volume 144. Academic press.

Fahmy, I.M., Nassef, L., Hefny, H.A., 2014. Predicted energy-efficient bee-inspired routing (peebr) path selection optimization. Journal of Wireless Networking and Communications 4, 33–41.

Fu, C., Zhang, G., Yang, J., Liu, X., 2011. Study on the contract characteristics of internet architecture. Enterprise Information Systems 5, 495–513.

Gao, M., Wang, Q., Arafin, M.T., Lyu, Y., Qu, G., 2017. Approximate computing for low power and security in the internet of things. Computer , 27–34.

Gazis, V., Goertz, M., Huber, M., Leonardi, A., Mathioudakis, K., Wiesmaier, A., Zeiger, F., 2015. Short paper: Iot: Challenges, projects, architectures, in: 2015 18th International Conference on Intelligence in Next Generation Networks, IEEE. pp. 145–147.

Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M., 2013. Internet of things (iot): A vision, architectural elements, and future directions. Future generation computer systems 29, 1645–1660.

Guo, B., Zhang, D., Wang, Z., 2011. Living with internet of things: The emergence of embedded intelligence, in: 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing, IEEE. pp. 297–304.

Jayakumar, H., Raha, A., Kim, Y., Sutar, S., Lee, W.S., Raghunathan, V., 2016. Energy-efficient system design for iot devices, in: 2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE. pp. 298–301.

John, A., Rajput, A., Babu, K.V., 2017. Energy saving cluster head selection in wireless sensor networks for internet of things applications, in: 2017 International Conference on Communication and Signal Processing (ICCSP), IEEE. pp. 0034–0038.

Kamalinejad, P., Mahapatra, C., Sheng, Z., Mirabbasi, S., Leung, V.C., Guan, Y.L., 2015. Wireless energy harvesting for the internet of things. IEEE Communications Magazine 53, 102–108.

Karim, A., Hassan, A., Akanda, M., et al., 2018. Monitoring food storage humidity and temperature data using iot. MOJ Food Process Technol 6, 400–404.

Kaur, A., Kaur, A., 2012. Comparison of mamdani-type and sugeno-type fuzzy inference systems for air conditioning system. International journal of soft computing and engineering 2, 323–325.

Kaur, J., Kaur, K., 2017. A fuzzy approach for an iot-based automated employee performance appraisal. Computers, Materials and Continua 53, 24–38.

Khatibi, A., Khatibi, O., 2018. An improved energy-aware clustering method for the regional queries in the internet of things. arXiv preprint arXiv:1806.11553 .

Lee, I., Lee, K., 2015. The internet of things (iot): Applications, investments, and challenges for enterprises. Business Horizons 58, 431–440.

Listyorini, T., Rahim, R., 2018. A prototype fire detection implemented using the internet of things and fuzzy logic. World Trans. Eng. Technol. Educ 16, 42–46.

Mahalle, P.N., Thakre, P.A., Prasad, N.R., Prasad, R., 2013. A fuzzy approach to trust based access control in internet of things, in: Wireless VITAE 2013, IEEE. pp. 1–5.

Mahmoud, M.S., Mohamad, A.A., 2016. A study of efficient power consumption wireless communication techniques/modules for internet of things (iot) applications .

Maksimovic, M., Vujovic, V., Perisic, B., Milosevic, V., 2015. Developing a fuzzy logic based system for monitoring and early detection of residential fire based on thermistor sensors. Comput. Sci. Inf. Syst. 12, 63–89.

Meana-Llorián, D., García, C.G., G-bustelo, B.C.P., Lovelle, J.M.C., Garcia-Fernandez, N., 2017. Iofclime: The fuzzy logic and the internet of things to control indoor temperature regarding the outdoor ambient conditions. Future Generation Computer Systems 76, 275–284.

City of Melbourne, A., 2017. Sensor readings, with temperature, light, humidity every 5 minutes at 8 locations (trial, 2014 to 2015). Also available athttps://data.melbourne.vic.gov.au/Environment/Sensor-readings-with-temperature-light-humidity-ev/ ez6b-syvw.

Purwanto, F.H., Utami, E., Pramono, E., 2018. Design of server room temperature and humidity control system using fuzzy logic based on microcontroller, in: 2018 International Conference on Information and Communications Technology (ICOIACT), IEEE. pp. 390–395.

Ross, T.J., 2005. Fuzzy logic with engineering applications. John Wiley & Sons.

Santamaria, A.F., Raimondo, P., De Rango, F., Serianni, A., 2016. A two stages fuzzy logic approach for internet of things (iot) wearable devices, in: 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), IEEE. pp. 1–6.

Shafique, M., Hafiz, R., Rehman, S., El-Harouni, W., Henkel, J., 2016. Cross-layer approximate computing: From logic to architectures, in: Proceedings of the 53rd Annual Design Automation Conference, ACM. p. 99.

Shah, P.A.A., Habib, M., Sajjad, T., Umar, M., Babar, M., 2016. Applications and challenges faced by internet of things-a survey, in: International Conference on Future Intelligent Vehicular Technologies, Springer. pp. 182–188.

Shahrestani, S., 2017. Internet of Things and Smart Environments: Assistive Technologies for Disability, Dementia, and Aging. Springer.

Singhala, P., Shah, D., Patel, B., 2014. Temperature control using fuzzy logic. arXiv preprint arXiv:1402.3654

Suganthi, L., Iniyan, S., Samuel, A.A., 2015. Applications of fuzzy logic in renewable energy systems–a review. Renewable and sustainable energy reviews 48, 585–607.

Whitmore, A., Agarwal, A., Da Xu, L., 2015. The internet of things — a survey of topics and trends. Information Systems Frontiers 17, 261–274.

Zadeh, L.A., 1965. Fuzzy sets. Information and control 8, 338–353.

A Computational Intelligent Approach for Managing Fish Farming

Sameh A. Elsayed Ahmed M. Gadallah Hesham A. Hefny

Cairo University-Faculty of Graduate Studies for Statistical Research

{ eng.samelsayed, ahmgad10} @yahoo.com, hehefny@ieee.com

Abstract: Water changes in its parameters like dissolved oxygen, acidity and temperature. These ongoing changes have more impact on fish farming. In other words many sites are not suitable for fish farming because of the bad water quality .Instead, fish farming become more suitable at other new sites or adjusted sites. This paper provides an intelligent fuzzy approach for selecting fish farming places. The approach consists of four phases ,The first phase for preparing water quality parameters of spatial database and the required water quality for fish farming .The second phase defines a set of fuzzy membership functions for defining the matching degree of water quality of sites with the required fish farming water quality. In the third phase the approach perform fuzzy query on the database. The fourth phase performs clustering of sites according to their suitability degree. The proposed approach outputs will be the more suitable sites for fish farming and sites compatible with international standard of water quality. The approach has been applied on the water quality parameters dissolved oxygen, acidity and temperature in Nile River of Egypt. The final results are presented to a set of researcher in the general authority for fish resources development and they were agreed with the results.

Keywords: Fuzzy logic, Data mining, Fuzzy query, Fish Farming Database, Water quality, Spatial database.

1. Introduction

Selecting the most suitable sites for fish farming is one of the management methods for fish farming. Water quality (WQ) is a measure of the physical chemical and biological properties of the water .Sites are different in WQ level so it is required to select the best WQ sites before start fish farming. Fish needs an appropriate degree of water parameters like dissolved oxygen (DO), acidity called (PH) and temperature (Salama et al 2016;Xu & Chen 2014). DO is amount of gaseous oxygen dissolved present in water. It is an important parameter in evaluating water quality. A DO level that is too low can harm fish life and affect WQ. It is measured in mg/l. DO should not be less than 5 mg/l for most fish (Sampuangthong et al 2014 ; Zaprudnova et al 2015). The parameter PH describes the acidity or alkalinity of water. The PH value represents the instantaneous hydrogen that affects biological and chemical of WQ. Most fish require water with PH between 6.5 and 8.4, fish are harmed if PH become below 4.8 or more than 9.2. Temperature effects fish breading and health, if temperature is very low or very high it is lead the diseases and mortality and loss of weight of fish. Most fish has some tolerances

in water parameters increasing or decreasing. Fish has tolerance to low PH, Zaprudnova et al 2015.The value of global fish production reached 171 million tons in 2016, which is the highest rate of production ,which indicates that fish farming is of great importance in increasing fish production, FAO 2018.

Fuzzy logic can handle partial truth, the query returns the values as degree of membership rather than crisp value true or false, Zadeh 1965, so fuzzy logic can be used to deal with the tolerance of fish in water parameters over the traditional methods. And fuzzy logic is more accurate in results over the traditional methods. Fuzzy queries have appeared to cope with the necessity to soften the Boolean logic in database queries. A fuzzy query system is an interface to users to get information from database using natural language sentences.

The study use data mining with fuzzy logic to evaluate the sites suitability degree according to WQ parameters of the sites and fish farming parameters like DO, PH and temperature. The sites are in the Nile River of Egypt. The study also compares WQ parameters of sites with the international standard. The next section will identify the problem; the next is the related work then the suggested methodology and a case study.

2. The Problem

Aquaculture are now facing a big challenge due to bad water quality parameters, climate change, pollution, Caldo and Dadios 2013, and that it is inappropriate in some places for fish farming. Fish farming requires suitable water quality parameters that are too low in some places. So it is very important to select the best sites with high WQ before start fish farming. To avoid loss of fish weight, reduce killed fish, reduce fish diseases and increase fish production.

3. Related works

Many approaches have worked on the problem of WQ and fish farming, some approaches show methods of WQ assessment, but these studies don't introduce a method for solving the problems of inappropriate of WQ for fish farming (Kateriya & Tiwari 2016 ; Sampuangthong & Leelasantitham 2014 ; Hidayah et al 2011). Other studies show the effects of change of water parameters like dissolved oxygen and PH on fish health Rana & Rani 2015. The paper, Huang et al 2010, simulates the effect of cage fish farming on the environment. Other approaches that work on assessment of water parameters, these types of studies introduce some solution for bad WQ that inappropriate for fish farming like add some material to water to improve WQ parameters(Ezeanya et al 2015; Xu & Chen 2014).Other approaches determine or select sites that have the most suitable WQ for fish farming(Salama et al 2016; Kasoz et al 2016 ; Yalcuk & Postalcioglu 2015; Zadeh 1965).

Unfortunately, all of the above approaches do not provide weight or matching degrees of WQ suitability for fish farming, also these don't compare the measures of WQ with the fish farming WQ requirements. Authors of (Salama et al 2016 ; Kasoz et al 2016) present

method to select the suitable site for fish farming. It evaluates sites according to WQ, some sites are suitable for fish farming and others are not. But this evaluation is crisp and no flexibility in it so a site with 0.9 degree is not suitable. Although this degree may be very appropriate for fish farming.

Authors of papers (Lv & Zhou 2013 ; Yalcuk & Postalcioglu 2015; Xu & Chen 2014) evaluate WQ in different sites using fuzzy logic, but don't compare sites WQ with the fish farming WQ requirements. Other study like, Ezeanya et al 2015, add some material to enhance WQ but this is a cost.

This paper introduces a new approach that evaluates the suitability of WQ for fish farming. It determines the sites suitability in fuzzy manner. It performs a fuzzy query on the selected spatial sites according to WQ requirements appropriate for fish farming .The result of this fuzzy query will be represented as degree of membership rather than accepted sites or not accepted sites for fish farming. Sites that reach or exceed a determined threshold will be accepted for fish farming. After that the study makes fuzzy clustering of spatial sites according to sites suitability degree. Also the approach compared the sites with international standard of WQ.

4. The proposed approach

The proposed computational intelligent approach for managing fish farming consists of four phases "Figure.1 The structure of the proposed approach".

4.1 Preparation of WQ parameters

This phase used to collect and prepare the data of WQ parameters of the selected sites and WQ parameters required for fish farming. Data cleaning so the noise and inconsistent data are removed.

4.2 Define the suitable WQ requirements represented as fuzzy membership functions

The membership function for a fuzzy set A on the universe of discourse "X" is defined as $\mu A: X \rightarrow [0,1]$, where each element of X is mapped to a value between "0" and "1". This value, called membership value or degree of membership or matching degree, quantifies the grade of membership of the element in "X" to the fuzzy set "A", Zadeh 1965. This phase defines the fuzzy membership functions that define the relation between the fish farming WQ parameters and WQ parameters existing in a specific site.

4.2.1. Define the fuzzy membership function of "PH" and Temperature degree for a specific fish type

- PH degrees in [b,c] are the most suitable with full matching degree equal to "1".
- PH degrees in [a,b[and [c,d[are partially suitable with matching degree in [0,1[.
- PH degrees greater than "d" or less than "a" are not suitable, its matching degree is "0".
- This can be described as trapezoidal fuzzy membership function, Eq (1).

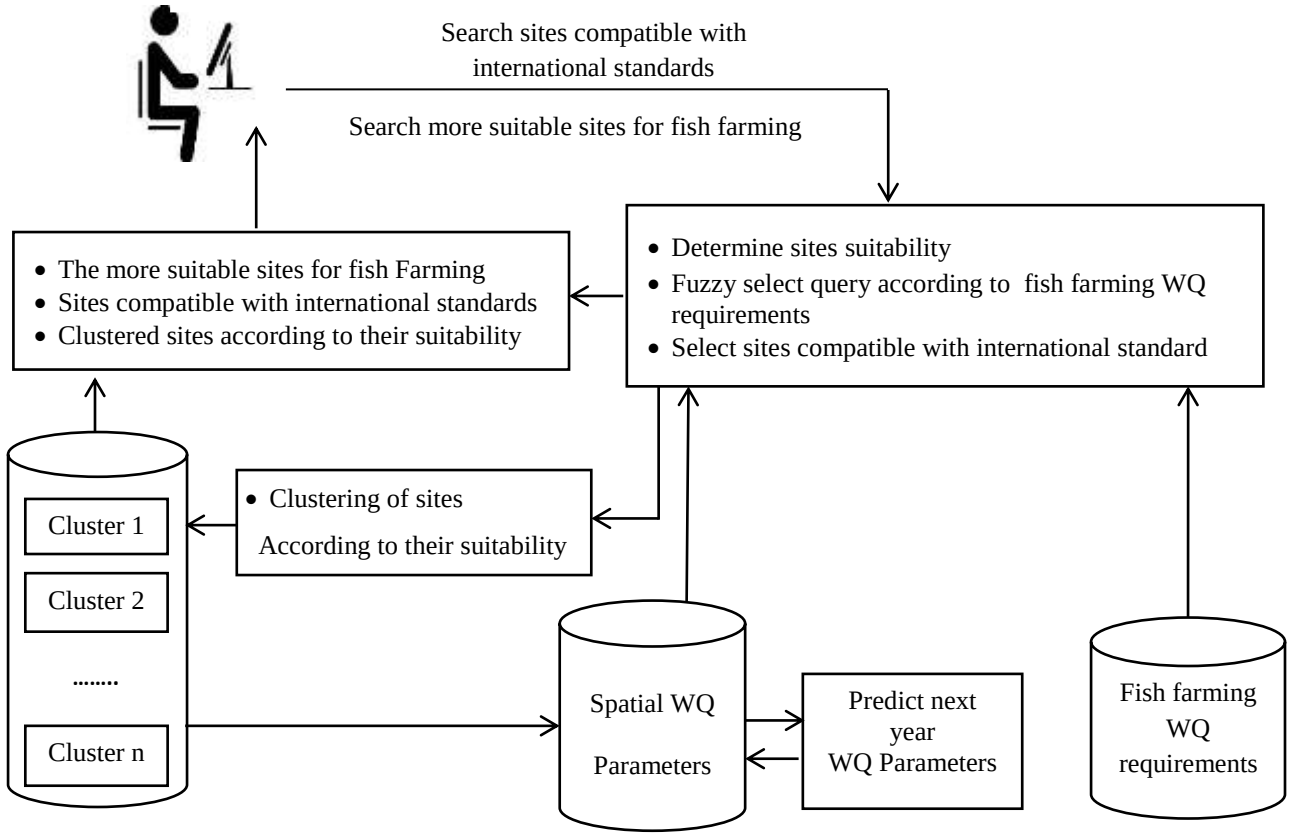
$$\mu_{\text{PH or Temp\_}}\text{ suitability (x)} = \begin{cases} 0, & (x < a) or (x > d) \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \frac{d-x}{d-c}, & c \le x \le d \end{cases} \qquad \text{Eq (1)}$$

4.2.2. Define the fuzzy membership function of DO degree for a specific fish type

- DO degree greater than "b" is the most suitable with full matching degree equal to "1".
- DO degrees in [a, b] are partially suitable with matching degree in [0,1].
- DO degree less than "a" are not suitable, and its matching degree is "0".
- This can be described as linear fuzzy membership function, Eq. (2).

$$\mu_{\text{DO\_}}\text{ suitability (x)} = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases} \qquad \text{Eq. (2)}$$

Fig (1) .The structure of the proposed approach



4.3 Perform fuzzy Query

This phase takes the databases of WQ parameters of the selected sites and the database of WQ parameters of fish farming requirements and the accepted threshold for the membership function. Then the phase calculates the matching degree of each site for fish farming. The phase selects sites with matching degree larger than or equals the accepted threshold. Also the phase selects sites compatible with international standards this phase is performed by the following fuzzy selection algorithm "Algorithm 1".

Algorithm 1. Fuzzy selection algorithm

***Inputs*:** WQ parameters of sites, WQ parameters require for fish farming, accepted threshold for suitability membership function, WQ international standard for fish farming.

Outputs: Accepted sites for fish farming, sites compatible with WQ international standard for fish farming.

For *each site*

For *each fish type*

PH_matching_degree=PH_matching (min_PH,"from-to"_PH, max_PH, Fish_required_PH)

DO_matching_degree=DO_matching(min_DO,Greater than_DO, Fish_Required_DO)

Temp_matching_degree= Temp _matching (min_ Temp,"from-to"_ Temp, max_ Temp, Fish_required_ Temp)

Suitability =Min **(**PH_matching_degree, DO_matching_degree, Temp_matching_degree)

If *(Suitability >= threshold)*

Add Site to accepted sites

Else

Continue

If *(Suitability compatible with international standard)*
Add Site to accepted international standard sites

Else

Continue

End

End

4.4- Clustering of sites according to their suitability degree

Clustering is the assignment of objects to groups of similar objects Attributes can be numerical or categorical. The assignment can be hard, where each object belongs to one cluster, or fuzzy, where an object can belong to several clusters with a probability. The clusters can be overlapping, though typically they are disjoint, Romera et al 2017.After determining the suitability degree of sites, the study will cluster the sites according to its suitability degree; these clusters can be introduced to decision maker.

5. An illustrative case study

To show how the approach work, we are going to a case study of spatial database of some sites along Nile River and database of fish that live in Nile River and show the suitability degree of these sites. “Table 1.The selected sites and its parameters” the table show the main parameters PH, DO, Temp.” Table 2.The selected fish of freshwater and their required parameters” the table show the selected fish live in freshwater and their required parameters. PH “min” is the minimum required PH for fish ,PH "from-to" is the most suitable PH for fish , “max” is the maximum required PH for fish. DO “min” is the minimum required DO for fish, DO “greater than” is the most suitable for fish. Table 3 show the calculated suitability degree.

Table 1.The selected sites and its parameters

Site	Symbol	PH	DO	Temp
Aswan	S1	7.925625	5	26.254375
Asyut	S2	6.8	7.20625	21.095625
Beni Suef	S3	7.3375	5.775	21.62375
Cairo	S4	7.0125	5.9125	21.72
Damietta	S5	7.65625	4.9875	19.418125
El Giza	S6	7.3625	6.01875	20.535625
Mansura	S7	5.8875	6.15	20.12
El Minya	S8	6.94375	6.2375	20.930625
Girga	S9	6.48125	6.24375	20.298125
Idfu	S10	7.3875	7.06875	24.3825
Kom Ombo	S11	5.025	3.44375	24.91875
Luxor	S12	6.84375	4.775	24.1025
Maghagh	S13	6.75625	5.475	21.404375
Qena	S14	7.63125	6.69375	23.363125
Sohag	S15	6.9875	6.29375	23.8925
Rahawy	S16	7.5	5.1625	23.93625
Sabal	S17	6.73125	3.8875	22.76125
Tala	S18	6.675	5.2	24.201875

Table 2.The selected fish of freshwater and their required parameters

Fish Name / WQ	Symbol	PH				DO Mg/liter		Temp °C			
		Min	from	to	Max	Min	greater than	Min	from	to	Max
Shrimp	F1	4.5	6.5	8	10	3.5	5	9	14	29	38
African butter catfish	F2	4	6	7.8	9	3	5.3	10	23	27	37
Anguilla	F3	5	7.4	8	10	2.8	5	9	23	30	39
Fresh water Sardines	F4	4	6.2	6.5	8	2	4	7	12	15	26
African catfish	F5	4	6.5	8	10	4	6	9	20	30	40
Electric catfish	F6	4.5	7	8	10	2.9	4	11	23	30	39
Bagrus docmac	F7	4.4	6.4	8.2	10	3	4.5	10	21	27	38
Lates niloticus	F8	4	6.2	8.5	10.2	3	4	11	21	27	37
Bottle nose	F9	3.5	5.5	8	10	3.5	5	13	23	30	38
Hypoph thalmichthys	F10	3	5.5	7.5	9.5	2.5	4	13	20	24	33
Barbus bynni	F11	3.5	6	8	10	2.7	4	10	22	27	35
Nile lebo	F12	4.2	7	8.5	10.5	4	6	10	22	28	38
Grass carp	F13	4	6	8.6	10.4	5	8	12	20	30	40
T.Zilli	F14	4	6	7.5	9	5	7	11	26	32	40
T.gallilea	F15	4	6	8.4	10	4	7	13	26	31	39
T.Aurea	F16	4	6	8	10	6	9	10	21	29	37
T.Niletica	F17	4	6	9	10.5	4	9	12	22	28	36

Table 3. The calculated suitability degree

Site /Fish	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
F1	1.00	1.00	1.00	1.00	0.99	1.00	0.69	1.00	0.99	1.00	0.00	0.85	1.00	1.00	1.00	1.00	0.26	1.00
F2	0.87	0.85	0.89	0.90	0.72	0.81	0.78	0.84	0.79	1.00	0.19	0.77	0.88	1.00	1.00	0.94	0.39	0.96
F3	1.00	0.75	0.90	0.84	0.74	0.82	0.37	0.81	0.62	0.99	0.01	0.77	0.73	1.00	0.83	1.00	0.49	0.70
F4	0.00	0.45	0.40	0.39	0.23	0.43	0.53	0.46	0.52	0.15	0.10	0.17	0.42	0.24	0.19	0.19	0.29	0.16
F5	0.50	1.00	0.89	0.96	0.49	1.00	0.76	1.00	0.99	1.00	0.00	0.39	0.74	1.00	1.00	0.58	0.00	0.60
F6	1.00	0.84	0.89	0.89	0.70	0.79	0.56	0.83	0.77	1.00	0.21	0.94	0.87	1.00	1.00	1.00	0.89	0.87
F7	1.00	1.00	1.00	1.00	0.86	0.96	0.74	0.99	0.94	1.00	0.30	1.00	1.00	1.00	1.00	1.00	0.59	1.00
F8	1.00	1.00	1.00	1.00	0.84	0.95	0.86	0.99	0.93	1.00	0.44	1.00	1.00	1.00	1.00	1.00	0.89	1.00
F9	1.00	0.81	0.86	0.87	0.64	0.75	0.71	0.79	0.73	1.00	0.00	0.85	0.84	1.00	1.00	1.00	0.26	1.00
F10	0.75	1.00	1.00	1.00	0.92	1.00	1.00	1.00	1.00	0.96	0.63	0.99	1.00	0.93	1.00	1.00	0.92	0.98
F11	1.00	0.92	0.97	0.98	0.78	0.88	0.84	0.91	0.86	1.00	0.57	1.00	0.95	1.00	1.00	1.00	0.91	1.00
F12	0.50	0.92	0.89	0.96	0.49	0.88	0.60	0.91	0.81	1.00	0.00	0.39	0.74	1.00	1.00	0.58	0.00	0.60
F13	0.00	0.74	0.26	0.30	0.00	0.34	0.38	0.41	0.41	0.69	0.00	0.00	0.16	0.56	0.43	0.05	0.00	0.07
F14	0.00	0.67	0.39	0.46	0.00	0.51	0.58	0.62	0.62	0.89	0.00	0.00	0.24	0.82	0.65	0.08	0.00	0.10
F15	0.33	0.62	0.59	0.64	0.33	0.58	0.55	0.61	0.56	0.88	0.00	0.26	0.49	0.80	0.76	0.39	0.00	0.40
F16	0.00	0.40	0.00	0.00	0.00	0.01	0.05	0.08	0.08	0.36	0.00	0.00	0.00	0.23	0.10	0.00	0.00	0.00
F17	0.20	0.64	0.36	0.38	0.20	0.40	0.43	0.45	0.45	0.61	0.00	0.16	0.30	0.54	0.46	0.23	0.00	0.24

Consider table.3, the suitability degree of some sites reach to 1.0 so the fish farming in these sites is suitable for fish farming without WQ enhancements or treatments. Other sites suitability is less than 1.0 but with accepted threshold, this threshold >=0.8 and is determined by experts, so sites with suitability from (0.80 to 1.0) can be used for fish farming. In the other hand experts determined sites with suitability from (0.70 to less than 0.80) can be used for fish farming after making some WQ enhancements or water treatments. So sites with suitability from (0.70 to less than 0.80) can be added to fish farming sites. Other sites with suitability less than 0.70 must be avoided in fish farming , if it is necessary to use these sites for fish farming it need a big enhancement and water treatment to be suitable for fish farming.

Consider table.4, Clustering of sites according to suitability degree, the description describe the cluster and some advices about it. The percentage of "class I" is about "47.7 %" in this approach compared with about "25.16 %".

Table 4. Clustering of sites

Class	Suitability	Range of Suitability	Description	Percentage of sites in cluster
Class I	Fit	From (0.80 to 1.0)	Accepted threshold for fish farming is >=0.80	≃ 47.71 %
Class II	Good	From (0.70 to less than 0.80)	Less than the accepted threshold and can be used for fish farming after making some WQ enhancements or water treatments.	≃ 6.5 %
Class III	Bad	From (0.50 to less than 0.70)	Avoided in fish farming, if it is necessary to use these sites for fish farming it need a big enhancement and water treatment to become suitable for fish farming.	≃ 11.11 %
Class IV	Very bad	From 0.30 to less than 0.50	Hard enhancement and treatment of water.	≃ 12.42 %
Class V	Unfit	Less than 0.30	Unfit to use.	≃ 20.92 %

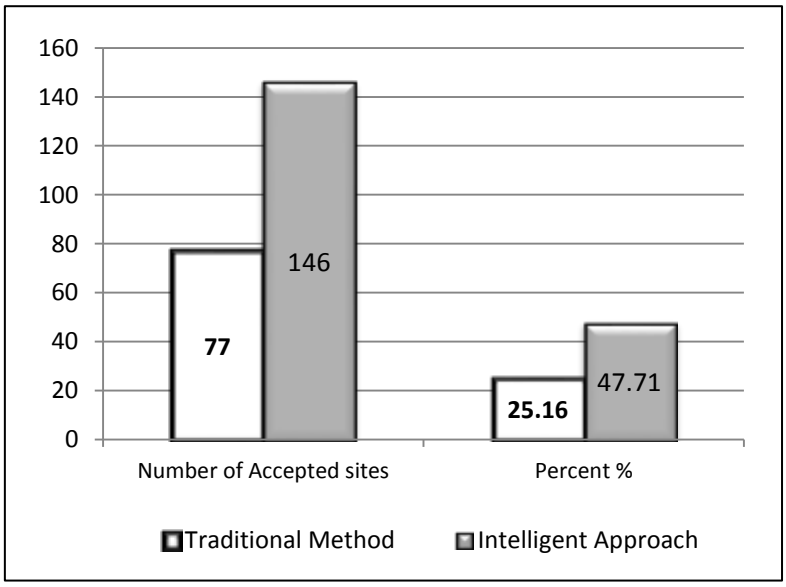
Consider table.5 WQ parameters of sites compared with WQ of international standard , Oram 2014, the mean value of DO in Nile River not compatible with international standard of WQ but the mean of DO is appropriate for fish farming, while the other parameters PH and temperature are compatible with international standard of WQ.

Table.5 WQ parameters of sites and international standard

WQ parameter	Unit	Mean value (Nile River)	International standard
PH	PH	6.94	From 6.0 to 9.0
Temperature	^{0}C	22.5	From 15 to 30
DO	mg/l	5.64	From 4 to 5

Consider the results of the intelligent approach "Figure 2. comparing the results of the proposed approach with the traditional methods" where the accepted threshold for sites suitability degree is ">=0.8" the percent of accepted sites for fish farming is "47.71%"in case of the intelligent approach, compared with the percent "25.16%" of traditional methods .

Figure.2 Compare the results of the intelligent approach with the traditional methods



6. Conclusions and future work

This paper introduced a computational intelligent approach depending on fuzzy logic and data mining for managing fish farming based on the available historical WQ parameters database of sites, and the WQ parameters database required for fish farming, the approach determined the suitability degree of sites, the results can be introduced to experts and fish farmers to help them in selecting the best suitable sites for fish farming. The paper also compared the WQ parameters with the international standard. In the future some enhancement will be added to the approach like adding some other WQ parameters such as dissolved Ammonia, Nitrate, Phosphate and carbon Dioxide to get better degree of accuracy. Also the study will find the most WQ parameter affect the suitability degree of sites. By

finding the most WQ parameters affect the suitability degree of sites, so enhancement will be in these parameters just and the number of accepted sites will be increased.

The study also can make the database of site parameters and the database of fish parameters with weights or degrees of effects if the parameters are different in its effects in fish farming this will lead to more accuracy results.

7. References

Caldo, R.B.,&Dadios, E.P., 2013. Fuzzy Logic Control of Water Quality Monitoring and Surveillance for Aquatic Life Preservation in Taal Lake, 2013.

Campbell,J.E. ,&Shin,M.,2012. Geographic Information System Basics",PP. 8-32,Vol.1.0, 2012.

Chen, G., &Pham,T.,2001, Introduction to Fuzzy Sets, Fuzzy Logic and Fuzzy Control Systems, University of Houston ,Houston, Texas 2001, pp. 57-86.

Ezeanya, N.C., Chukwuma,G.O., Nwaigwe, K.N.,& Egwuonwu**, C.C.**, 2015.Standard Water Quality Requirements And Management Strategies For Fish Farming (A Case Study Of Otamiri River), International Journal of Research in Engineering and Technology, Vol. 04 Issue: 03 , Mar-2015.

FAO. 2018. The state of world fisheries and aquaculture, [http://www.fao.org/state-of-fisheries-aquaculture].Accessed 22 October 2019.

Hidayah, S.N., Tahir,N., Rusop,M.,& Rizam,S.,2011. Development of Fuzzy Fish Pond Water Quality Model,IEEE colloquium on humanities ,science and engineering research(chuser) , pp. 556-561,Dec 5-6-2011.

Huang, H. Jia, X. Lin, Q. Guo, G. ,& Liu,y., 2010. Application of Adaptive Neuro-fuzzy Inference System to Carrying Capacity Assessment for Cage Fish Farm in Daya Bay,China,in 2010 IEEE Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE. pp. 815-818.

Kasoz,N., Opi,H. ,&Iwe,G., Enima,C., Nkambo,M.,Turyashemererwa,M., Naluwayiro, J., & Sadik,K., 2016. Site suitability assessment of selected bays along the Albert Nile for Cage Aquaculture in West Nile region of Uganda, International Journal of Fisheries and Aquaculture, Vol. 8(9),September 2016.

Kateriya, B., &Tiwari,R., 2016. River Water Quality Analysis and Treatment Using Soft Computing Technique: A Survey ,IEEE,International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, INDIA, Jan. 07 – 09, 2016.

Lv, H.,& Zhou, F., 2013. Identification of the Yancheng Region Water Quality Using GIS and Fuzzy Synthetic Evaluation Approach, 2013.

Oram,B. Dissolved oxygen in water,2014 [https://www.water research.net /index.php /dissovled-oxygen-in-water], Accessed 15 May 2019.

Rana, D.S, &Rani,S., 2015. Fuzzy Logic Based Control System for Fresh Water Aquaculture: A MATLAB based Simulation Approach, Serbian Journal Of Electrical Engineering, Vol. 12, No. 2, June 2015, pp. 171-182.

Romera, J.M., Gutiérrez, J.G., Ballesteros, M.M. ,&Santos,J. R.,2017. An approach to validity indices for clustering techniques in Big Data, springer, 5 - October 2017.

Salama, A.J., Satheesh,S., Balqadi ,A.A., &Kitto,M.R., 2016. Identifying Suitable Fin Fish Cage Farming Sites in the Eastern Red Sea Coast, Saudi Arabia, Springer International Publishing Switzerland , Thalassas, 32:1–9, 4 February 2016.

Sampuangthong, K., Kiattisin, S. ,& Leelasantitham, A., 2014. Evaluation Levels of Water Quality in Maeklong Basin Using Fuzzy Logic,IEEE,Biomedical Engineering International Conference (BMEiCON) ,2014.

Xu, T., &Chen, F., 2014. An Embedded Fuzzy Decision System for Aquaculture, IEEE Workshop on Electronics, Computer and Applications, pp. 351–353, 2014.

Yalcuk, A., & Postalcioglu, S., 2015. Evaluation of pool water quality of trout farms by fuzzy logic: monitoring of pool water quality for trout farms, Int. J. Environ. Sci. Technol.,(2015) 12:1503–1514, Azad University (IAU), 19 March 2014.

Zadeh, L.A.,1965. Fuzzy sets. Information and control, vol. 8, pp. 338–353 ,1965

Zaprudnova, R.A. Kamshilov, I.M. ,&Chalov, Y.P.2015. Functional Properties Of Hemoglobin During The Adaptation Of Fish To Low Environmental Ph, Inland Water Biology Vol. 8 No. 2, pp. 188–194, 2015.

Improving the Performance of Colored QR Code

Islam M. El-Sheref[1] Ahmed H. Asad[1] Fatma A. El-Licy[1]

Abstract

Many attempt have been performed to resove the problems associated with Color and image distortion of the QR colored code. In particular, due to its vast popularity and usefull applications, which encourage its enhancement to improve the capacity and relegate the effects of the image and color distortion. The paper presents a short history for the QR code and the current research advances for improving quick response code storage capacity. The main goal of the paper is to establish a methodology to improve the performance of the QR color code, while, maintainning its colored image undestorted.

Keywords: 2-D barcodes, QR Codes - Colored QR Codes - Error Correcting Codes - Image multiplexing - Data Capacity Enhancement - Multiplexing QR Color Coding.

1. Introduction

QR is short for quick response two-dimensional code that have become widely popular along with its useful applications. Its specifications offer many more advantages than that of the one-dimensional code. The reason why they are more useful than a standard barcode is that they can store (and digitally present) much more data, with diverse format and types including URL links, geo coordinates, and text. The other key feature of QR Codes is that it is supported by smart systems and many of the modern cell phones, therefore, no special hand-held scanner is required.

The Quick Response (QR) code is invented by the one of major Toyota group companies in 1994 and was initially used for tracking inventory in vehicle parts manufacturing [see Ahlawat et al. (2017)], and approved as an ISO international standard (ISO/IEC18004) in June 2000 [see Soon, 2008]. A QR code is a type of matrix or two-dimensional barcode that stores up to 4296 characters of information and it is designed to be read by smartphones [see Tiwari (2016)]. QR stands for "Quick Response" indicating that the code contents should be decoded very quickly at high speed. A QR code consists of black modules arranged in a square pattern on a white background. The information encoded may be text, a URL or other data [see Singh and

[1] Computer Sciences department, Faculty of Graduate studies for Statistical Research, Cairo University

Singh (2016)]. The idea behind the development of the QR code is the limitation of the standard Universal Product Code (UPC) barcode information capacity (that holds, only, 20 characters) [see Kinjal et al. (2014)]. QR codes system became popular due to its fast readability and greater storage capacity [see Rizqi et al. (2018), Bjardwaj et al. (2016), Goyal et al. (2016), Čović et al. (2016)].

The main objective of this paper is to improve the performance of the colored QR code.

1.1. The History of QR Code Development

In 1970, IBM developed UPC symbols consisting of 13 digits of numbers to enable automatic input into computers. These UPC symbols are still widely used for Point-Of-Sale (POS) system. In 1974, Code 39 which can encode (symbolize) approximately 30 digits of alphanumeric characters was developed. Then in the early 1980s, multi-staged symbol codes where approximately 100 digits of characters can be stored such as Code 16K and Code 49 were developed. As informatization rapidly developed in the recent years, requests had mounted for symbols which can store more information and represent languages other than English. To enable this, a symbol with even higher density than multi-staged symbols was required. As a result, QR Code, which can contain 7,000 digits of characters at maximum including Kanji characters (Chinese characters used in Japan) was developed in 1994 [see Soon (2008)].

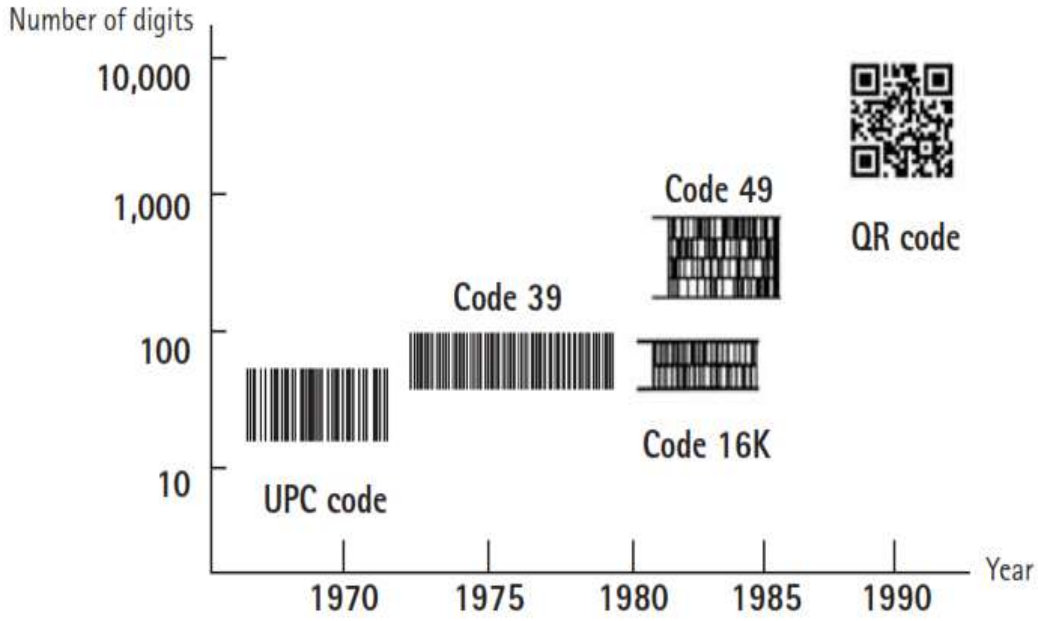


Figure 1: The history of symbols

1.2. Types of Two Dimensional Code

There are many types of 2-D Code, each of which, has its unique structure. The structure of the QR code is shown in Figure 2-(A), while, Figure 2-(B), depicts the structure of several other types of two-dimensional codes. Yet, QR codes are utilized most often for its higher data storage capacity. QR codes have become widely popular along with its useful applications. Its specifications offer many more advantages than that of the one-dimensional code. The Advantages of QR Code include:

- Reduced space.
- Durability against soil and damage.
- High data capacity.
- More supported languages.
- Supporting of 360-degree reading.

1.3. Application on QR Code

Based on the flexibility of QR code, there are many applications on QR Code, including:

- Mobile operating systems.
- Virtual stores.
- Payment.
- Website login, Wi-Fi network login and TOTP.
- Robot Navigation.
- Education.

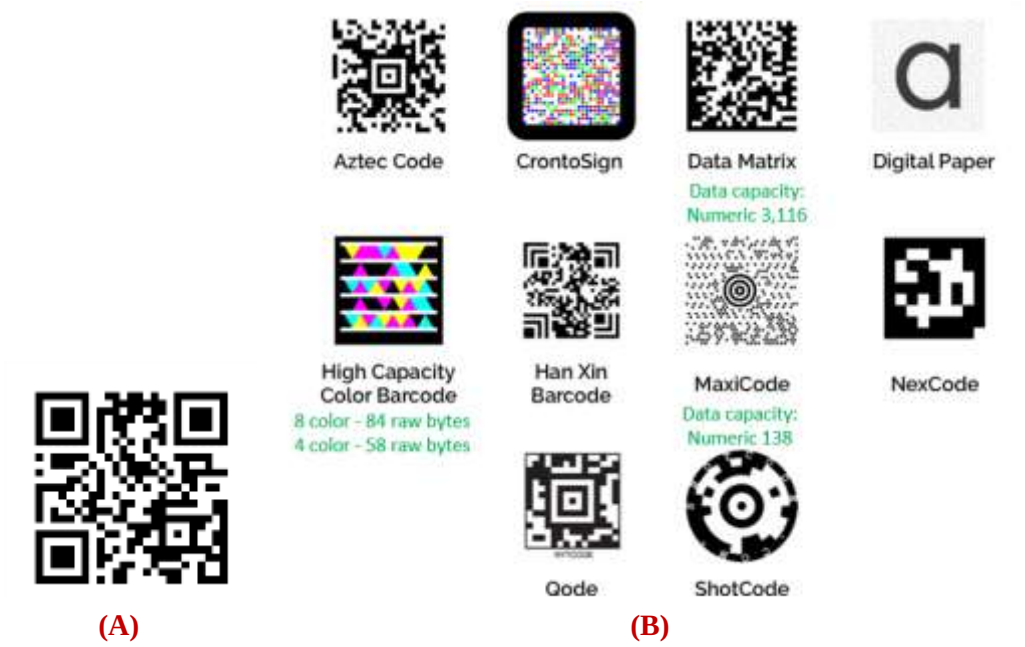


Figure 2: A. QR Code, B. and other type of 2D code

1.4. QR Code Structure

The image that represents a QR code is composed of three main constructions, namely: function pattern, encoding region and quiet zone.

1.4.1. Function Patterns

Must be placed in specific areas of the QR code to ensure that it can be correctly identified by QR code scanners. The function patterns are:

1. Finder pattern: are the special position-detection patterns located in three corners (upper left, upper right, and lower left) of each symbol.
2. Separators: They are the one-module wide areas of whitespace between each finder pattern and encoding region.
3. Timing patterns: There are 2 timing patterns (horizontal timing pattern and vertical timing pattern), they are consisting of alternating dark and light modules. These patterns are helpful in determining the symbol density, module coordinates and version information area.
4. Alignment patterns: They are a construction of three modules: A 5 × 5 dark modules, surrounding a 3 × 3 light modules and a single dark module in the center. QR codes, that are version 2 and higher, must have alignment patterns and the number of alignment patterns depends on the symbol version.

1.4.2. Encoding Region

Encoding region contains the information about the coded data, shown in Figure 3:

- Version information
- Format information: contains format error correction, mask pattern and error correction level (EC Levels) as illustrated in Figure 3, and detailed in Section 1.4.
- Error correction codewords: contains the encoded data after adding error correction codeword

1.4.3. Quiet Zone

It is a wide area containing no data. It occupies a width equivalent to four times that of the module.

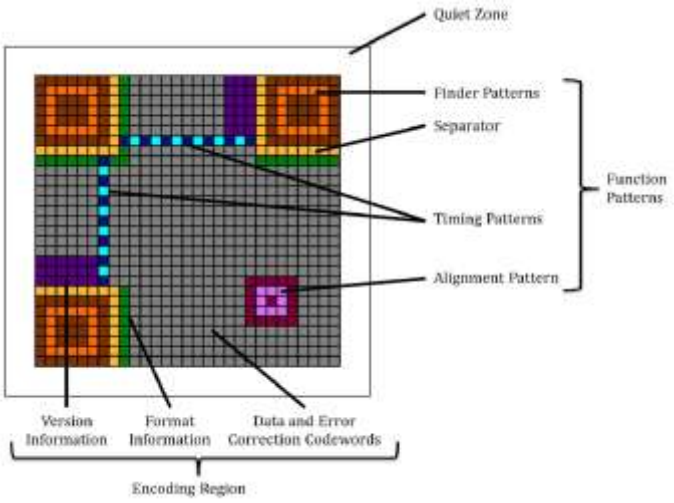


Figure 3: Structure of QR Code

1.5. Format Information

The left area in Figure 3, illustrates the left half region of the QR code structure. The format information is coded in three location:

- The vertical bar parallel to the upper left finder.
- The horizontal bar below the left finder pattern.
- The vertical bar parallel to the lower left finder pattern.

Error correction level:

Codewords are 8 bits long and use the Reed–Solomon error correction algorithm with four error correction levels. The higher the error correction level, the less storage capacity. Table 1 lists the approximate error correction capability at each of the four levels. For example, at the error correction Level M (Medium), 15% of codewords can be restored.

Mask Pattern:

It is used to break up patterns in the data area that might confuse a scanner, such as large blank areas or misleading features that look like the locator marks. The mask patterns are defined on a grid that is repeated as necessary to cover the whole symbol. Modules corresponding to the dark areas of the mask are inverted to white. Figure 4, illustrate the masking patterns.

1.5 QR Code Versions and Storage Capacity

The symbol versions of the QR Code may range from version 1 to version 40, each of which has a different module configuration or number of modules. The module refers to the black

and white dots that make up QR code [see Ahlawat et al. (2017), Tiwari (2016), Singh and Singh (2016), Kinjal et al. (2014)]. Table 2, illustrates the versions of QR Code

Table 1: Error correction levels and percentage of correction

S No.	Error-Correction Level	Approximate Amount of Correction
1.	L	7%
2.	M	15%
3.	Q	25%
4.	H	30%

1.6. QR Code Working Overview

The QR code system consists of a QR code encoder and decoder. The encoder is responsible for encoding data and generation the QR Code, while the decoder decodes the data from the QR code. Figure 5, shows an overview of the QR code working. The plain text, URL, or other data are given to the QR code encoder, generate the required QR code. A scanner (decoder) however, is utilized to decode QR code back to its original data.

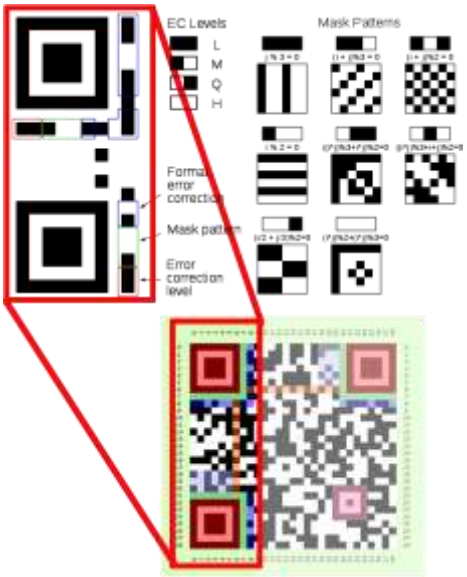


Figure 4: Format information

Table 2: Data Capacity of QR Code version 40

Version	Modules	ECC Level	Data Bits (mixed)	Numeric	Alpha-numeric	Binary	Kanji
40	177x177	L	23,648	7,089	4,296	2,953	1,817
		M	18,672	5,596	3,391	2,331	1,435
		Q	13,328	3,993	2,420	1,663	1,024
		H	10,208	3,057	1,852	1,273	784

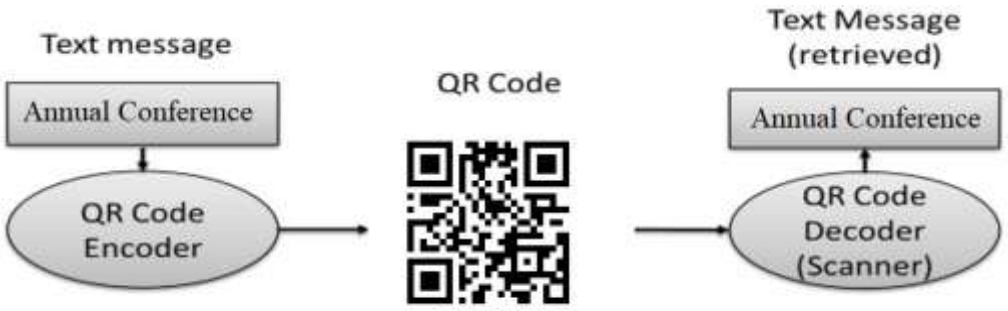


Figure 5: Working (overview) of QR Code

2. Improving QR Capabilities

There are many applications that focus in improving QR code data capacity and its tolerance to distortion. That is influenced by QR code structural flexibility. Some of these techniques focus on increasing data capacity including data hiding, data compression, Multiplexing and colored QR codes techniques [see Gupta et al. (2018)].

2.1 Data Hiding Techniques

Data hiding techniques distort the QR code during the hiding process, which can be resolved by using additional algorithms for the correction of image distortions. The achieved expanssion of the data capacity, however, is limited [see Paul et al. (2018)].

2.2 Data Compression Techniques

Another way to increase the data capacity is to compress the data before generating a QR code, results, however, shows that data compression technique offers compression up to 52% of the original size [see Abas et al. (2017)].

In this work, the authors employed colored coding for improving the QR code performance and capacity, therefore, the following Section will discuss the work related to colored QR code and color code Multiplexing.

3. Related Work

The research, in both Academia and Industry have been active to improve both the capacity and the tolerance of the QR code. The follwing are some of the related publications.

3.1 Coloured QR Codes "CQR Code-5"

Vizcarra [see Melgar et al. (2012)] purposed a colored QR code structure. It is designed to employ 5 different RGB colors (red, green, blue, black and white), which enables twice as much as that of the traditional binary QR codes. Each information module represents *2* bits, while the traditional utilize 1 bit for each module. Black modules are used only for alignment purposes. The "red, green, blue and white" colors are chosen because of their maximum equidistance on the RGB color space, as shown in Figure 6.

CQR Code-5 has the following advantages:

- Increasing of data storage capacity, Two times the traditional QR Code.
- Applying the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

CQR Code-5 Disadvantage is that it ignores the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting in capture process.

3.2 CQR Code-9:

Melgar [see Melgar et al. (2016)] proposed CQR Code-9. It can store up to *2,024* information bits which is twice that of CQR Code-5 and *4* times that of classic QR code. Each information module represents 3 bits.

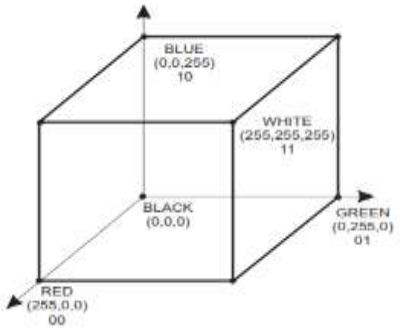


Figure 6: The equidistance on RGB color space

Advantages of CQR Code-9:

- Increasing of data storage capacity, Three times the traditional QR Code.

- applying the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

Disadvantage of CQR Code-9, though, is that it ignores the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting during the capture process.

3.3 Per-Colorant-Channel Color Barcodes:

Blasinki [see Blasinki (2013)], exploit the spectral diversity afforded by the cyan (C), magenta (M), and yellow (Y) print colorant channels, commonly, used for color printing and the complementary red (R), green (G), and blue (B) channels, respectively, used for capturing color images. Specifically, exploiting this spectral diversity to realize a three-fold increase in the data rate by encoding independent data in the C, M, and Y print colorant channels and decoding the data from the complementary R, G, and B channels captured via a mobile phone camera. Developing interference cancellation algorithm to mitigate the effect of cross-channel interference among the print colorant.

Advantages of Per-Colorant-Channel:

- Increasing of data storage capacity, Three times the traditional QR Code.
- Using the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.
- Avoiding interference between CMYK and RGB color systems.

Disadvantage of Per-Colorant-Channel is that it Ignores the color distortions caused by camera sensor and surrounding lighting during the capture process.

3.4 Multiplexing to Increase Information in QR Code:

Vongpradhip [see Vongpradhip (2013)] purpose Multiplexing QR Codes, in encoding the original data is divided into smaller parts; each part will form QR code pattern in its standard form. Each QR code pattern is, then, multiplexed and represented as module in the final QR code with black and white special symbols. The multiplexed QR code is decoded to give back the number of QR code patterns that was multiplexed.

In Decoding QR code with special symbols (that was multiplexed) is decoded to give back the number of QR code patterns that was multiplexed. These QR code patterns can be read by the general QR code reader and the data can be concatenated back to form its original information. The number of required symbols is defined by 2^n , where n is the number of QR code patterns.

The main advantage of Multiplexing is the increase of data storage capacity.

Disadvantages of Multiplexing are:

- Increasing of data storage capacity is limited by the number of available shapes that a pattern recognition algorithm can recognize.
- A pattern recognition algorithm is required to detect these black and white symbols which make decoding procedure very much different from that of the basic QR code.

3.5 Extending the Storage Capacity for Faster QR-Code:

Gupta [see Gupta (2018)] purpose new approach for extending the storage and faster QR decoding. Encoding: Convert part of data to normal QR code to generate 1st QR code, in which the black bits are one third instead of full of 1st QR code. Convert another part of data to normal QR code to generate 2nd QR code, in which black bits are instead of full of 2nd QR code, then rotate 2nd QR code in ***90*** degrees and place it on 1st QR code. The author performs the decoding by utilizing bar code decoding system, The horizontal and vertical regions are scanned and the black modules with size less than one third that of the standard are ignored.

Advantages of Gupta approach are:

- Increase of data storage capacity up to 2 times that of traditional binary QR Codes.
- Faster decoding than classic QR code.

Disadvantage of Gupta approach:

- Waste of some space, especially, and if data when data gets bigger then waste of space gets more visible. As shown in Figure 17.
- Unrecognizable, when the image's size is not big enough for the data It will not be recognized if data is large and its\s image is presented in a small area.
- More vulnerable to light glare and illumination issue.
- Is Require additional step (barcode decoding).

3.6 Multiplexing With Color Coding:

Galiyawala [see Galiyawala and Kinjal Pandya (2014)] intended *MUX* method in which several QR codes are mixed to generate QR code as shown in Figure 7. This technique offers improvement of the data storage capacity up to *24 times* compared to a standard QR code of the same size. In Encoding; Original information is divided into "*n*" parts, which are used to generate "*n*" individual QR codes, so it needs 2^n distinct encoding colors. In Decoding; the color value is demultiplexed using the table to get the three-separate classic QR codes.

Advantages of Multiplexing with Color Coding:

- Increasing of data storage capacity up to *24* times that of traditional binary QR Codes.
- Using the same localization and geometry correction algorithms for monochrome barcodes on the grayscale image.

Disadvantage of Multiplexing with Color Coding:

- Ignoring the color distortions caused by printer while printing QR code or by camera sensor and surrounding lighting during the capture process.
- The processing time for multiplexing and especially that for demultiplexing increased exponentially when merging more than *11* images.

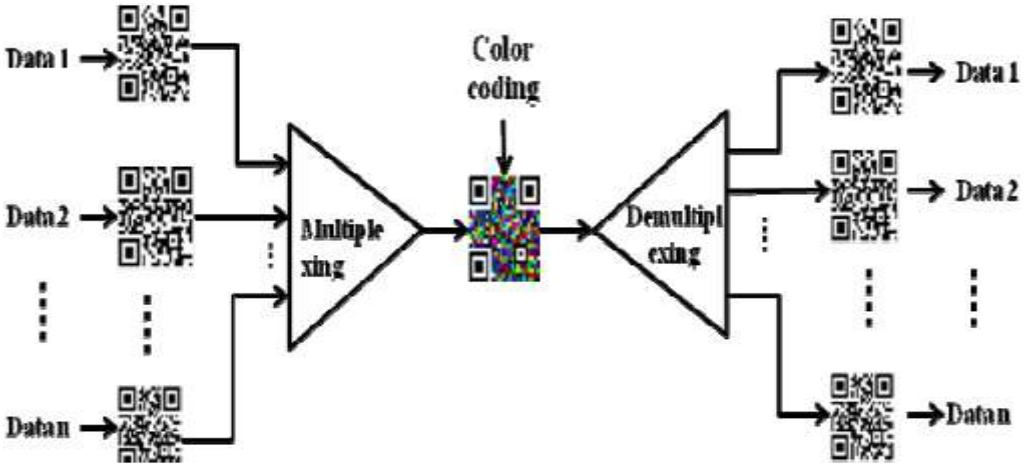


Figure 7: Block diagram of multiplexing with color coding technique.

4. Proposed System

One of the best ways for increasing data storage capacity of QR Code is by using color coding. The current state of art support this approach, yet, most of the current researchs ignored the accociated problems of colour distortion while printing QR code or by camera sensor and surrounding lighting during the capture process. Some of them, however, added enhancing algorithms to overcome some of the tolerance problems, which affected QR performance.

The main idea of the proposed system is multiplexing *n* of QR Code as three- Dimensional array of data, in which, the third dimension support the module's color, as shown in Figure 8. The modules in the virtical column (the thick red rectangle in Figure 8.(B) indicates a string of bits that is recognized by the decoding algorithm as an RGB colour code. The integrity of the color is guaranteed by including folding length in a predefined set of modules to be referenced in the decoding stage.

4.1 System Architecture

Figure 9 illustrates a block diagram for the proposed system.

4.2 System Implementation:

The two main algorithms involved in the system are the one used in the encoding and that for decoding of the QR colored code.

(A) **(B)**

Figure 8: (A) classic QR Code. **(B)** 8 levels of QR Codes.

"Encoding" psuedo code:

Input: data to be encoded and the parameters (version of QR code and Data size)
Output: a colored modules
Step 1. Divid data into *n* QR Code according to the input parameters.
Step 2. Print the value of *n* in the first data module in black and white.
Step 3. Calculate number of divisions required in *RGB* space.
Step 3. for each column of data module do
- **3.1 Convert value into decimal.**
- **3.2 Calculate number of steps in (*X, Y, Z*).**
- **3.3 Get equivalent color in *RGB* system.**
- **3.4 Print colored module.**

"Decoding stage" psuedo code:

Input: colored module and the *n* of QR Code.
Output: column of data module.
Step 1. Take input data (Array of colored module, *n* of QR Code).

Step 2. Calculate number of divisions required in RGB (X, Y, Z) space.
Step 3. For each colored module.
3.1 Calculate number of steps in (X, Y, Z).
3.2 calculate Decimal value for position.
3.3 calculate the corresponding Binary value.
3.4 Generate column of black and white data module.
Step 4. Generate *n* classic QR Codes.

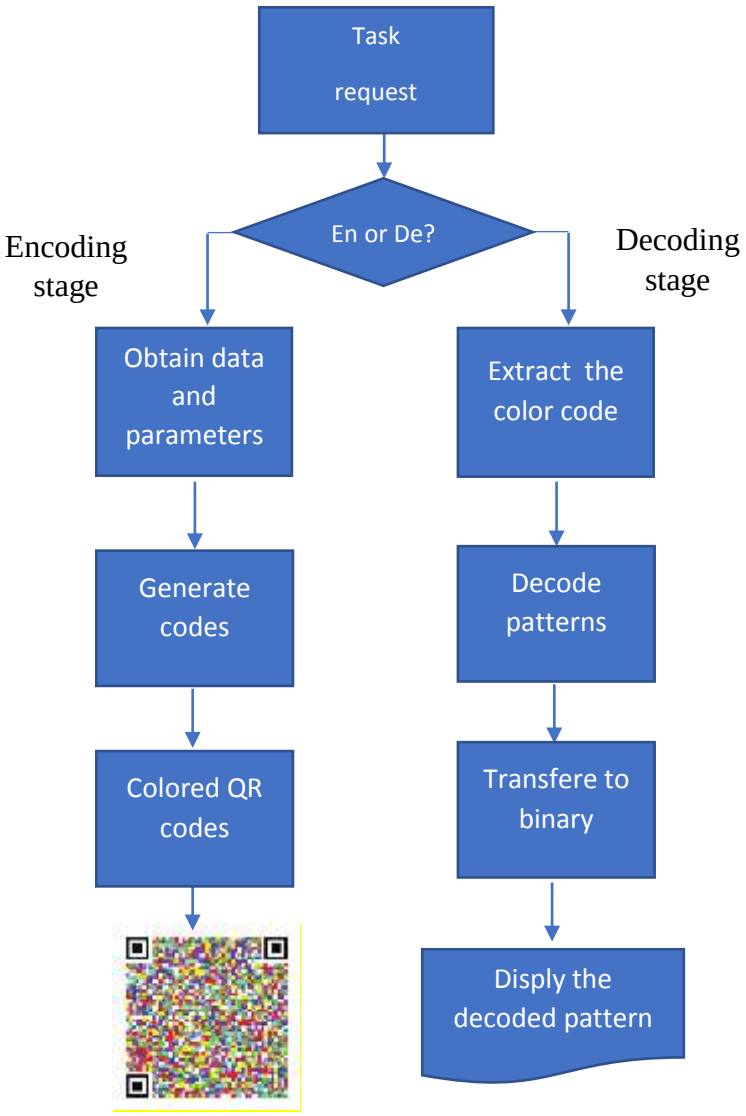


Figure 9: Block diagram for stage in the proposed system.

***n* is number of QR code which is determined in encoding stage and stored in QR code as black and white module at the first of data nodule**

5. System Performance Evaluation:

An experiment was performed to test the proposed system performance as follows: A randomly generated information of typical size was applied as input to generate a QR color codes for different QR code level (2, 3, 4,…, 14). The system assumed version number 40-L (of the QR code which has 177 X 177 modules) as general parameter, and occepts the folding as a special parameter for each QR code partition.
Figure 10 and Table 3, illustrate the time performance of the proposed system. A very interesting observation is that decoding time is almost equals to that of encoding time for all the QR code folds except those greater than 5 and less than 12 folds, during which the decoding consumes lesser time.

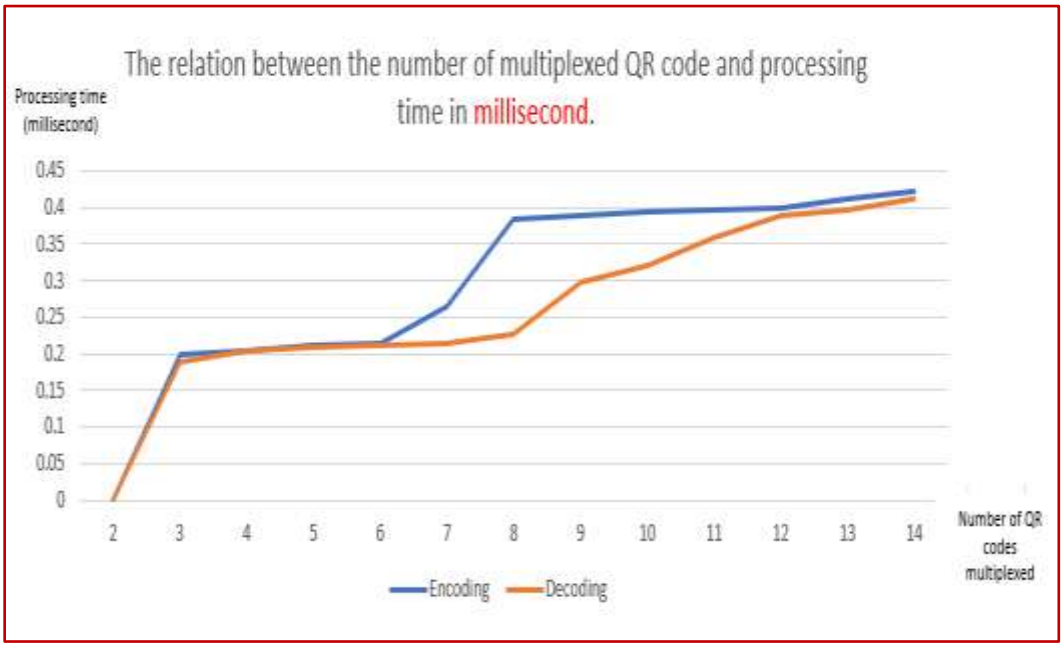


Figure 10: Time performance of the proposed system for version 40-L QR code

Table 3: Processing time for different QR sizes of the proposed system,

No.	Number of QR codes multiplexed	Assigned distinct colors(2^n)	Encoding time Millisecond	Decoding time Millisecond
1	2	4	0	0
2	3	8	0.198	0.188
3	4	16	0.205	0.203
4	5	32	0.212	0.209
5	6	64	0.213	0.211
6	7	128	0.265	0.215
7	8	256	0.383	0.226
8	9	512	0.389	0.298
9	10	1024	0.393	0.320

10	11	2048	0.396	0.359
11	12	4096	0.400	0.389
12	13	8192	0.411	0.396
13	14	16384	0.423	0.412

6. Comparison with the Classic Multiplexing Color Coding System

Table 4 shows the execution time of the clasical system [Galiyawala 2014] for multiplexing number of QR codes. While, Figure 11 shows the the encoding and decoding execution time for Muliplexing various sets of QR codes in the classical system of Galiyawala. The results of Galiyawala 2014 were ploted along with the results of the same set of QR code folds of the proposed system, as depected in Figure 12. The Figure illustrates the time performance of the proposed system versus that of the Classical Multiplexing system [Galiyawala 2014]. The plot was built with the time in milliseconds, yet, the results of the classical system was scaled down using $\log_{10}$ of the original values, due to the exponintial differences in the corresponding values.

Table 4: Processing time of colored QR Multiplexing [Galiyawala 2014]

Sr. No.	Number of QR codes multiplexed	Assigned distinct colors (2^n)	Processing time (seconds)	
			During Encoding	*During Decoding*
1	2	4	2.922	2.892
2	3	8	3.194	2.988
3	4	16	3.364	2.982
4	5	32	3.131	3.084
5	6	64	3.297	3.381
6	7	128	3.489	3.624
7	8	256	3.911	4.359
8	9	512	4.811	6.128
9	10	1024	6.229	11.406
10	11	2048	9.453	28.398
11	12	4096	15.787	91.393
12	13	8192	27.940	323.198
13	14	16384	53.157	1236.105

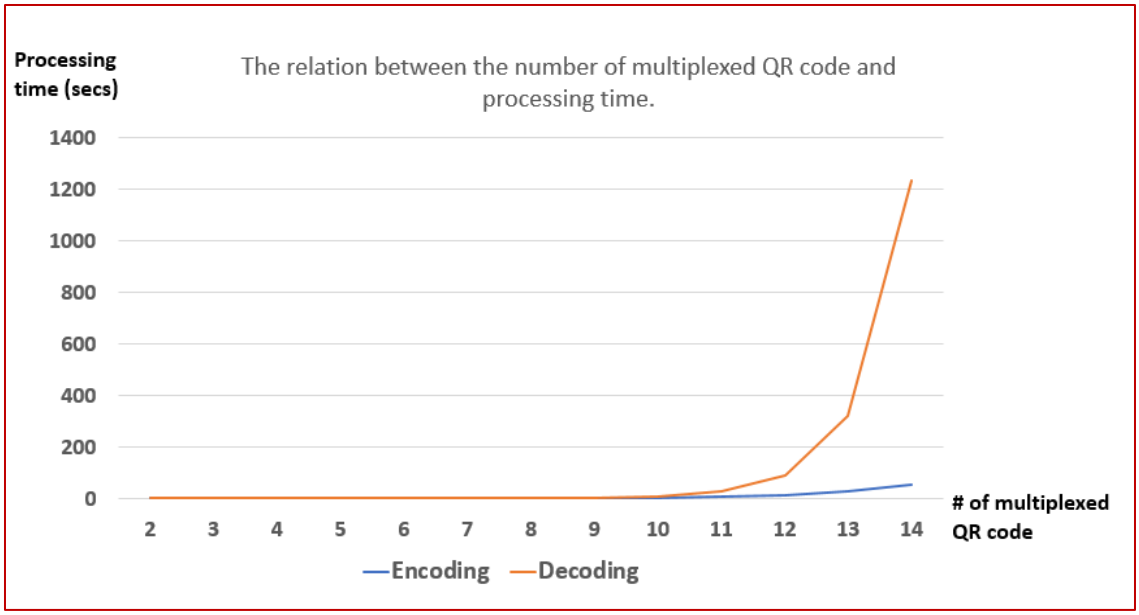


Figure 11: Performance of the classical system [Galiyawala 2014]

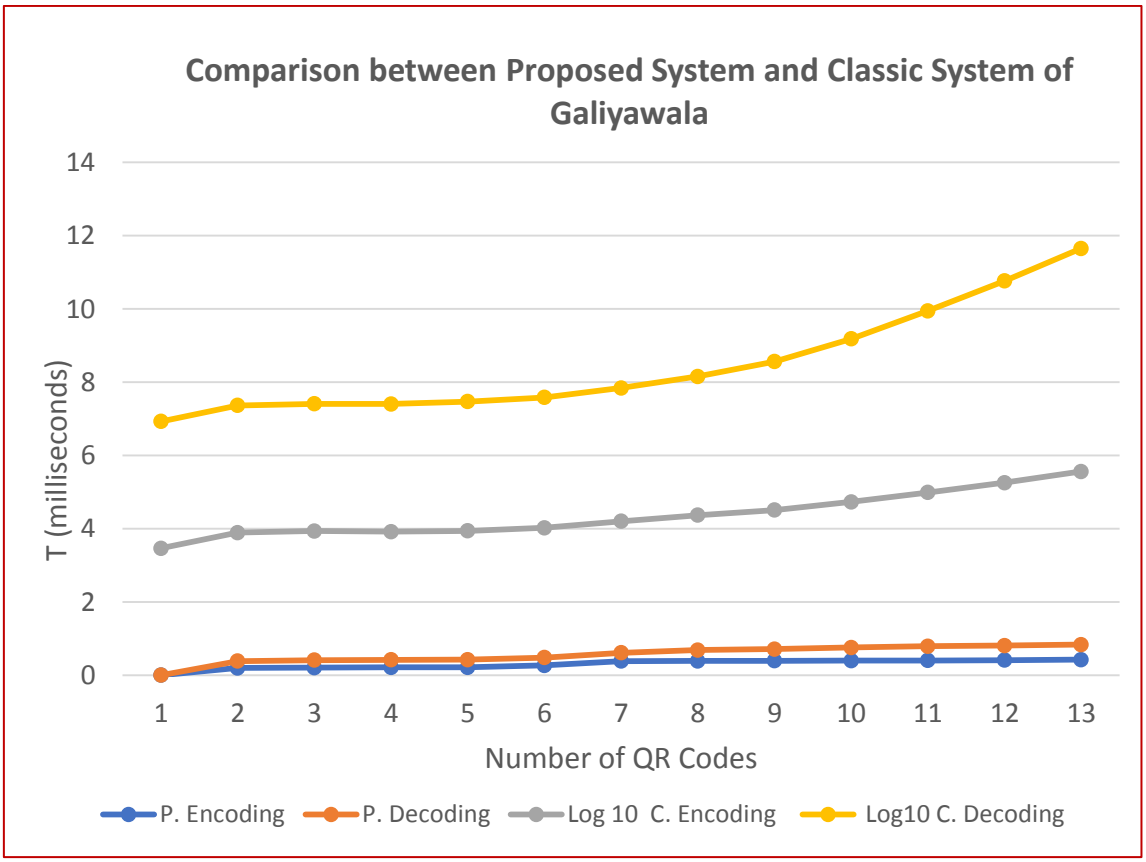


Figure 12: Performance comparison between the proposed and the classical system of Galiyawala

7. Conclusion

This paper introduced the quick response (QR) code and its applications with the possibilities of its utilization in different areas that is yet to be explored. The technology has a firm ground for research aspects. More and more experiments are done with QR Codes in different aspects like enhancing the storage capacity, better recognition, reducing redundancy in order to save space, possibility of encoding different kind of data. The paper established a system to improve the performance of the color coded QR. The system based on utilizing a color scheme extension as a third dimenssion of the QR code. The proposed system proved to be very efficient, and the processing time depends, only on the number of folds, (information partitions). The proposed system has the same property of multitude the capacity of the QR code, (as that of the classical), Galiyawala, with the added feature of maintaining the undistorted modules' color.

References

Abas A., Yusof Y. and Kabir F., 2017, "Expanding the data capacity of QR codes using multiple compression algorithms and base64 (encode/decode)", Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9, no. 2-2, pp. 41-47.

Bhardwaj N., Kumar R., Verma R., Jindal A. and Bhondekar A. P., 2016, "Decoding algorithm for color QR code: A mobile scanner application", International Conference on Recent Trends in Information Technology (ICRTIT), India, pp. 1-6.

Goyal S., Yadav S. and Mathuria M., 2016, "Exploring concept of QR code and its benefits in digital education system", International Conference on Advances in Computing, Communications and Informatics (ICACCI), India, vol. 3, issue 5, pp. 452-456.

Gupta K. D., Ahsan M. and Andrei S., 2018, "Extending the storage capacity and noise reduction of a faster QR-code", Broad Research in Artificial Intelligence and Neuroscience (BRAIN), vol. 9, issue 1, pp. 59-71.

Henryk Blasinski, April 2013, "Per-Colorant-Channel Color Barcodes for Mobile Applications: An Interference Cancellation Framework", vol. 22, no. 4, pp. 1498-1511.

Hiren J. Galiyawala and Kinjal H. Pandya, 2014, “To Increase Data Capacity of QR Code Using Multiplexing with Color Coding: An example of Embedding Speech Signal in QR Code”, IEEE India Conference (INDICON), pp. 1-6.

Kinjal H. Pandya and Hiren J. Galiyawala, "A survey on QR codes: in context of research and application” , 2014, International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol. 4, issue 3, pp. 258-262.

Kishor Datta Gupta, Md Manjurul Ahsan and Stefan Andrei, 2018, “Extending the Storage Capacity and Noise Reduction of a Faster QR-Code”, Lamar University, Beaumont, USA pp. 59-71.

Max E. Vizcarra Melgar, Alexandre Zaghetto, Bruno Macchiavello, Anderson C. A. Nascimento, 2012, "CQR CODES: COLORED QUICK-RESPONSE CODES", IEEE Second International Conference on Consumer Electronics - Berlin (ICCE-Berlin), pp. 3321-325.

Max E. Vizcarra Melgar, Mylène C. Q. Farias, Flávio de Barros Vidal, Alexandre Zaghetto, 2016, “A High Density Colored 2D-Barcode: CQR Code-9”, SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 329-334.

Remya Paul, 2018, "A review on stegnography in QR codes", International Research Journal of Engineering and Technology (IRJET), vol. 5, issue 5, pp. 4294-4296.

Rizqi R. I., Rohama N. A. and dr. Nimkerdphol K., 2018, "Inventory management system using QR code on android a case Study in computer engineering department", Journal of Electrical Engineering and Computer Sciences (JEECS), vol. 3 no. 1, pp. 381-388.

Sartid Vongpradhip, 2013, “Use Multiplexing to Increase Information in QR Code”, April 26-28, 2013. Colombo, Sri Lanka, pp. 361-364.

Seema Ahlawat et al., 2017, "A review on QR codes: colored and image embedded" International Journal of Advanced Research in Computer Science (IJARCS), vol. 8, no. 5, pp. 410-413.

Singh A. and Singh P., 2016, "A review: QR codes and its image pre-processing method", International Journal of Science, Engineering and Technology Research (IJSETR), vol. 5, no. 6, pp.1955-1960.

Sumit Tiwari, 2016, "An introduction to QR code technology", IEEE International Conference on Information Technology (ICIT), pp. 39-44.

TJ Soon, 2008, "QR Code", Synthesis Journal.

Zlatko Čović, Ürmös Viktor, János Simon, Dalibor Dobrilović, Željko Stojanov, 2016, "Usage of QR Codes in Web Based System for the Electronic Market Research", 14th IEEE International Symposium on Intelligent Systems and Informatics (SISY), Serbia, pp. 187-192.

Recent Approaches for Automated Glaucoma Diagnosis in Retinal Images

Osama M. Kamara[1] Ahmed H. Asad[2] Hesham A. Hefny[3]

Abstr

Accurate automated system for glaucoma diagnosis is very important since glaucoma is the second disease that leads to vision loss. Therefore, it is necessary to utilize all computational intelligence approaches in order to accurately detect glaucoma at its early stages in retinal images. The most significant sign of glaucoma in retinal images, which help doctors for indicating if patient has normal or glaucomatous image, is the ratio of optic cup diameter to optic disc diameter. So, it is important to segment both of the optic disc and optic cup regions in retinal images. This paper presents the most recent approaches of glaucoma detection that were proposed in the last few years where these approaches were categorized into four categories based on image processing, machine learning, swarm intelligence and deep learning.

Keywords: Glaucoma diagnosis; Image processing; Machine learning; Swarm intelligence; Deep learning.

1. Introduction

Glaucoma is a chronic eye disease in which the optic nerve is progressively damaged. It is the second leading cause of blindness, and is predicted to affect around 80 million people by 2020, Quigley et al (2006), The progression of glaucoma disease leads to loss of vision, which occurs gradually over a long period of time. As the symptoms only occur when the disease is quite advanced, glaucoma is called the silent thief of sight. Glaucoma cannot be cured, but treatment can slow down its progression. Therefore, detecting glaucoma in time is critical, Costagliola et al (2009). There are three methods to detect glaucoma: 1) Assessment of raised intraocular pressure (IOP), 2) Assessment of abnormal visual field, 3) Assessment of damaged optic nerve head.

For the first detection method, the IOP measurement using noncontact tonometry (also known as the "air puff test") is neither specific nor sensitive enough to be an effective screening tool because glaucoma can be present with or without increased IOP. For the second method, a functional test through vision loss requires special equipment only present in tertiary hospitals and therefore unsuitable for screening. For the third detection method, the assessment of the damaged optic nerve head is both more promising, and superior to IOP measurement and visual field testing for glaucoma screening, Almazroa et al (2015). Although optic nerve head assessment can be done by a trained professional but manual assessment is subjective, time consuming and expensive. Therefore, automated optic nerve head assessment would be very beneficial. One strategy for automated optic nerve head assessment is to use image features for a binary classification between glaucomatous and healthy subjects.

1. Master student at department of computer science, faculty of graduate studies for statistical research, Cairo University, Egypt.

2. Assistant Professor at department of computer science,faculty of graduate studies for statistical research, Cairo University, Egypt.

3. Professor at department of computer science, faculty of graduate studies for statistical research, Cairo University, Egypt.

These features are normally computed at the image-level where the selection of features and classification strategy is difficult and very challenging. The other and more common strategy is based on clinical indicators, Nicolela et al (2016), such as the vertical cup-to-disc ratio (CDR), disc diameter, peri-papillary atrophy (PPA), as well as inferior, superior, nasal, and temporal (ISNT) zones rule in the optic disc area, Chrástek et al (2005). Although different ophthalmologists have different opinions on the usefulness of these factors, CDR is well accepted and commonly used where a larger CDR indicates a higher risk of glaucoma. The optic nerve head or the optic disc (OD), in short disc, is the location where 1.2 million ganglion cell axons exit the eye to form the optic nerve, through which visual information of the photo-receptors is transmitted to the brain. So, examining the OD helps clarify the relationship between the optic nerve cupping and loss of visual field in glaucoma, Harizman et al (2006). In 2-D images, the OD can be divided into two distinct zones; namely, a central bright zone called the optic cup (OC), in short cup, and a peripheral region, called the neuro-retinal rim. Figure 1 shows the major structures of the OD. The CDR is computed as the ratio of the vertical cup diameter (VCD) to vertical disc diameter (VDD) clinically. So accurate segmentations of OD and OC are essential for CDR measurement.

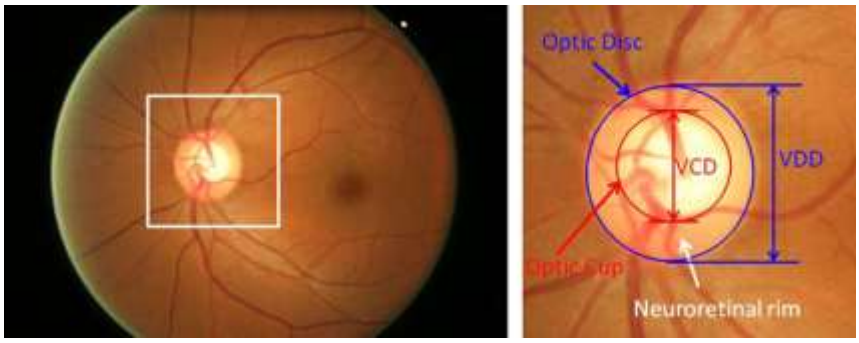


Figure 1: *Optic disc structure. The region enclosed by the blue line is the optic disc; the bright central region enclosed by the red line is the optic cup; and the region between the red and blue lines is the neuro-retinal rim.*

The remaining of this paper consists of four sections. Section two describes the metrics and datasets used to evaluate the performance of automated glaucoma detection approaches. Section three presents these approaches with discussion about them is provided in section four. Finally, the conclusion is in section five.

1. Performance Metrics and Datasets

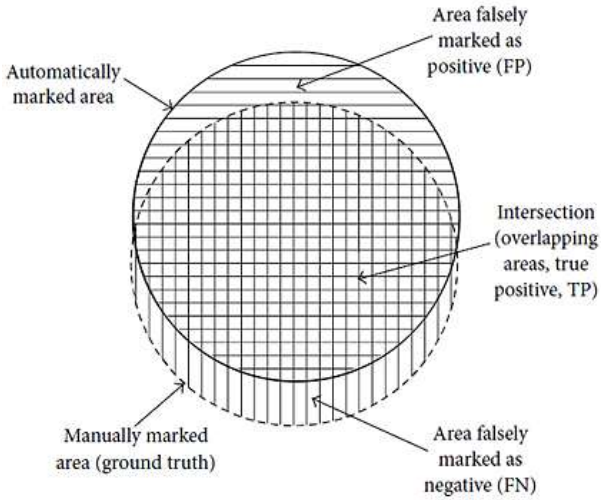


Figure 2*: the relation between ground truth and predicted area*

2.1 Metrics: There are different metrics that are used to measure the approaches performance for OD and OC segmentation in retinal image and determine if the image is normal or glaucomatous. The most used metrics are CDR error, sensitivity (SN), specificity (SP), accuracy (ACC), dice metric (DM), F-Score and AUC as shown in table 1. These metrics are calculated based on the relationships between the ground-truth area marked manually by an expert ophthalmologist and the segmented area marked automatically by an approach. The CDR error is the absolute difference between the ground-truth and predicted CDRs. The SN, SP, ACC and DM and F-Score are based on the calculation of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) areas. The TP area is the overlapping area between the manually marked ground-truth and automatically marked segmented areas while the FP area is the area that is classified only in the automatically marked segmented area. The FN area is the area that is classified only in the manually marked ground-truth area while the TN area is the area that is not classified in by both of the manually marked ground-truth and automatically marked segmented areas.

Table 1: *The most used metrics to measure optic disc and optic cup segmentation performance*

Metric	Equation	Description
CDR error	$\lvert CDR_{\text{ground-truth}} - CDR_{\text{pred}} \rvert$	Indicate the absolute difference between ground-truth and automatically predicted area
Sensitivity (SN)	TP / (TP + FN)	Indicator for the probability of glaucomatous classes that identified as glaucomatous
Specificity (SP)	TN / (TN + FP)	Indicator for the probability of normal classes that identified as normal
Accuracy (ACC)	(TP + TN) / (TP + FP + FN + TN)	Indicator for the performance and quality of an approach
Dice Metric (DM)	2*TP / (TP + FP + FN)	Indicator for the similarity between automatically predicted area and ground-truth area.
F-Score	2*TP / (2*TP + FP + FN)	Indicator for the accuracy in range of 0-1, it is better quality of performance when F-score is closed to 1.
AUC	(SN + SP) /2	Indicator for the classification performance with value of range 0-1, it is good classification when AUC value is closed to 1.

2.2 Datasets: Although there are various benchmark retinal images datasets available to evaluate the automated OD and OC segmentations approaches, but some researchers evaluated their approaches on locally collected datasets while others didn't mention, not available (NA), what is dataset they used to evaluate their approaches. In this subsection, we will brief a short description only for each benchmark dataset mentioned in this paper, and we mentioned some retinal images of these datasets in figure 3.

- ORIGA-light dataset, Online Retinal Fundus Image Database for Glaucoma Analysis and Research, Zhang et al (2010), contains 650 retinal images divided to 168 glaucomatous images acquired through Singapore Malay Eye Study (SiMES), Foong et al (2007).
- HRF dataset, public dataset of 45 retinal images at this moment contains 15 healthy images, 15 glaucomatous images and 15 diabetic retinopathy images, Budai et al (2013).
- Rim-One dataset is a database of 159 stereo fundus images with two ground truth of disc and cup collected at the Hospital Universitario de Canarias and divided by the experts into 85 healthy images and 74 glaucomatous images, Fumero et al (2015).
- SCES (Singapore Chinese Eye Study) dataset contains 1676 retinal images including 46 glaucomatous images, Hilal et al (2013).
- Drishti-GSI dataset, Sivaswamy et al (2014), contains 101 retinal images available online

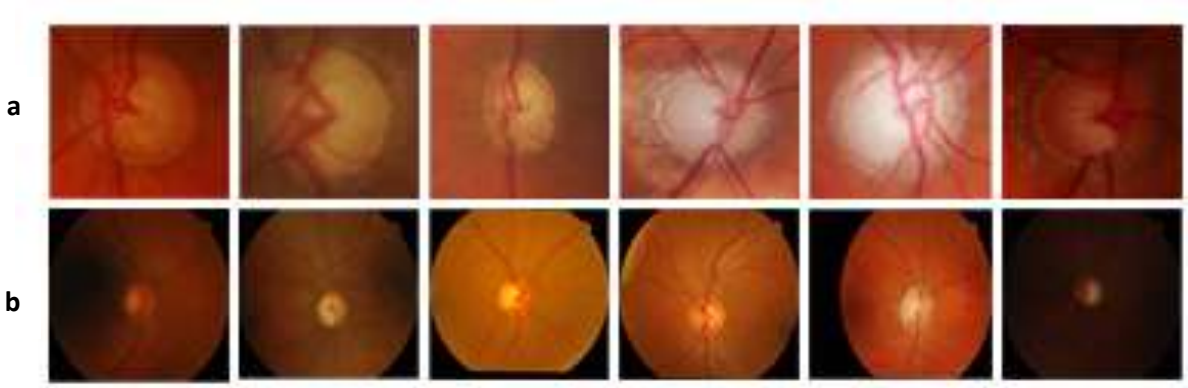


Figure 3: *retinal images from (a) Rim-one-1 dataset and (b) Drishti dataset*

3. Glaucoma Diagnosis Approaches

There are three basic steps in any automated glaucoma diagnosis system in retinal images. The first step is the preprocessing of retinal image to compensate for non-uniform illumination and to remove the distracting blood vessels. The second step is to localize the optic disc and hence optic cup for segmenting them. The third step is optic disc and optic cup segmentations by refining their boundaries for calculating the CDR or ISNT rule to diagnose glaucoma. In this section, the most recent automated glaucoma diagnosis approaches are categorized into four categories based on their underlying methods. The first category is image processing-based approaches, the second category is machine learning based approaches, the

third category is swarm intelligence-based approaches and the fourth category is deep convolutional neural network (CNN) based approaches.

3.1 Image Processing Based Approaches:

Table 2: *Glaucoma diagnosis approaches based on image processing*

Authors	Image Processing Method	Performance Metrics	Dataset	Number of Images
Hatanaka et al (2011)	Pixel Line Profile Analysis	CDR error: 11	Local	45
Yin et al (2012)	Statistical model that combines Hough Transformation and novel optimal channel selection	DM OD: 0.92 DM OC: 0.81	ORIGA	650
Hatanaka et al (2014)	Blood vessels bends detection based method	SN: 78.1% SP: 91.7%.	Local	94
Sedai et al. (2016)	Novel cascaded shape regression method	F-Score OD: 0.95 F-Score OC: 0.85	Local	50
Shyam et al (2016)	High pass filtering with morphological operations	SN: 0.742 SP: 0.943 ACC: 0.931	HRF	10
Natraj et al (2017)	Unique template-based correlation using Pearson-r coefficients and isotropic wavelet transform	SN: 0.86 SP: 0.86	HRF	10

Hatanaka et al (2011) proposed line profile analysis based approach in order to measure CDR, first step is to remove blood vessels then because disc is more brighter he uses the P-tile threshold method in order to extract optic disc then apply Canny Edge Detector on intensity image to enhance the edge then determine and extract the optic disc outline using the Spline Interpolation method. The second step is to calculate CDR depending on the optic disc profile of normal against glaucomatous cases, whereas normal cases profile appears as narrow with long skirts while glaucomatous profile is a broad with long skirts. This approach was tested on 45 retinal images including 23 glaucomatous images and disc extraction was 85% while rate of cup region was 67% because this approach failed to remove blood vessels in nasal side and the average error of abnormal and normal cases were 11 and 10 respectively.

Yin et al (2012) proposed statistical model-based approach that utilizes Circular Hough Transform with a Novel optimal selection for optic disc and optic cup segmentation. First for shape and appearance modeling, Yin used 24 landmark points which each adjacent point should give angle of 15o degrees with optic disc center in order to build the shape model. In order to reduce blood vessels effects Yin proposed a Novel optimal channel selection to select the best image "according to image contrast ratio R" with the least blood vessels information and avoid any evenly distributed intensity for accurately segmentation. The next step is to initialize the deformable model by estimating the optic disc position and size using the Canny Edge Detector then uses Circular Hough Transform to smooth the OD boundary, now is to extract OD boundary, so each landmark point will be moved towards the contour perpendicularly to its new position in iteratively way and finally apply the Direct Ellipse Fitting Method as smooth

algorithm. As Yin segmented OD in previous steps, similarly Yin proposed the same to segment OC: he apply active shape model to get the appearance model and segment OC in green channel as region-of-interest (ROI) and detect blood vessels using Local Entropy Thresholding then remove it by replacing each pixel in blood vessels detected by the median of intensity values in its neighborhood then use Median Filter to smooth it, then use Active Shape Model in green channel to extract OC boundary. Yin tested his proposed approach on ORIGA-light dataset which contain 650 images include 128 images are glaucomatous; Yin used 325 images as train and the other 325 images as test. Finally, the results measured in Dice Metric for OD is are 0.92 and for OC is 0.81 with error of 0.19.

Hatanaka et al (2014) improved his previous approach by proposing blood vessel bends detection based approach for automated OC segmentation process which consists of 3 steps, 1) OD segmentation process by first removing blood vessels then apply Canny Edge Detector on red channel and select 48 initial points on disc edge then apply Active Contour Model (ACM) to update points and then in order to extract disc outline he uses the Spline Interpolation method. 2) Optic Cup segmentation process, first get an initial OC outline by applying Pixel Line Profile Analysis method then detect blood vessels based on the concentration features determined from the density gradient and apply P-tile method and then track the blood vessel bends from disc outline to initial cup outline and divided bends into visible bends detected by the K-Curvature and invisible bends looked as endpoints on OD, and then apply the Spline Interpolation method to extract the Cup. 3) Final step is to calculate CDR. This approach was tested on 50 retinal images and sensitivity was 78.1% and specificity was 91.7%.

Sedai et al (2016) proposed Novel cascaded shape regression approach which OD and OC are represented as collection of landmark points then apply a regression function. First localize OD using its properties of yellowish color and OD has the maximum number of blood vessels in order to find ROI. Before to apply Insert Point Detection there is preprocessing to enhance blood vessels by applying Multiscale Line Detection Filter in green channel then Otsu's Thresholding to segment blood vessels then compute the mean intensity and vessels density inside the sliding window in order to select the window of OD area. Then apply Canny Edge operator to get binary image that contains the major intensity gradients and then apply Circular Hough Transform, now OD circle is detected and compute the 2 insert points of optic disc (0o degree on temporal side and 180o degree on Nasal). Finally, iteratively generate a collection of initial shapes based on PDM and updated by Shape Increment Vector predicted by each regressor. The proposed approach has been trained and tested on 50 patients and the results measured by F-Score are 0.85 for optic cup and 0.95 for optic disc. This approach should be tested on more images.

Shyam et al (2016) has proposed High Pass Filtering based approach to detect glaucoma by means of ISNT ratio. First, they extract green channel then perform contrast and vessels enhancement and then apply Low Pass Filter which vessels corresponds to high frequency components then the high pass filtered image is subtracted then thresholded to get pixels less than 0, then negative the enhanced image in order to perform Tophat reconstruction then apply

mask to filter each quadrant then calculate the ISNT ratio, the accuracy of this method is 0.931 and it was tested on HRF dataset which has 10 retinal images.

Natraj et al (2017) proposed a system used Neurortinal rim thickness and blood vessels in order to improve sensitivity. Natraj has proposed new feature named Vessel Risk Index (VRI) by measuring number of vessels and their diameters. First, they apply preprocessing to fundus images in order to remove noise using an Adaptive Weiner Filter on gray channel then uses Pearson-r Coefficient in order to extract image features and segment OD and OC then calculate CDR and RDR, then classify images to glaucomatous using values of CDR, RDR, superior and inferior neuroretinal rim thickness and VRI. This approach was tested on HRF dataset and local retinal images and the result while using all features in the methodology was great since sensitivity and specificity was 86.67 and 86.67 respectively but results while using feature of CDR only was 80%

3.2 Machine Learning Based Approaches:

Table 3: Glaucoma diagnosis approaches based on machine learning

Authors	Machine Learning Method	Result	Dataset	Number Of Images
Xu et al (2012)	Superpixel classification using support vector machine rely on prior knowledge	CDR error: 8.1	ORIGA	650
Cheng et al (2013)	Superpixel classification combined with deformable model	OD overlapping error: 10% OC overlapping error: 26%	SIMES SCES	650 1676
Xu et al (2014)	Low rank superpixel classification	CDR error: 7.9	ORIGA	650
Bechar et al (2018)	Semi-supervised superpixel classification	CDR error: 9.84	Rim-One	159
Deepika el al (2018)	Prepeocessing & blood vessels features classification	SVM accuracy: 91.67% ANFIS accuracy: 97.77%	HRF	67
Jain et al (2019)	Two-dimensional tensor empirical wavelet transforms and SVM classification	Accuracy: 98.7%	Local	505

Xu et al (2012) proposed a superpixel classification approach based on the Simple Linear Iterative Clustering (SLIC) method that rely on prior information from the test image without needing pre-labeled training set to segment optic disc into 512 superpixels. This method firstly removed blood vessels by applying the Bottom-Hat filtering algorithm. Secondly, the feature extraction step is applied which include position information and mean RGB colors where 256-bin histogram of each color channel is built. After that, a support vector machine (SVM) is used to perform superpixel classification process relying on prior knowledge of optic disc structure such that cup is considered within 1/5 of the disc radius. The SVM classifier is trained to label each superpixel if it is belonging to rim or cup using LIBLINEAR toolbox, and the labeling is refined depending on it is near to or far from a superpixel's position with respect to the OD center. Finally, CDR calculation step is performed. This approach was tested on ORIGA-light dataset with result 8.1 of CDR error.

Cheng et al (2013) proposed superpixel classification (Simple Linear Iterative Clustering - SLIC) based method combined with the deformable model based method in order to overcome the limitation of pixel classification based method. First, they applied feature extraction process by performing: a) histogram equalization to five color channels red, green, blue, hue and saturation. b) Center Surround Statistics (CSS) and generating Gaussian pyramids filter on multi scales (8 levels) with ratio from 1:1 (level 1) to 1:256 (level 8) to overcome the similarity of PPA region and optic disc. Second, they processed the initialization and deformation in order to get superpixel decision values by using SVM with linear kernel. To achieve smoothed decision values and get the binary decision, they applied the mean filter as smoothing with a threshold that is the average of values with assign (+1) as disc and (-1) as background or non-disc. Third, they applied fine tune to disc using elliptical Hough transform to obtain disc boundary. These previous three steps was performed to segment optic disc and in order to segment optic cup, they used similar method which is superpixel classification method that rely on prior knowledge in training instead of relying on blood vessel bends. So they applied the feature extraction process and computed histogram feature without red channel because it has little information about optic cup and applied the center surround statistics (CSS) with reservation of including distance between superpixel center and optic disc center in order to get accurate results if pallor is nonobvious. After that, they applied superpixel classification for optic cup estimation using the LIBSVM with linear kernel and applied the mean filter to smooth the output decision values in order to obtain binary decisions for all pixels and connect the largest objects to obtain OC boundary. The final step was calculating the CDR from the segmented OD and OC regions. This proposed method was tested on 2326 retinal images of two datasets (SIMES dataset which has 650 images includes 168 glaucomatous images and SCES dataset which has 1676 images includes 46 glaucomatous images), and the mean overlapping error of segmentation for disc was 10% and for cup was 26%.

Xu et al (2014) proposed a method as an improvement to their previous method, Xu et al (2012) by keeping all steps as them except that they formulated the labeling problem as a Low Rank Representation (LRR) and its selection parameter were determined automatically. This approach was tested on ORIGA-light dataset with result 7.9 of CDR error which is very closed to results of their previous approach, specifically their previous approach used the prior structure as well.

Bechar et al (2018) proposed a semi-supervised superpixel approach which has three main steps: 1) classify data to a few amount of labeled and a large amount of unlabeled using SLIC algorithm and the expert annotation. 2) feature extraction from each superpixel including different five color spaces (RGB, HSV, HSL, Luv, YUV) and spatial information of the OD and OC. 3) training the Co-forest classifier with few labeled and large number of unlabeled superpixels to generate a robust classifier then finally smoothing the segmented optic disc and optic cup by applying Active Geometric Shape Model (AGSM) to calculate the CDR. This approach was tested on RIM-ONE retinal fundus image database include 35 glaucomatous images, with results of CDR accuracy 90.16.

E.Deepika et al (2018) proposed classification and segmentation method based on blood vessels features, first he enhanced fundus images by applying median filter, wiener filter then to extract green channel and use CLAHE to extract blood vessel features then to apply each of SVM and ANFIS classifiers separately to detect fundus images if it is normal or abnormal and finally this method get accuracy as following: for SVM classifier accuracy, Sensitivity and Specificity are 91.67%, 90% and 93.33% respectively , and for ANFIS classifier accuracy, Sensitivity and Specificity are 97.77%, 85.7% and 92% respectively. This approach was tested on HRF database using 35 images as train set and 31 images as test set.

Jain et al (2019) proposed classification and segmentation method based on tensor empirical wavelet transform (tensor based EWT) features extraction and to classify using SVM classifier. This approach used Local dataset contains 250 glaucomatous images and 255 normal images. Its accuracy was 98.7%.

3.3 Swarm Intelligence Based Approaches:

Table 4: Glaucoma diagnosis approaches based on swarm intelligence

Authors	Swarm Intelligence Method	Result	Dataset	Number of Images
Arnay et al (2017)	Ant Colony Optimization using prior knowledge	OC overlapping error: 24.3%	Rim-One	159

Arnay et al (2017) proposed Ant Colony Optimization (ACO) based approach that has an advantage of being capable of accurately segment optic cup even if image has noise or weak pallor. The segmentation problem is formulated as optimization problem in order to find the shortest path of graph G. First Arnay defined the inner and outer exploration area using circle of radius rmin and rmax and its centers belong to Optic Disc, then agents start exploring from a seed point and stop if agent reaches a seed point again. The step of defining heuristic function uses the prior knowledge of vessel bends at and radial gradient of intensity that belong to rim area. Blood vessels are segmented using Multi-Scale vessel Enhancement Filtering then thinned segmented blood vessels into 1 pixel wide line in order to compute the intensity of each pixel in the curvature vessels, then performs the obtaining radial gradient process using dilated vessels mask to calculate only gradient of pixels which are not vessels. Reinforcing relevant vessel bends process is applied in order to determine the vessel bends that belong to OC only, finally the agents start constructing their solutions then apply smoothing the segmented cup. This method was tested on Rim-One dataset and the overlapping error is 24.3%.

3.4 Deep Convolutional Neural Network Based Approaches:

Table 5: *Glaucoma diagnosis approaches based on deep CNNs*

Authors	Deep Learning CCNs Method	Result	Dataset	Number Of Images
Diaz-Pinto et al (2019)	DCGAN model and semi-supervised	Specificity:0.7986 Sensitivity: 0.829	40 publicly datasets	Total: 86926

	learning method	AUC: 0.9017 F-score: 0.8429		images, 956 glaucomatous and 1401 normal
Li et al (2018)	Combining multiple deep features extracted using various 7 CNNs and SVM classifier	AUC: 0.8483	Origa	650
Li el al (2018)	Inception-v3 architecture deep learning features extraction	Accuracy of 92.9%, Sensitivity of 95.6% and Specificity of 92.9%.	online dataset LabelMe which contains 200000 images	**39745** Was selected & graded from random 70000 images
Claro et al (2019)	7 CNNs architectures & the descriptors (LBP, GLCM, HOG, Tamura, GLRIM and morphology) features space	Accuracy: 93.61%	5 datasets: (Dristh, Rim-one, Hrf, Jsiec, Acrima)	1675
Bajwa et al (2019)	Regions with Convolutional Neural Network (RCNN)	AUC: 0.874	7 public datasets include : Origa, HRF, Drive.	100.000 for testing localization and 505 images for glaucoma diagnosis

Diaz-Pinto et al (2019) proposed an approach based Deep Convolutional Generative Adversarial Networks (DCGAN) model which consist of two phases 1- semi-supervised learning phase, 2-generation phase using T-SNE plot features in order to synthesize retinal images as a two-player minimax game, the result of AUC is 0.9017. this approach uses 40 publicly datasets contain 86926 images (956 glaucomatous images and 1401 normal): which is very low number of glaucomatous retinal images.

Li et al (2018) proposed an approach based on deep learning visual features extraction using 7 CCNs: AlexNet, VGG-16, VGG-19, GooLeNet, ResNet-50, ResNet-101, ResNet-152, then to use SVM classifier. Annan to clarify that integrating of holistic features which get AUC: 0.8229 and local features which get AUC: 0.8359 obtain better result which holistic + local features get AUC: 0.8483, this approach used Origa dataset contains 650 fundus images divided into 90% train set and 10% test set.

Li et al (2018) proposed an approach based on Inception-v3 architecture deep learning features extraction, because of big differences between selected images Zhixi et al. perform preprocessing to normalize images, he scaled pixel values to range from 0-1and resized it into 299x299 matrix and subtracted Local space average color to avoid color constancy issue, in this method Inception-v3 architecture which has 22 layers and composed of 11 inception modules was adopted and used for learning phase. This approach used random 70000 fundus images from online dataset LabelMe then selected 48116 images which have visible OD for the labelling as

glaucomatous Optic Neuropathy then only 39745 was fully graded into Glaucomatous and non-glaucomatous image (divided into 8000 images as test set for validation and 31745 images as training set). Result was an Accuracy of 92.9%, Sensitivity of 95.6% and Specificity of 92.9%.

Claro et al (2019) proposed an approach based on texture and shape features extraction using Convolutional Neural Networks (CNNs) and from analysis, it is good result when creating classification model from total 30682 concatenated features as these descriptors LBP, GLCM, HOG, Tamura, GLRLM, Morphology and Transfer learning then apply the gain ratio algorithm to rank and select features then to apply RF classifier. This approach was applied on databases Drishti, Rim-one-1, Rim-one-2 and Rim-one-3, total 873 images, the result of accuracy is 93.35 % and AUC 0.98.

Bajwa et al (2019) proposed an approach based on Regions with Convolutional Neural Networks (CNNs), Bajwa tested his proposed approach on only 551 images (412 healthy images + 139 glaucomatous images) from 7 public datasets include the large publicly ORIGA, HRF, Drive datasets, this approach achieved Area Under the Receiver Operating Characteristic Curve (AUC): 0.874.

4. Discussion

Because we restricted our paper to only the most recent approaches in the last few years, so we categorized the glaucoma detection approaches that we collected into four main categories: image processing, machine learning, metaheuristic and deep learning depending on the used method in each paper. All these approaches perform the same steps which are firstly localizing the OD, secondly segmenting the OD and OC and finally calculating the CDR. The first step of localization and second step of segmentation face difficult challenges such as noise, non-uniform illumination, blood vessels distracters, big similarity in color and intensity between OC, OD and PPA regions, plus the big variations among retinal images in the same dataset as shown in figure 4.

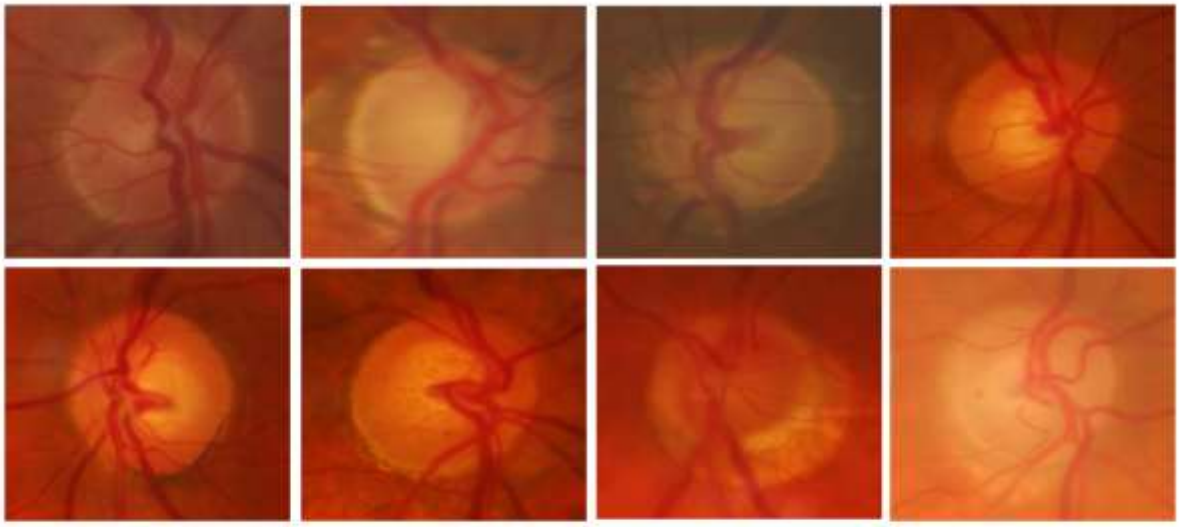


Figure 4: *some retinal images from Rim-one-1 dataset indicate some of segmentation challenges such as noise, non-uniform illumination, blood vessels distracters, big similarity in color and intensity between OC, OD and PPA regions.*

For the first category that includes approaches based on image processing methods, although it is the largest category of most recent approaches, but performances of half of them are not evaluated on benchmark datasets rather than the used local datasets which also have insufficient number of glaucomatous retinal images – one of them includes only six images – for good evaluation. Even the second half of these approaches was evaluated on too small benchmark datasets – HRF dataset has only 10 images – except Yin et al (2012), who used significant benchmark dataset of 650 images, so their paper is still the most cited in the literature.

For the second category that includes approaches based on machine learning methods, their performances were evaluated on respectful benchmark datasets, but since machine learning methods rely on experts annotations in the training images then each method will not has the same results and accuracy for different datasets, specially each dataset has a big variations in each image which is shortage of their ability to be used for glaucoma diagnosis in all datasets.

For the third category that includes approaches based on swarm intelligence methods, Arnay et al (2017) has challenged the various of noise and weak pallor in retinal images successfully but after evaluation the overlapping error was 24.3%, while in Bechar et al (2017), CDR error was 9.84 and in Cheng et al (2013) overlapping error was 26% which means that swarm intelligence methods need improvement for better performance than machine learning methods and to be evaluated on many datasets.

For the fourth category approaches based on deep CNNs, they must have used with very big datasets which most of its retinal images are not related to glaucoma diagnosis and only limited number of images are classified as glaucomatous or normal, like Diaz-Pinto et al (2019) method which tested on 86926 images while only 956 are glaucomatous images and Bajwa et al (2019)

method which used 100.000 images but only 505 images are classified as glaucomatous, and their results were not more accurate 93.61% and max AUC: 0.9017.

5. Conclusion

Glaucoma detection in retinal images is still currently and important point to many researchers. In this paper, we presented only some of the recent approaches in the last few years for automated glaucoma diagnosis in retinal images. These approaches were divided them into four categories that were based on image processing, machine learning, swarm intelligence and deep learning methods.

Many of introduced approaches in this paper should be evaluated on benchmark datasets especially the image processing based category or more benchmark datasets to assure their convergence in order to be run and used in clinic devices, since current datasets such as ORIGA, HRF and Rim-One do not have enough characteristics and many retinal images of them have low resolution.

We noticed good results are obtained when using combined many features with preprocessing and classification methods. So, in future works, we recommend that researchers utilize many features of retinal images and to normalize it for faster running, combine many preprocessing and classification methods and finally to use public datasets that we can comprise results with each other in order to enhance glaucoma detection performance in retinal images.

References:

Almazroa, A., Burman, R., Raahemifar, K., & Lakshminarayanan, V. (2015). Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. Journal of ophthalmology, 2015.

Arnay, R., Fumero, F., & Sigut, J. (2017). Ant colony optimization-based method for optic cup segmentation in retinal images. Applied Soft Computing, 52, 409-417.

Bajwa, M. N., Malik, M. I., Siddiqui, S. A., Dengel, A., Shafait, F., Neumeier, W., & Ahmed, S. (2019). Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. BMC medical informatics and decision making, 19(1), 136.

Bechar, M. E. A., Settouti, N., Barra, V., & Chikh, M. A. (2018). Semi-supervised superpixel classification for medical images segmentation: application to detection of glaucoma disease. Multidimensional Systems and Signal Processing, 29(3), 979-998.

Budai, A., Bock, R., Maier, A., Hornegger, J., & Michelson, G. (2013). Robust vessel segmentation in fundus images. International journal of biomedical imaging, 2013.

Cheng, J., Liu, J., Xu, Y., Yin, F., Wong, D. W. K., Tan, N. M., ... & Wong, T. Y. (2013). Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. IEEE transactions on medical imaging, 32(6), 1019-1032.

Chrástek, R., Wolf, M., Donath, K., Niemann, H., Paulus, D., Hothorn, T., ... & Michelson, G. (2005). Automated segmentation of the optic nerve head for diagnosis of glaucoma. Medical image analysis, 9(4), 297-314.

Claro, M., Veras, R., Santana, A., Araújo, F., Silva, R., Almeida, J., & Leite, D. (2019). An hybrid feature space from texture information and transfer learning for glaucoma classification. Journal of Visual Communication and Image Representation, 64, 102597.

Costagliola, C., Dell'Omo, R., Romano, M. R., Rinaldi, M., Zeppa, L., & Parmeggiani, F. (2009). Pharmacotherapy of intraocular pressure: part I. Parasympathomimetic, sympathomimetic and sympatholytics. Expert opinion on Pharmacotherapy, 10(16), 2663-2677.

Deepika, E., & Maheswari, S. (2018, January). Earlier glaucoma detection using blood vessel segmentation and classification. In 2018 2nd International Conference on Inventive Systems and Control (ICISC) (pp. 484-490). IEEE.

Diaz-Pinto, A., Colomer, A., Naranjo, V., Morales, S., Xu, Y., & Frangi, A. F. (2019). Retinal Image Synthesis and Semi-supervised Learning for Glaucoma Assessment. IEEE transactions on medical imaging.

Foong, A. W., Saw, S. M., Loo, J. L., Shen, S., Loon, S. C., Rosman, M., ... & Wong, T. Y. (2007). Rationale and methodology for a population-based study of eye diseases in Malay people: The Singapore Malay eye study (SiMES). Ophthalmic epidemiology, 14(1), 25-35.

Fumero, F., Sigut, J., Alayón, S., González-Hernández, M., & González de la Rosa, M. (2015). Interactive tool and database for optic disc and cup segmentation of stereo and monocular retinal fundus images.

Harizman, N., Oliveira, C., Chiang, A., Tello, C., Marmor, M., Ritch, R., & Liebmann, J. M. (2006). The ISNT rule and differentiation of normal from glaucomatous eyes. Archives of ophthalmology, 124(11), 1579-1583.

Hatanaka, Y., Nagahata, Y., Muramatsu, C., Okumura, S., Ogohara, K., Sawada, A., ... & Fujita, H. (2014, August). Improved automated optic cup segmentation based on detection of blood vessel bends in retinal fundus images. In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 126-129). IEEE.

Hatanaka, Y., Noudo, A., Muramatsu, C., Sawada, A., Hara, T., Yamamoto, T., & Fujita, H. (2011, August). Automatic measurement of cup to disc ratio based on line profile analysis in retinal images. In 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3387-3390). IEEE.

Hilal, S., Ikram, M. K., Saini, M., Tan, C. S., Catindig, J. A., Dong, Y. H., ... & Qiu, A. (2013). Prevalence of cognitive impairment in Chinese: epidemiology of dementia in Singapore study. J Neurol Neurosurg Psychiatry, 84(6), 686-692.

http://www5.cs.fau.de/research/data/fundus-images

Jain, S., & Salau, A. O. (2019). Detection of glaucoma using two dimensional tensor empirical wavelet transform. SN Applied Sciences, 1(11), 1417..

Li, A., Wang, Y., Cheng, J., & Liu, J. (2018, April). Combining multiple deep features for glaucoma classification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 985-989). IEEE.

Li, Z., He, Y., Keel, S., Meng, W., Chang, R. T., & He, M. (2018). Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology, 125(8), 1199-1206.

Nicolela, M. T., & Vianna, J. R. (2016). Optic nerve: clinical examination. In Pearls of Glaucoma Management (pp. 17-26). Springer, Berlin, Heidelberg.

Quigley, H. A., & Broman, A. T. (2006). The number of people with glaucoma worldwide in 2010 and 2020. British journal of ophthalmology, 90(3), 262-267.

Sedai, S., Roy, P. K., Mahapatra, D., & Garnavi, R. (2016, August). Segmentation of optic disc and optic cup in retinal fundus images using shape regression. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 3260-3264). IEEE.

Shyam, L., & Kumar, G. S. (2016, July). Blood vessel segmentation in fundus images and detection of Glaucoma. In 2016 International Conference on Communication Systems and Networks (ComNet) (pp. 34-38). IEEE.

Sivaswamy, J., Krishnadas, S. R., Joshi, G. D., Jain, M., & Tabish, A. U. S. (2014, April). Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In 2014 IEEE 11th international symposium on biomedical imaging (ISBI) (pp. 53-56). IEEE.

Vijapur, N. A., & Kunte, R. S. R. (2017). Sensitized glaucoma detection using a unique template based correlation filter and undecimated isotropic wavelet transform. Journal of Medical and Biological Engineering, 37(3), 365-373.

Xu, Y., Duan, L., Lin, S., Chen, X., Wong, D. W. K., Wong, T. Y., & Liu, J. (2014, September). Optic cup segmentation for glaucoma detection using low-rank superpixel representation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 788-795). Springer, Cham.

Xu, Y., Liu, J., Lin, S., Xu, D., Cheung, C. Y., Aung, T., & Wong, T. Y. (2012, October). Efficient optic cup detection from intra-image learning with retinal structure priors.

In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 58-65). Springer, Berlin, Heidelberg.

Yin, F., Liu, J., Wong, D. W. K., Tan, N. M., Cheung, C., Baskaran, M., ... & Wong, T. Y. (2012, June). Automated segmentation of optic disc and optic cup in fundus images for glaucoma diagnosis. In 2012 25th IEEE international symposium on computer-based medical systems (CBMS) (pp. 1-6). IEEE.

Zhang, Z., Yin, F. S., Liu, J., Wong, W. K., Tan, N. M., Lee, B. H., ... & Wong, T. Y. (2010, August). Origa-light: An online retinal fundus image database for glaucoma analysis and research. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (pp. 3065-3068). IEEE.

An Enhanced Approach for Privacy Preserving Data Mining

Hajar. H. Redha[1] Ahmed M. Gadallah[2] Hesham A. Hefny[3]

1. Master student at department of computer science, faculty of graduate studies for statistical research, Cairo University, Egypt.
2. Assistant professor at department of computer science, faculty of graduate studies for statistical research, Cairo University, Egypt.
3. Professor at department of computer science, faculty graduate studies for statistical research, Cairo University, Egypt.

ABSTRACT

Many organizations, public, private, or educational institutes collect and collate data to have the ability to provide a range of services. The broad range of services will need a vast amount of private and confidential information to be collected and analyzed for use; maintaining such data privacy using data mining techniques became more challenging. Various PPDM methods help in the process of preserving such data private; thus, they are advised for use. Yet, there is still a need for more flexible approaches to allow Privacy Preserving when applying Data Mining techniques. In this work, PPDM techniques and efficient solutions for everyday problems faced by traditional PPDM techniques are thoroughly outlined. The proposed includes normalization, categorization, and substitution of sensitive attributes. It makes sure that data is entirely private with decreasing the loss of available information. An illustrative case study shows the benefits of the proposed approach.

Key Words

Knowledge discovery (KD), Data Privacy, Data Mining, Privacy Preserving Data Mining (PPDM).

I. INTRODUCTION

Generally, data mining is an essential tool to extract valuable patterns or knowledge from huge amounts of data to help in the process of monitoring and decision making. Data Mining has been defined as the procedure of finding intriguing knowledge from a colossal volume of data that has been stored in databases or stored in large data warehouse or any data archives.

The main objective of data mining is to form descriptive or predictive models from data. Predictive models are used to predict unknown or future data, whereas Descriptive models attempt to turn patterns into human-readable descriptions of the data. Data Mining can also be referred to as KDD (Knowledge discovery from data). KDD typically refers to the process composed of the following sequence of steps **[**Han and Kamber**]:**

(a) **Data cleaning**: Removes noise and inconsistent data.
(b) **Data integration**: Multiple data sources may be combined.
(c) **Data selection**: Data relevant to the analysis task are retrieved from the database.
(d) **Data transformation or normalization**: Data is transformed into forms appropriate for mining.
(e) **Data mining**: Intelligent Methods are applied to extract data patterns.
(f) **Pattern evaluation**: To identify and evaluate patterns representing knowledge.
(g) **Knowledge presentation**: Knowledge representation techniques are used to present mined knowledge to the user.

Commonly, data mining has various algorithms and techniques like Classification, Association Rule Mining, Clustering, Regression, Decision Trees, Prediction, etc. The first three are the most common approaches in machine learning, where the first two are supervised learning techniques, and the Clustering is an unsupervised learning mechanism.

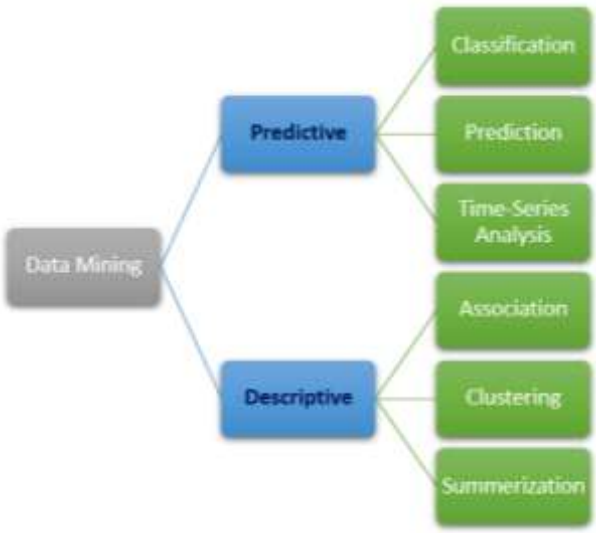


Figure 1. Taxonomy of data mining techniques

Privacy-preserving in data mining represents one of the fastest-growing research areas to improve the efficiency of existing techniques. Data mining techniques are applied to several application domains like healthcare and insurance, financial organizations, transportation, banking, education, and e-Commerce. Privacy preservation is a crucial concern in data mining as the secrecy of sensitive information must be maintained while sharing the data through different un-trusted parties (side). Privacy-preserving data mining PPDM protects sensitive data without losing information or reduce the benefit of mining process. The main idea of information privacy is to have control over the collection and handling of one's personal data. Some benefits of the information technologies are only possible through the collection and analysis of data. To protect from information leakage, privacy preservation methods have been developed to protect the owner's exposure by modifying the original data. However, transforming the data may also reduce its utility, resulting in inaccurate or even infeasible extraction of knowledge through data mining. This is the paradigm known as privacy-preserving data mining PPDM. Several different taxonomies for PPDM methods have been proposed. The three main approaches of Privacy preservation are Perturbation, Anonymization, and Cryptography. The types of **PPDM** techniques have been classified based on data lifecycle phase such as:

A. Data Collection

To ensure privacy at data collection time, the sensory device transforms the raw data by randomizing the captured values before sending it to the collector. The assumption is that the entity collecting the data is not to be trusted. Therefore, and to prevent privacy disclosure, the original values are never stored and used only in the transformation process. Consequently, randomization must be performed individually for each captured value.

B. Randomization

In this method and most common method, the original data values are modified by adding noise using known statistical distribution so that when data mining algorithms are used, the original data distribution may be reconstructed, but not the original (individual) values. Thus, the randomization process in data mining encompasses the following steps:

- Randomization at data collection
- Distribution reconstruction (subtracting the noise distribution from first step).
- Data mining on the reconstructed data.

This randomization method is known as randomization with additive noise.

Another way to add noise is by multiplying noise with a known statistical distribution known as randomization using multiplicative noise. Since original data values are modified into perturbed data. The randomization method requires specific data mining algorithms that can leverage knowledge discovery from the distribution of data but not from original values.

Data modification may be applied at other phases than at data collection, and other methods besides additive and multiplicative noise do exist. In fact, randomization is considered to be a subset of the perturbation operations. Though the collector is not trusted, so, the original data must not be stored nor buffered after the transformation.

C. Privacy Preserving Data Publishing

The main Consideration of PPDM is twofold, first sensitive raw data like identifiers, name, addresses, etc., should be hidden or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, the data which has to be mined from the database by using data mining algorithms should also be precluded because the main purpose of PPDM is to develop techniques for modifying the original data in some way so that. The private data and private knowledge remain private even after the mining process. It is necessary to know about the basic form of data in the data table, and its attributes different types. Table 1 gives four types of attributes that may exist in a data [Mailk., Gahzi and Ali][Raj and Kulkarni].

Table 1 Types of attributes in a data table

Attributes	Description
Explicit identifiers	These are attributes explicit identifiers an individual such as Name, Address, Social Security Number (SSN).
Quasi identifiers	These are attributes that do not explicitly identify an individual in isolation, but can nevertheless be used in combination by joining them with publicly available information such as Gender, Age, Zip code.
Sensitive identifiers	These are attributes that are considered private by most individuals such as Salary and disease.
Non-Sensitive identifiers	These are attributes that creates no problem if revealed even to untrustworthy parties.

Also, these techniques study different transformation methods associated with privacy, e.g., randomization, k-anonymity, 1-diversity, t-closeness, slicing, and also handles like how perturbed. Data can be used in conjunction with association rule mining.

D. Data Mining Output Privacy

The outputs of the data mining algorithms may be extremely revealing, even without explicit access to the original dataset. An adversary may query such applications and infer sensitive information about the underlying data. Below, a description of the most common techniques to preserve privacy to the output of the data mining is presented.

- **Association Rule Hiding**

In association rule data mining, some rules may explicitly disclose private information about an individual or a set of individuals. Association rule hiding is a privacy-preserving technique whose objective is to mine all non-sensitive rules, while no sensitive rule is discovered.

- **Downgrading Classifier Effectiveness**

Classifier applications may also leak information to adversary users. A good example is the membership inference attacks, in which an adversary determines if a record is in the training dataset (original data). Since some rule-based classifiers use association rule mining methods as subroutines, association rule hiding methods are also applied to downgrade the effectiveness of a classifier.

- **Query Auditing**

In Some cases, some entities may provide access to the original dataset, allowing exclusively statistical queries to the data. Users can only query the aggregate data from the dataset but not the individual or group records. Still some queries may reveal original or private data or private information.

Query auditing technique provides privacy by using two approaches: Query inference control and Query auditing. In Query inference control, either the original data or the output of the query is perturbed. In Query auditing, one or more queries are denied from a sequence of queries.

II. RELATED WORK

Amar Paul el. in [Singh and Parihar] presented and focused on effective methods that can be used for providing better data utility and handling high-dimensional data. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats we consider slicing where each attribute is in exactly one column. An extension is the notion of overlapping slicing, which duplicates an attribute in more than one column. Our experiments show that random grouping is not very effective. The Proposed grouping algorithm is optimized L-diversity slicing check algorithm obtains the more effective tuple grouping and Provides secure data. Another direction is to design data mining tasks using the anonymized data computed by various anonymization techniques. Another important advantage of slicing is that it can handle high-dimensional data. The data analyst has to make the uniform distribution assumption that every value in a generalized interval is equally possible this significantly is twofold: firstly, it reduces the data utility of the generalized data; and secondly, each attribute is generalized separately correlations between different attributes are lost. This is an inherent problem of generalization that prevents effective analysis of attribute correlations. Bucketization first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket

Commonly, generalization for k-anonymity loses considerable amount of information especially for high dimensional data [Brickell and Shmatikov]. Limitations of k-anonymity are:

- Not hide whether a given individual is in the database
- Reveals individuals' sensitive attributes.
- Not protect against attacks based on background knowledge
- Mere knowledge of the k-anonymization algorithm can violate privacy.
- Cannot be applied to high-dimensional data without complete loss of utility

On the other hand, the work presented in [Lakshmi and Siddamsetti] focuses mainly on using a method of combination of randomization and k-anonymity techniques to preserve privacy and reduce Information loss and increase private gain. It makes

the attacker challenging to find the background and homogeneity attack. And also, it provides less information loss and more privacy for sensitive data. The adopted strategy is separated into two algorithms.

Algorithm 1 is utilized to perform randomization on dataset utilizing property transitional likelihood matrix, and calculation two is used to perform k-anonymity on a randomized yield dataset. In other words, Algorithm 2 employments the result of calculation one as input. In Algorithm 1, the quasi identifier, sensitive attribute, and key attribute are selected from the original dataset, then suppressed key attribute, After removing the key attributes, we get table contain Age, Gender and Pincode attributes are considered as quasi-identifiers, and disease attribute is considered as a sensitive attribute. Then given an input to algorithm 2 categorize the sensitive values into two class high (H) and low (L) then classify and move each tuple belong to categorize class (H) or (L) and don't anonymize the sensitive attribute Disease contains four values specifically cancer, HIV+, Bronchitis, and Gastritis. Among which HIV+ and Cancer are categorized into High Sensitive class (H) Bronchitis and Gastritis are classified into low sensitive class (L).

In contrary, K. Ilavarasi et, in [Ilavarasi, Sathiyabhama and Poorani], introduced different techniques that describe the taxonomy of Privacy preservation in data mining. The three main approaches of privacy preservation are Perturbation, Anonymization, and cryptography

The perturbation method for categorical data can be used by organizations to prevent or limit disclosure of confidential data for identifiable records when the data are provided to analysts for classification. There are many kinds of data perturbation methods that include additive, multiplicative, matrix multiplicative, categorical data, data swapping, resampling, micro aggregation, data shuffling. Anonymization reduces the risk of identity disclosure, whereas the data remains still realistic. Two main Privacy-Preserving approaches are k-anonymity and 1-diversity. Generalization is one of the conventional anonymization techniques. Bucketization Anatomy is one of the new techniques for publishing sensitive data. Protects privacy by releasing all the Quasi-Identifiers and sensitive values in two separate tables. This technique provides an effective data analysis than generalization. And it also achieves privacy-preserving publication by capturing the exact QI-distribution. Bucketization has better utility than generalization, but bucketization does not prevent membership disclosure and does not have any clear separation between QIs and SAs. Anonymization through permutation is carried

out because of the following reasons. The individual's identity can be recovered in three ways.

- link between the identifier and quasi-identifiers in the public database **P**

- link between the QIs in **P** and those in the deidentified micro data **D**
- link between QIs and the sensitive value **D**

The associations of the above links will ensure privacy. Domain generalization weakens only the second and third links. Instead of using domain generalization, we can permute the association between the quasi-identifiers and the sensitive attributes. Even if an attacker can link an individual's identifier with tuple's QI, he will not be able to know with certainty the exact value of the individual's sensitive attribute. Is one of the novel techniques which better preserves the data utility than generalization. Slicing also protects membership disclosure than bucketization. High dimensional data can be handled better by slicing based anonymization.

III. THE PROPOSED APPROACH

The proposed approach presented in this work for privacy preserving data mining incorporates the following steps:

Step 1: Data Collection

The data collection stage is the basic stage for gathering and defining the data set.

Step 2: Generalization

Data generalization summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute age) by categorizing data ranges into small identifiable sub-groups. Given the large amount of data stored in databases, it is useful to be able to describe concepts in concise and succinct terms at generalized (rather than low) levels of abstraction

Step 3: Segmentation

Divide the values and convert them from Binary value or Bit Datatype to (Yes, No)

Step 4: Normalization

Normalization is necessary when dealing with attributes with a different distribution of values. Alternatively, it could lead to a less efficient method of picking an equally important value while dismissing another. In contrast, poor models would be generated when multiple values with different range are present.

For this reason, normalizing values to being them to the same level would assist in the performance of data mining operations dramatically.

Step 5: Substation

Replace sensitive features with other equivalent features to preserve their characteristics

Step 6: Calculate association rules

Finding association rules using the Apriori algorithm represents a classical approach in data mining. It is used mainly to find the frequent itemset and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions.

Algorithm1: the proposed data mining algorithm

Input: Patients data set and other related data.

Output: A set of association rules indicating the effects of correlated attributes of diabetes degrees with privacy preserving.

Begin:

1. Extract Dataset from Original Tables.
2. Create new table to import data after substitution.
3. Define Identity Column in new table auto generated when insert row.
4. Select the quasi identifier, key attributes and sensitive attribute from original dataset.
5. Remove/suppress the key attributes.
6. Substitution of each sensitive column (**address)** by other correlated attributes **(Avg_Temp_Desc).**
7. Normalization of each attribute values like the attributes (**eAG)** to make its collected values from different patients' sites have the same measurement unit.

8. Generalization the Quasi identifier value (**Age , Weight , AvgTemp**) at multiple levels.
9. Merge all collected datasets into one data set (new table).
10. Apply the association rules mining technique.

End Algorithm

IV. An Illustrative Case Study

This case study is concerned with mining the association rules for the diabetes patients' correlated factors or variables with the diabetes degrees.

1. Data Collection

The data collection stage is the basic stage for gathering and defining the data set. In this work, data is collected from two sources:

(a) Data from a descriptive correlational design was used with a convenience sample of Lebanese adults with type 2 diabetes recruited from a major hospital in Beirut, Lebanon. [Sukkarieh-Haraty, Egede, Kharma and Bassil]

(b) A questionnaire was created using the Google Model, which targets diabetic patients in Kuwait and Egypt, taking into account the existence of the same basic fields found in diabetes data in Lebanon.

The following set of data tables (from Table 2 to Table 7) gives the important characteristics or attributes of patients under study.

Table 2. Occupation Type

ID	Description
1	Employed
2	UnEmployed
3	Unable to work due to health Condition

Table 3. Education Level

ID	Description
1	<=High School
2	Technical
3	>=Undergraduate degree

Table 4. Country Income Level

Table 5. Social Status

ID	Name	Income Description
61	Egypt	Lower middle-income
98	Jordan	Upper middle-income
104	Kuwait	High income
108	Lebanon	Upper middle-income
167	S. Arabia	High income

ID	Status
1	Single
2	Married
3	Divorced
4	Widowed

Table 6. Country Climate Data

ID	Avg Temp Highest	Avg Temp Lowest	Avg Temp	Avg Temp Desc
61	32.59	10.95	22	Moderate
98	31.79	4.99	19	Moderate
104	38.84	8.07	24	High
108	29.44	4.55	17	Low
167	36.50	12.76	25	High

Table 7. Person Income Level

ID	Description
1	high
2	Middle
3	Low

2. Generalization:

Data generalization summarizes data by replacing relatively low-level values (e.g., numeric values for an attribute age) by categorizing data ranges into small identifiable sub-groups. Given the large amount of data stored in databases, it is useful to be able to describe concepts in concise and succinct terms at generalized (rather than low) levels of abstraction for Example:

- For the Age attribute:
  Its values can be categorized or generalized as follows:
  - (Age: between 12 and 35 ⟶ Age_Group: Young)
  - (Age: between 36 and 55 ⟶ Age_Group: Middle)
  - (Age: > 55 ⟶ Age_Group: Older)
- Weight
  - (Weight: <=55 ⟶ Weight_Group: Low)
  - (Weight: between 56 and 79 ⟶ Weight_Group: Moderate)

- (Weight: between 80 and 109 → Weight_Group: High)
- (Weight: between >=110 → Weight_Group: Very High)

- AvgTemp
  - (AvgTemp <= 18 → Avg_Temp_Deac: Low)
  - (AvgTemp BETWEEN 19 AND 23→Avg_Temp_Deac: Moderate)
  - (AvgTemp > 23 → Avg_Temp_Deac: High)

- HbA1C (Diabetes) [Hanas and John]
  - (HbA1C between 5.7 and 6.3 → HbA1C_Group: Low)
  - (HbA1C between 6.4 and 7.0 → HbA1C_Group: Medium)
  - (HbA1C > 7.0 → HbA1C_Group: High)

3. Segmentation

Divide the values and convert them from Binary value or Bit Datatype to (Yes, No)

– Binary: Smoking (Yes/No)

4. Normalization

All average glucose, or eAG readings were standardized Health care providers can now report A1C results to patients using the same units (mg/dl or mmol/l) that patients see routinely in blood glucose measurements the relationship between A1C and eAG is described by the formula:[Osborn, Bains and Egede][Soros, Chalew, McCarter, Shepard And Hempe]

- eAG = 28.7 X A1C(HbA1c) – 46.7
- Avg.Plasma Blood Glucose (mg/dl) = (HbA1c * 35.6) - 77.3
- Avg.Plasma Blood Glucose (mmol/L) = (HbA1c * 1.98) - 4.29

Table 8. Extract Dataset from Original Tables in Database

ID	Name	Age	Gender	Weight	Country	Education Level	Social Status	Occupation	Smoking	Income	eAG
1	Jassim	43	M	90	Kuwait	Undergraduate degree	Married	Employed	Yes	Middle	9.00
2	Amani	38	F	60	Kuwait	Technical	Married	Employed	No	Middle	22.00
3	Lina	39	F	73	Lebanon	Technical	Married	UnEmployed	Yes	Middle	9.40
4	Tony	49	M	70	Lebanon	High School	Divorced	Employed	Yes	Middle	8.40
5	Fathy	69	M	85	Egypt	Undergraduate degree	Married	UnEmployed	No	Middle	190.0
6	Alaa	59	M	60	Egypt	<11 Years	Married	Employed	No	high	9.00

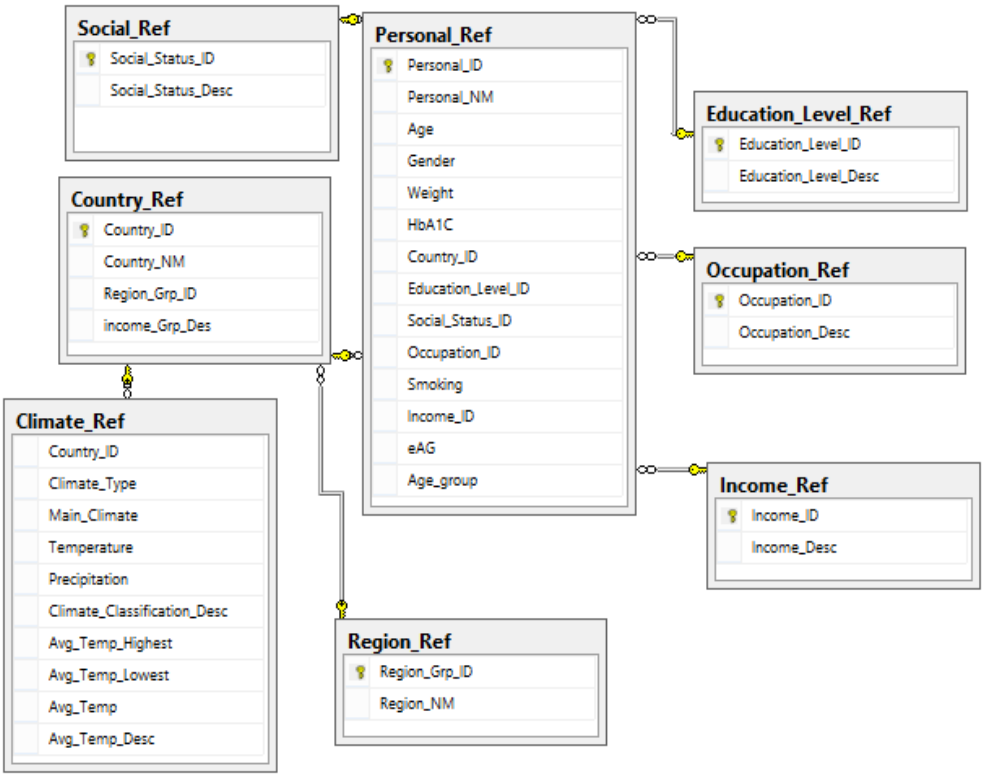


Figure 2 Original data set and related data tables

Table 9. Select the quasi identifier, key attributes and sensitive attribute

Table 10. After removing key attribute

Key Attribute		Sensitive attribute		Quasi identifier		
ID	Name	Country Name	eAG	Age	Gender	Weight
26	Jassim	Kuwait	9.00	43	M	90
446	Amani	Kuwait	22.00	38	F	60
180	Lina	Lebanon	9.40	39	F	73
245	Tony	Lebanon	8.40	49	M	70
748	Fathy	Egypt	190.0	69	M	85
869	Alaa	Egypt	9.00	59	M	60

Sensitive attribute	Quasi identifier			
Country Name	eAG	Age	Gender	Weight
Kuwait	9.00	43	M	90
Kuwait	22.00	38	F	60
Lebanon	9.40	39	F	73
Lebanon	8.40	49	M	70
Egypt	190.0	69	M	85
Egypt	9.00	59	M	60

Table 11. Substitution sensitive attribute with equivalent attributes

Sensitive attribute	Quasi identifier			
Avg Temp	eAG	Age	Gender	Weight
High	9.00	43	M	90
High	22.00	38	F	60
Low	9.40	39	F	73
Low	8.40	49	M	70
Moderate	190.0	69	M	85
Moderate	9.00	59	M	60

Table 12. Categorize Quasi identifier attributes with equivalent class (e.g., Weight)

Sensitive attribute	Quasi identifier				
Avg Temp Desc	HbA1C	HbA1C Group	Age	Gender	Weight
High	7.30	High	Middle	M	High
High	15.40	High	Middle	F	Moderate
Low	7.50	High	Middle	F	Moderate
Low	6.90	Medium	Middle	M	Moderate
Moderate	8.20	High	Older	M	High
Moderate	7.30	High	Older	M	Moderate

Table 13. Insert All Data to New Created Table in SQL

Age Grp	Gender	HbA1C Group	Weight Group	Education Level	Social Status	Occupation	Smoking	Income	Avg Temp
Older	F	Low	Moderate	<=High School	Divorced	UnEmployed	False	Middle	High
Middle	F	High	High	>=Undergraduate degree	Married	Employed	False	high	High
Older	M	High	High	Technical	Married	UnEmployed	False	Middle	Moderate
Older	F	Medium	High	<=High School	Widowed	UnEmployed	False	Middle	High
Older	M	Medium	Low	Technical	Married	UnEmployed	False	Middle	Low
Older	F	High	Moderate	<=High School	Married	UnEmployed	True	Low	Low
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Step 5: Association Rules Mining

In this step, Microsoft Azure Machine Learning Studio is used to discover the association rules between the diabetes degrees and the other affecting factors. Table 14 shows the final ready data set for mining. On the other hand, Fig.3 and Table 15 show a sample of discovered association rules.

Table 14. Retrieve dataset from Database

HbA1C Group	Age Group	Education Level	Income	Weight Group	Average Temp	Smoker	Occupation

High	Middle	<=High School	high	Moderate	Low	false	UnEmployed
High	Middle	<=High School	high	High	Low	false	UnEmployed
Medium	Middle	>=Undergraduate degree	high	Moderate	Low	true	Ineligible for work
High	Middle	>=Undergraduate degree	high	High	High	true	Employed
Low	Middle	>=Undergraduate degree	Middle	High	High	false	Employed
Low	Middle	>=Undergraduate degree	Middle	Very High	Moderate	false	Employed

Figure 3. Discover Association Rules using ML Azure

Table 15. Discover Association Rules using ML Azure -convert to CSV

id	Lhs	rhs	support	confidence	lift
1	{Education_Level_Desc=Technical,Income_Desc=Middle Income,Weight_Group=Very High Weight}	{Diabetes=High}	0.012927054	1	2.039548023
2	{Income_Desc=Low Income,Weight_Group=High Weight,Avg_Temp_Desc=High Temperature}	{Diabetes=High}	0.010156971	1	2.039548023
3	{Education_Level_Desc=Technical,Income_Desc=Middle Income,Weight_Group=Very High Weight,Avg_Temp_Desc=High Temperature}	{Diabetes=High}	0.011080332	1	2.039548023

4	{Education_Level_Desc=<=High School,Income_Desc=Middle Income,Weight_Group=Moderate Weight,Avg_Temp_Desc=Moderate Temperature}	{Diabetes=Medium}	0.020313943	1	3.88172043
5	{Education_Level_Desc=Technical,Income_Desc=Low Income,Weight_Group=High Weight}	{Diabetes=High}	0.010156971	0.916666667	1.869585687
6	{Income_Desc=Middle Income,Weight_Group=Moderate Weight,Avg_Temp_Desc=Moderate Temperature}	{Diabetes=Medium}	0.020313943	0.785714286	3.049923195
7	{Education_Level_Desc=>=Undergraduate degree,Income_Desc=Middle Income,Weight_Group=Moderate Weight,Avg_Temp_Desc=Low Temperature}	{Diabetes=Low}	0.013850416	0.681818182	2.829153605
8	{Education_Level_Desc=<=High School,Income_Desc=high Income,Weight_Group=High Weight,Avg_Temp_Desc=Low Temperature}	{Diabetes=Low}	0.018467221	0.606060606	2.514803204
9	{Education_Level_Desc=<=High School,Income_Desc=high Income,Weight_Group=High Weight}	{Diabetes=Low}	0.019390582	0.6	2.489655172
10	{Education_Level_Desc=>=Undergraduate degree,Income_Desc=high Income,Weight_Group=High Weight,Avg_Temp_Desc=High Temperature}	{Diabetes=Medium}	0.010156971	0.578947368	2.247311828
11	{Income_Desc=high Income,Avg_Temp_Desc=Moderate Temperature}	{Diabetes=High}	0.022160665	0.857142857	1.748184019

III. CONCLUSION

Commonly, privacy preserving data mining represents a crucial problem that needs flexible solutions without loss of information. The proposed approach in this paper outlines an efficient, convenient, and robust method that give various opportunities for privacy-preserving data mining. It enables an organization to share their information with guarantee for preserving the data confidentiality and with loss of information. This mission is achieved by applying normalization and categorization of Quasi identifier, and substitution processes for data's sensitive attributes. The illustrative case study shows the added value of the proposed approach against other previous ones. This approach will help preserve sensitive data and prevent unauthorized access to such information without losing its implicit values.

REFRENCES

Han. Jiawei, and Kamber. Micheline, "Data Mining: Concepts and Techniques", pp. 5-7

Mailk. Majid Bahir, Gahzi. M. Asger, Ali. Rashid, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", IEEE Third International Conference on Computer and Communication Technology, 2012.

Raj. Ronica and Kulkarni. Veena: A Study on Privacy Preserving Data Mining:Techniques, Challenges and Future Prospects

Singh. Amar Paul, Ms. Parihar. Dhanshri "A Review of Privacy Preserving Data Publishing Technique", in International Journal of Emerging Research in Management & Technology **ISSN**:2278-9359 Volume-2, Issue-6, JUNE 2013.

Brickell. J and Shmatikov. V, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008

Lakshmi. CH Venkata, Siddamsetti. Swapna "An Enhanced Approach for Privacy Preserving Data Mining (PPDM)", **IJRTER** | International Journal of Recent Trends in Engineering & Research **ISSN**(ONLINE): 2455-1457, IMPACT FACTOR:4.101

Ilavarasi. K, Sathiyabhama. B, Poorani. S "A Survey on Privacy Preserving Data Mining Techniques", **IJCSBI.ORG**|International Journal of Computer Science and Business Informatics, **ISSN**: 1694-2108 | Vol. 7, No. 1. NOVEMBER 2013

Sukkarieh-Haraty. Ola, Egede. Leonard E, Kharma. Joelle Abi, Bassil. Maya: Psychometric properties of the Arabic version of the 12-item diabetes fatalism scale January 11, 2018.

Hanas R, John G, Behalf of the International HbA1c Consensus Committee 2010 consensus statement on the worldwide standardization of the hemoglobin A1c measurement. Clin Chem. 2010;56:1362–4. [PubMed] [Google Scholar]

Consensus Committee Consensus statement on the worldwide standardization of the hemoglobin A1C measurement: the American Diabetes Association, European Association for the Study of Diabetes, International Federation of Clinical Chemistry and Laboratory Medicine, and the International Diabetes Federation. Diabetes Care. 2007;30:2399–400. [PubMed] [Google Scholar]

Osborn CY, Bains SS, Egede LE. Health literacy, diabetes self-care, and glycemic control in adults with type 2 diabetes. Diabetes Technol The. 2010; 12: 913–919..

Soros AA, Chalew SA, McCarter RJ, Shepard R, Hempe JM : Hemoglobin glycation index: a robust measure of hemoglobin A1c bias in pediatric type 1 diabetes patients. Pediatr Diabetes. 18 January 2010 [Epub ahead of print][Google Scholar]

A Survey on Vehicle Detection

Eslam H.Fouda[1], Ahmed H. Asad[2], Hesham A. Hefny[3]

[1]M.Sc. Student
[2,3]Computer Science Department, Faculty of Graduate Studies for Statistical Research, Cairo University

ABSTRACT

Vehicle detection is the process of localizing vehicles of different categories on roads. It is used in some different applications in context of intelligence transportation systems like calculating vehicle speed, density, volume, traffic flow rate, travelling time or congestion level. Also it can be applied for vehicle tracking, parking area monitoring, road traffic monitoring and management etc. Different methods handle this problem over many years. The aim of this survey is to assist the new researchers and provide guidance to them with the cutting-edge efforts and state-of-art methods in this very fast developing research field.

1. Introduction

Object detection is one the most common problems in the field of computer vision. It is the first step of image or video processing. Applications that uses object detection method are image retrieval, security, video surveillance, automatic vehicle parking systems, medical diagnosis and treatment, remote sensing, underwater sensing and civil infrastructure. Vehicle detection is an application of object detection on road scenes that contain other different categories like navigation paths, pedestrians, stop signs, traffic lights, motorcycles, trees and other objects. Vehicle detection has many applications such as automatic parking system to decide which vehicle type if allowed, automatic counting and information gathering about the vehicles coming in or going out of the parking lots, aerial surveillance of the vehicular objects in the urban areas, vehicle tracking on the roads across the countries. This paper focuses on methods process videos captured from road scenes for detecting vehicles on these roads

This paper is organized as follows: section 2 classifies vehicle detection methods into three classes and discusses each class of methods deeply. Section 3 gives comparison among the three classes in terms of pros and cons of each class, and evaluation of their results. Finally, the conclusion section displays some challenges in vehicle detection that still remain.

2. Vehicle Detection Challenges

The most common problems in the task of vehicle detection are the problems as shown in figures (1),(2), of different type of vehicles having different size and shape and different situations like occlusion, lightening and different weather conditions which is making harder to classify and detect moving vehicles in the dynamic scenes.

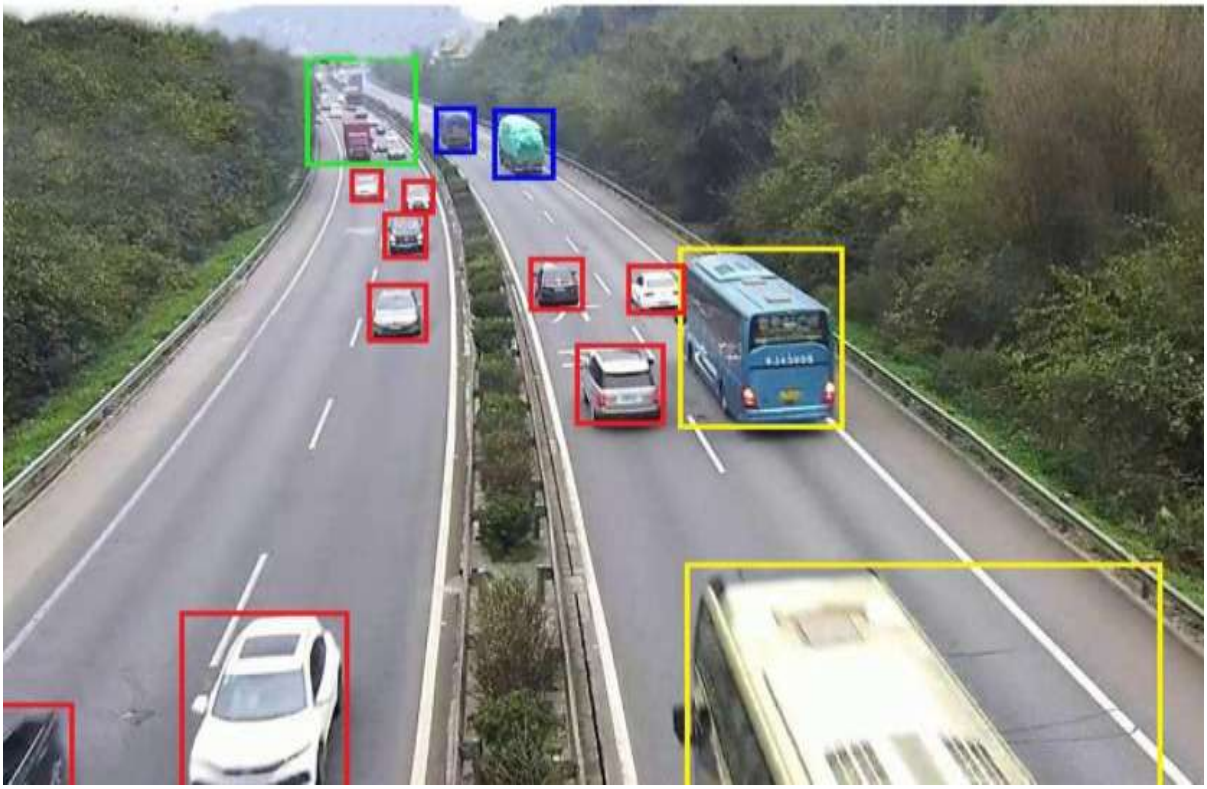


Figure (1): Traffic images and videos contain vehicles with a large variance of scales.

Figure (2): Vehicles detections in variant lighting and weather conditions.

3. Benchmark Datasets

It is difficult to compare the results of different real-time vehicle detection methods because their performances were measured on different datasets so it is very important to take a quick look on some common automated driving datasets as shown in table 1.

Table 1: Common automated driving datasets.

Dataset	Brief Description	Task	Format
KITTI Creator: The Karlsruhe Institute of Technology	open sourced six hours of data captured while driving in Karlsruhe	vision benchmarks	video
Oxford Robot Car Creator: Oxford University	1000 km recorded driving in central Oxford over the period of 1.5 years and features almost 20 million images	non-commercial academic use	video and images
Apollo Creator: Baidu company	annotated traffic sign in 74 thousand video frames, vehicle log data from demonstrations, training data for multi-sensor localization and scenarios for their simulation environment	semantic segmentation and classification	video
Udacity Creator: Udacity company	data recordings from more than ten hours of driving and annotated driving datasets where objects in video have been marked with surrounding boxes	open source tools, Udacity publishes programming challenges to further the development	video

4. Performance Measures

While recent research has been primarily focused on improving accuracy, for actual deployment in an autonomous vehicle, there are other issues of image object detection that are equally critical:

a) Accuracy. More specifically, the detector ideally should achieve 100% recall with high precision on objects of interest.

Precision: of examples that the classifier recognizes as vehicles what percentage actually are vehicles?

Precision = True Positive/ (True Positive + False Negative) or

Precision = True Positive/ Total Positive results

mAP: mean average precision averages the precision over easy, moderate, and hard scenes.

Recall: of all examples that really are vehicles what percentage were correctly recognized by the classifier?

Recall = True Positive/ (True Positive + False Negative) or

Recall = True Positive/ Total vehicle cases

F1 score: A standard way to combine precision and recall ”average of precision and recall”

F1 score = 2/ ((1/precision) + (1/recall))
F-measure = 2 ×((Recall × Precision) / (Recall + Precision))

Intersection over Union (IoU): A function used to evaluate the vehicle detection algorithm and improve it. This is one way to map localization to accuracy. The higher IoU, the more accurate bounding box.

IoU = (Ground truth bounding box size ∩ Algorithm bounding box size) /
(Ground truth bounding box size U Algorithm bounding box size)

Some other used measures:
Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)

Compactness = $Perimeter^2$ / Area
Aspect Ratio = Height / Width
Area Ratio = Area / Region of Interest

b) Speed. The detector should have real-time or faster inference speed to reduce the latency of the vehicle control loop. It is measured by frames per second (fps) or time per image.

c) Small model size. Smaller model size brings benefits of more efficient distributed training, less communication overhead to export new models to clients through wireless update, less energy consumption and more feasible embedded system deployment.

d) Energy efficiency. Desktop and rack systems may have the luxury of burning 250W of power for neural network computation, but embedded processors targeting mobile applications have an even smaller power budget and must fit in the 3W–10W range.

5. Vehicle Detection Methods

Detection of a moving vehicle from a captured video is a challenging task in the field of computer vision. Many methods were proposed for moving vehicle detection. In this section, various approaches proposed by various researchers have been discussed.

One of the earliest works in object detection to achieve real-time detection with high accuracy was the cascaded detector. Cascaded detector is an algorithm that has been widely used to implement sliding window detectors. Ohn-Bar and Trivedi (2015) used cascaded detector architecture to implement sliding window for handling vehicle detection of orientation and occlusion by learning an ensemble of detection models from visual and geometrical clusters of object instances. An AdaBoost detection scheme was employed as in ElKerdawy et al. (2014) with pixel lookup features for fast detection. Learning models at multiple resolution was shown to significantly improve detection/orientation estimation performance. This method achieved 66.42% mAP of car detection of KITTI dataset. The main current limitation of detector cascades is the difficulty of implementing multiclass detectors under this architecture.

There are two main methods that detect different classes from captured video: Motion-Based Features and Appearance Based Features.

5.1 Motion based Methods

Under this class of methods there exists two main techniques background subtraction and optical flow. At the end of motion based methods discussion a comparison of these methods is presented in table (2) :

5.1.1 Background Subtraction

There is a set of ideas called background subtraction. These methods subtract image frame containing vehicles from the background or previous frame and then processed to obtain vehicle regions. In these methods connected region of the vehicle that is "blob "associated with each vehicle is identified. The process of extracting moving foreground objects input image from stored background image static image called non-adaptive methods or model the background frame generated form image series captured by a static camera "video" which called adaptive methods like background subtraction, frame differencing, single Gaussian model and mixture Gaussian models.

Cho et al. (2019) proposed a novel algorithm, which can operate in moving camera environments. In addition, an area-efficient hardware design for the proposed algorithm was presented to test the algorithm in real-time processing. In this algorithm the background is modeled using the GMM algorithm. Then, the object detection process is done by a background subtraction operation first followed by Brox optical flow method that applied to two consecutive frames to extract the flow vectors. The algorithm used 27.2 fps and 480*704 image size. They achieved 98% precision and 95 % recall.

5.1.2 Optical Flow

Optical flow is the instantaneous speed of pixels on the image surface, which corresponds to moving objects in 3-D space using static or moving cameras. The main idea of optical flow is to match pixels between image frames using temporal and gradient information.

Agarwal et al. (2016) applied optical flow to an image and get flow vectors of the points corresponding to the moving objects in the sequence of video frame taken from the static camera. The algorithm averaged flow vectors are and generated the optical flow vectors over frames. For the better accuracy of the detection morphological erosion and dilation is performed. Kanade-Lucas-Tomasi (KLT) optical flow approach has been chosen for the estimation of optical flow because of its high accuracy and its basic principle that uses the change of intensity between two consecutive video frames for motion detection. Now the filtering is done to smooth out the boundaries of the moving object using median filters. The algorithm was tested on datasets available online, real time videos and also on videos recorded manually. The algorithm achieved 89.15% precision for car detection.

Table 2: A comparison of motion based methods

Method	Year	Dataset	Precision (mAP)
GMM+Optical Flow	2019	Platform, hardware system design	98%
Optical Flow	2016	datasets available online	89.15%

5.2 Appearance Based methods

In appearance based method, moving vehicle is detected by extracting the features from the vehicle regions such as rear lamps, license plate, rear view mirrors, edges and corners of vehicles etc. These features are processed to track the vehicles correctly. It may be extracted

using Statistical Learning Methods or Convolutional Neural Networks (CNN). The first step in appearance based method is vehicle segmentation. In this step, background mask is established and foreground objects are separated from the background. In the next step, features of vehicles such as, color, shape, and texture can be extracted as shown in figure (3) and employed for modeling. Further vehicles are classified by extracting the features like area ratio, compactness etc.

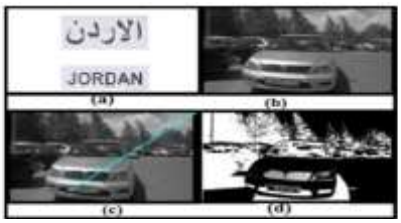
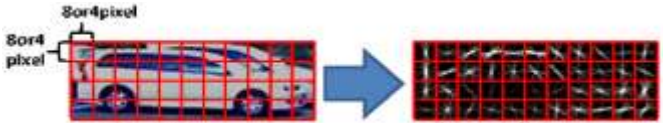


Figure (3): Template matching between the template image given in (a) and the input image given in (b). The matching results are shown in (c), where the blue lines are the successfully matched features. (d) the binarized image.

5.2.1 Statistical Learning Methods Based on Handcrafted Features

In Statistical Learning Methods like Histogram Oriented Gradients (HOG) as in Manzoor et al. (2017), Scale Invariant Feature Transformation (SIFT) as in Manzoor et al. (2017), and Haar-like as in ElKerdawy et al (2014) are commonly used for vehicle detection followed by classifiers like Support Vector Machine (SVM) as in SIFT+HOG by Manzoor et al. (2017) and AOG+SVM by Li et al. (2014) and Adaboost as in CascadedDetector+Adaboost by Ohn-Bar and Trivedi (2015) and Haar-features+Adaboost by ElKerdawy et al (2014). More advanced algorithms like Deformable Parts Model (DPM) as in Yebes et al. (2014) and And-Or Graph as in Li et al. (2014) that explore the underlying structures of vehicles and use hand-crafted features to describe each part of vehicles. At the end of statistical learning based methods discussion a comparison of these methods is presented in table (3):

Manzoor et al. (2017) presented Vehicle Make and Model Recognition System (VMMR) using Bag of SIFT Features. The method extracted image features from vehicle images and create feature vectors to represent the dataset. The system used HOG and GIST to represent the images and SVM and Random Forest (RF) to classify the vehicles. The proposed method was tested against publicly available National Taiwan Ocean University-Make and Model Recognition (NTOU-MMR) dataset. RandomForest -VMMR (HOG) achieved mAP of 94.53% and using 35.7 fps. SVM-VMMR (HOG) achieved mAP of 97.89% and using 13.9 fps. But the dataset contains frontal vehicle images. The HOG Descriptor methods figure (4) collects information from sequences of frames of scatter background but not efficient to detect the objects with the similar color to the background.



(a) Image (b) HOG feature map

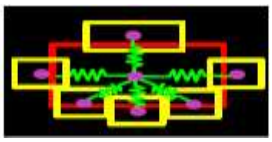
Figure (4): HOG feature map for vehicle detection

Li et al. (2014) proposed a method of learning reconfigurable hierarchical And-Or Graph (AOG) to integrate context and occlusion for car detection. The And-Or model represents the regularities of car-to-car context and occlusion patterns at three levels: a) N-car layouts, b) single car and c) car parts. The learning process consists of two stages. (1) learn the structure

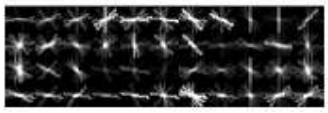
of the And-Or model with its three components. (2) train the model parameters (for appearance, deformation and bias) using Weak-Label Structural SVM. The method achieved 67.63% mAP for car detection in the KITTI benchmark.

Yebes et al. (2014) used a Discriminative Part-based Model (DPM) from images as part of the object detection and orientation estimation detector as shown in figure (5). The evaluation algorithm and metrics, the selection of a clean but representative subset of training samples and the DPM tuning are key factors to learn an object detector in a supervised fashion. They provided evidence of subtle differences in performance depending on these aspects. The method achieves 72.05% mAP for car detection in the KITTI benchmark.

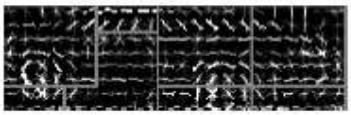
(a) Vehicle side model root and six parts



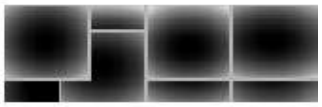
(b) Parts deformation



(c) Root filter



(d) Part filters



(e) Deformation costs

Figure (5): Structure of deformable object model

ElKerdawy et al (2014) presented a real-time vision framework that used Haar-like features and Adaboost cascade classifier generating a general rule that detected and tracked vehicles from stationary camera. The framework consists of three main stages. Vehicles were first detected using Haar-like features. In the second phase, an adaptive appearance-based model was built to distinguish between objects and back-ground and to dynamically keep track of the detected vehicles. This model was also used in the third phase of data association to fuse the detection and tracking results. The advantage of this framework that it was sensitive to vertical, horizontal and symmetric structure. The disadvantage was that it has a high computational efficiency. The paper used self-collected dataset to test the performance. The framework achieves 93% accuracy and overall average running time of 0.2017 seconds.

Table 3: A comparison of statistical learning based methods

Method	Year	Dataset	Precision (mAP)
SIFT+HOG	2017	NTOU-MMR	97.89%
AOG+SVM	2014	KITTI	67.63%
DPM	2014	KITTI	72.05%
Haar-features+Adaboost	2014	self-collected	93% accuracy

5.2.2 Convolutional Neural Network-Based Features

CNNs-based object detection algorithms fall into two categories. The first category is built upon a two-stage pipeline as in Adapted Faster R-CNN by Gao et al. (2017); ARVDS+DCNN by Azevedo et al. (2019); MS-CNN by Cai et al. (2016) of the squares or rectangles. However, this grid-based paradigm cannot obtain comparable accuracy to two-stage detection pipeline, as the grids have too strong spatial constraints to predict small objects appeared as groups. In this regard, most of the existing methods employ the two-stage detection pipeline. At the end of convolutional neural network-based methods discussion a comparison of these methods is presented in table (4):

Most existing solutions are designed based on two types of pyramid representations. Some works apply the concept of image pyramid like Gao et al. (2017) and Azevedo et al. (2019), which exploit the multi-scale input images to make the network fit all the scales. The main disadvantage of this representation is its large computational cost prohibiting its application to real-time detection tasks.
Gao et al. (2017) adapted the Faster R-CNN framework in a vehicle detection task to work well on a vehicle detection task for datasets presenting natural scale distribution and unbiased real-world object frequency. They had proposed several modifications on the network architecture's convolutional layers and region proposal selections. By selecting better scales in the region proposal input and by combining feature maps through careful design of the convolutional neural network, they had improved performance on smaller objects. They had significantly increased detection mAP for KITTI dataset car class from 76.3% on their baseline Faster R-CNN detector to 83.6% in their improved detector.

Azevedo et al. (2019), they proposed the augmented-range vehicle detection system (ARVDS) to detect cars in a long range camera view using a deep convolutional neural network (DCNN) for the IARA self-driving car. They employed a bio-inspired foveated technique where the DCNN receives as input an image and crops of the image aligned with the projection of waypoints of the self-driving car's future path whose sizes increase with the distance from the self-driving car. They employed an overlap filter to discard detections of the same car in different crops of the same image based on the percentage of overlap of detections' bounding boxes. They evaluated the performance using the self-driving car an annotated evaluation dataset. Experimental results show that ARVDS increases the Average Precision (AP) of long range car detection from 29.51% (using a single whole image) to 63.15%. This method is able to detect vehicles at a large distance expending much less processing power than would be necessary if it had to examine the whole input image in several scales.

The other representation is the feature pyramid, which exploits the information extracted from multi-layer feature maps like Cai et al. (2016). The first attempt is to use the high resolution shallow layers to detect small objects, while using low resolution deep layers to detect large objects.

Cai et al. (2016) proposed a unified deep convolutional neural network, denoted the MS-CNN, for fast multi-scale object detection. The MS-CNN consists of a proposal sub-network and a detection sub-network. In the proposal sub-network, detection is performed at multiple output layers, each focusing on objects within certain scale ranges so that receptive fields match objects of different scales. These complementary scale-specific detectors are combined to produce a strong multi-scale object detector. The MS-CNN 85.11% of mAP with 0.4 seconds per image on KITTI Benchmark. The main disadvantage of these methods is that the feature maps in the shallow layers lack of the semantic information, they usually fail to distinguish the small objects accurately.

Some other works combine multi-layer feature maps together to train a network like Kon et al. (2016) presented HyperNet, a fully trainable deep architecture for joint region proposal generation and object detection. HyperNet provided an efficient combination framework for deep but semantic, intermediate but complementary, and shallow but high-resolution CNN features. HyperNet achieved 65.9% mAP and 96.16% recall on PASCAL Visual Object Classes 2012 test set (with IoU = 0.5) with video speed of 5 fps. The advantage of the

proposed architecture is its ability to produce small number of object proposals while guaranteeing high recalls.

These attempts take full advantage of deep layer information to tackle scale variations. However, due to the down-sampling operations used in the network, small objects cannot maintain sufficient spatial information in the deep layers, and thus they are still difficult to be detected.
To better maintain the deep feature maps of small objects, another solution like Zhang et al. (2019) and Liu et al. (2016) is proposed to use the high-resolution feature maps and the up-sampled deep feature maps together to predict small objects.

Liu et al. (2016) have introduced SSD, a fast single-shot object detector for multiple categories, discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location at the top of the network. At prediction time, the network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape. Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes. They built SSD models with at least an order of magnitude more box predictions sampling location, scale, and aspect ratio, than existing methods. SSD300 used 58.78 fps and achieved 74.18% mAP on DETRAC Benchmark. SSD512 used 27.75 fps and achieved 76.83% mAP on DETRAC Benchmark.

Zhang et al. (2019) developed DP-SSD, a single deep neural network for vehicle detection. The method proposed the use of different feature extractors for localization and classification tasks, and enhance these two feature extractors through deconvolution (D) and pooling (P) between layers in the feature pyramid. In addition, the network concatenates the feature pyramid of conventional SSD and extended the scope of the default box by adjusting its scale so that smaller default boxes to detect small vehicles more accurately. DP-SSD512 achieves 83.11% mAP with 0.1 second runtime. DP-SSD300 achieves 80.67% mAP with 0.07 second runtime on KITTI dataset. For the DETRAC test set, DP-SSD with the input size of 300*300 achieves 75.43% mAP at speed of 50.47 fps, and the framework with a 512*512 sized input reaches 77.94% mAP at 25.12 fps. The main problem of this solution is that the upsampling operation is performed on the entire feature map, which requires more memory resources and additional computational costs. While extra information leads to better accuracy, high computational cost is inevitable, which is unacceptable for our real-time vehicle detection task.

Hu et al. (2018) presented the SINet for fast detecting vehicles with a large variance of scales with two new techniques. First, a context-aware Regions of Interest (RoI) pooling to maintain the contextual information and original structure of small scale objects. Second, a multi-branch decision network to minimize the intra-class distance of features. These lightweight techniques bring zero extra time complexity but prominent detection accuracy improvement. The proposed techniques can be equipped with any deep network architectures and keep them trained end-to-end. SINet achieves 85.81% mAP with 0.11 second per image on (KITTI) benchmark dataset.

Another interesting work by Redmon et al. (2018) is YOLO, that reframes vehicle detection as a single regression problem. A single convolutional network that uses the initial convolutional layers to extract features directly from the entire image and the top fully connected layers simultaneously predicts multiple bounding boxes and class probabilities for those boxes as shown in figure (6). The non-max suppression method is applied at the end of

the network to merge the overlapping bounding boxes of the same object. To increase the accuracy YOLOv2 replaced the fully connected layers with an anchor box. YOLOv3 used a residual network and combines multi-scale prediction network. YOLOv3 achieves 87.42 map with 24.8ms per image on KITTI dataset. This method is relatively faster computational speed but it doesn't handle multiple aspect ratios which hurts accuracy, produces incorrect localization, each grid cell only predicts one class and it's hard to detect small objects appear in a group.

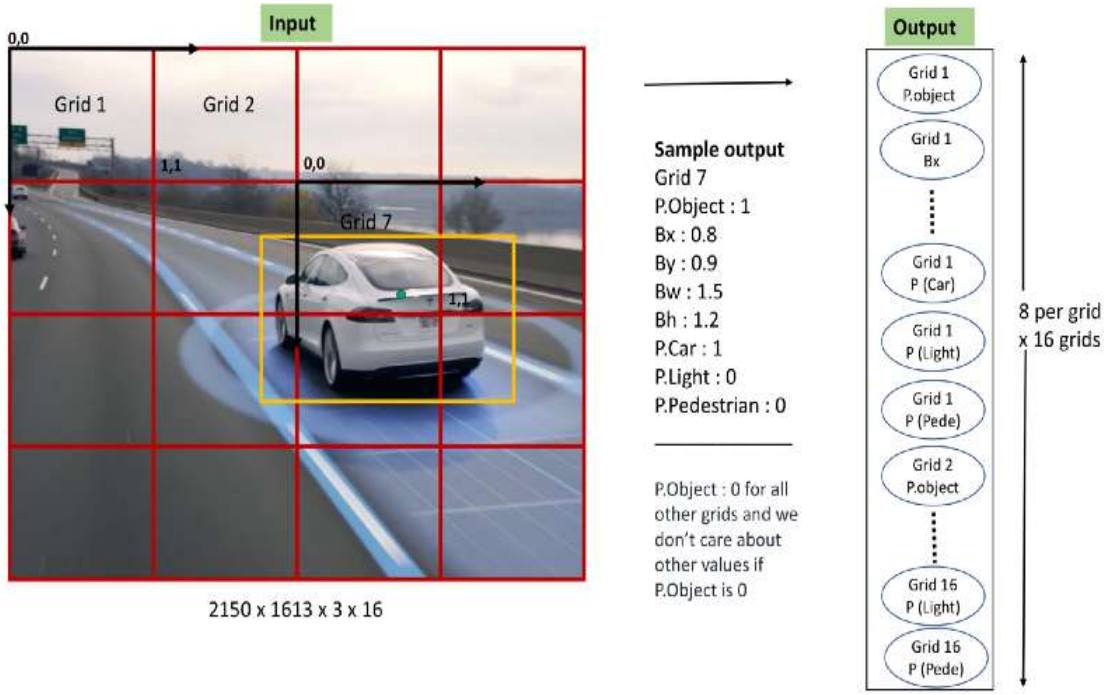


Figure (6): Bounding boxes, input and output for YOLO

Table 4: A comparison of convolutional neural network-based methods

Method	Year	Dataset	Precision (mAP)
Adapted Faster R-CNN	2017	KITTI	83.6%
ARVDS+DCNN	2019	IARA self-driving car	63.15%
MS-CNN	2016	KITTI	85.11%
HyperNet	2016	PASCAL VOC 2012	65.9%
SSD	2016	DETRAC	76.83%
DP-SSD	2019	DETRAC	80.67%
SINet	2018	KITTI	85.81%
YOLOv3	2018	KITTI	87.42%

6. Comparison and Evaluation

Motion Based Methods as the early works used to detect vehicles that are useful for free flowing traffic and insensitive to occlusions as shown in figure (7) but this kind of approaches is unable to distinguish the fine-grained categories of the moving objects such as car, bus, van or person. They can't recognize stationary vehicles as shown in figure (8). In addition, these methods need lots of complex post-processing algorithms like shadow detection as shown in figure (9) and occluded vehicle recognition to refine the detection results.

(a) input image (b) result

Figure (7): Background subtraction applied on occluded scene

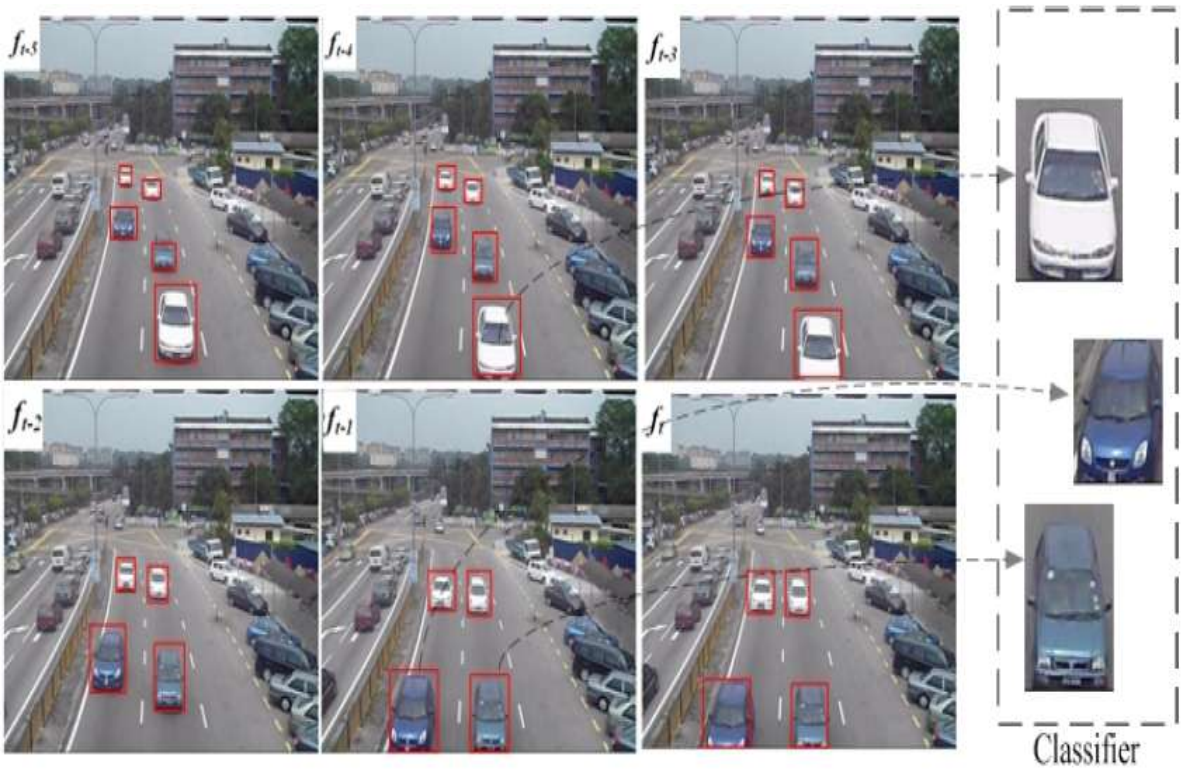


Figure (8) undetected stationary vehicles

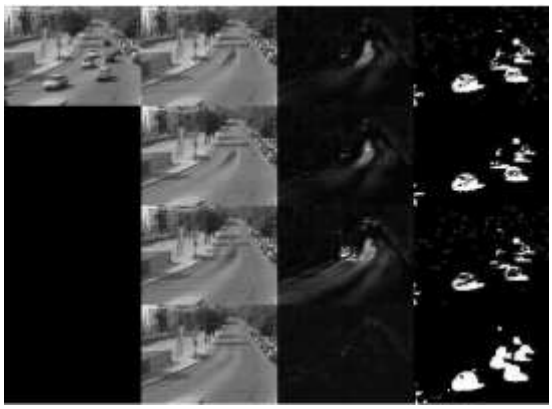


Figure (9) Shows shadow effects

Then, The statistical learning methods based on the handcrafted features are applied to detect the vehicles from the images directly. These methods have low complexity and it is fast and can recognize stationary vehicles as shown in figure (10) but if the features are not grouped accurately, then there may be failure in the detection of vehicle. And they have limited ability of feature representation, which is difficult to handle complex scenarios as shown in figure (11).

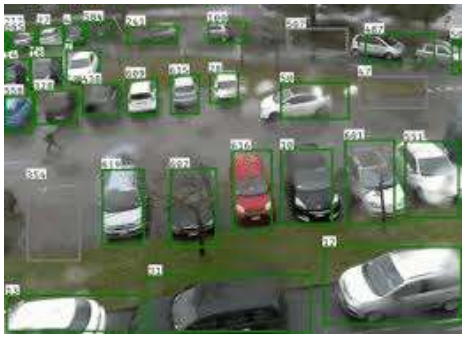


Figure (10) Detection of stationary vehicles

Figure (11) complex scenarios of vehicle detection

Now, features learned by the deep convolutional neural network show strong representation of the semantic meanings of vehicles as shown in figure (12), which make a great contribution to the state-of-the art vehicles detectors. Although these methods overperform lots of hand-crafted vehicle detection methods, the vehicles with a large variance of scales are still difficult to be detected accurately in a real-time manner as shown in figure (13) due to the scale sensitive convolutional features.

Figure (12) Semantic meanings of vehicles

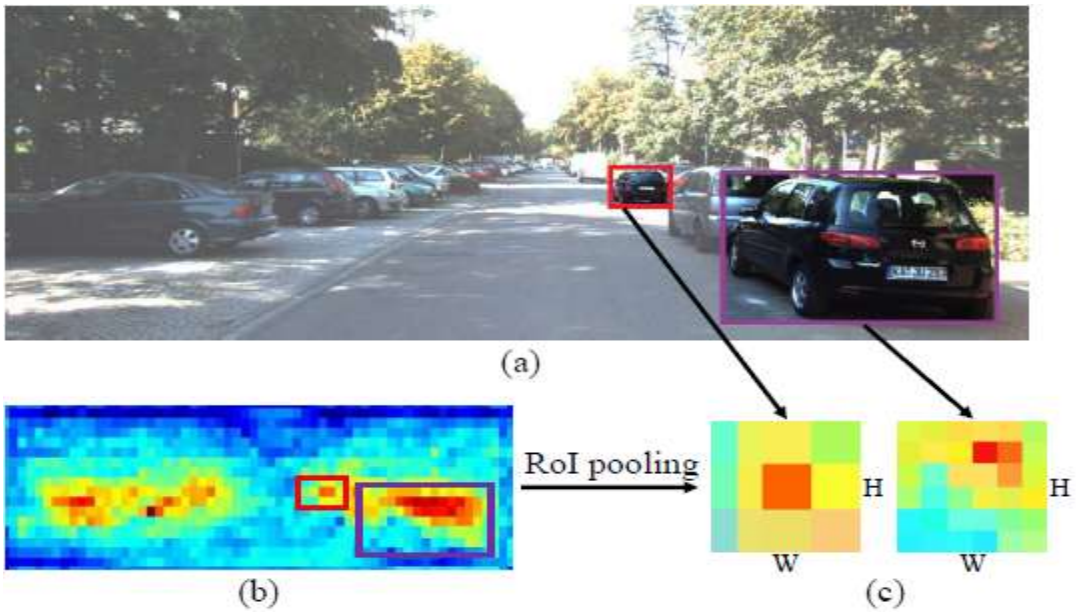


Figure (13) The scale-sensitive problem. (a) An image includes both large and small vehicles. (b) The feature representations of small and large vehicles in deep layer are largely different. (c) Traditional RoI pooling introduces noise as it simply replicates the values on the feature map for the small vehicle.

7. Conclusion

The most common problems in the task of vehicle detection are the problems such as heavily biased of the data set and different type of vehicles having different size and shape and different situations like occlusion, lightening and different weather conditions which is making harder to classify and detect vehicles. As a result of survey, no method outperforms the other ones at least in the same class. It is necessary to improve the robustness against the effects of the environment such as noise, illumination changes, occlusions etc. It is a big challenge for researchers to make a decision on which detection is more suitable. Combination of these methods can be used for the accurate detection of vehicles. Recent studies show CNN as a promising method for vehicles detection.

8. References

Agarwal, A. Gupta, S. & Singh, D. (2016). Review of optical flow technique for moving object detection. in IEEE Conference on CVPR.

Azevedo, P. Panceri, S. Guidolini, R. & Cardoso, V. (2019). Bio-Inspired Foveated Technique for Augmented-Range Vehicle Detection Using Deep Neural Networks. In IEEE.

Cai, Z. Fan, Q. Feris, R. & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. In ECCV,.

Cho, J. Jung, Y. Kim, D. Lee, S. and Jung, Y. (2019). Moving Object Detection Based on Optical Flow Estimation and a Gaussian Mixture Model for Advanced Driver Assistance Systems. Published in Multidisciplinary Digital Publishing Institute (MDPI).

Elkerdawi, S. Sayed, R. & ElHelw M. (2014). Real-time Vehicle detection and tracking using Haar-like features and compressive tracking. In 1st Iberian Robotics Conference pp.381-390.

Everingham, M. Eslami, S. Gool, L. Williams, C. Winn, J. Zisserman, A. (2012). The Pascal Visual Object Classes Challenge – a Retrospective. International Journal of Computer Vision.

Gao, Y. Guo, S. Huang, K. Chen, J. Gong, Q. Zou, Y. Bai, T & Overet, G. (2017). Scale Optimization for Full-Image-CNN Vehicle Detection. In IV, IEEE, pp. 785-791.

Geiger, A. Lenz, P. Stiller, C. & Urtasun, R. (2013). Vision meets robotics: The kitti dataset, The International Journal of Robotics Research. vol. 32, no. 11, pp. 1231–1237.

Hu, X. Xu, X. Xiao, Y. Chen, H. He, S. Qin, J. & Heng, P. (2018). SINet: A Scale-insensitive Convolutional Neural Network for Fast Vehicle Detection. In IEEE Transactions on Intelligent Transportation Systems (T-ITS), in Computing of Research Repository (CoRR), vol 1, pp. 1–10.

Huang, X. Cheng, X. Geng, Q. Cao, B. Zhou, D. Wang, P. Lin, Y. & Yang, R. (2018). The Apollo Scape Dataset for Autonomous Driving. In Institute of Electrical and Electronics Engineers (IEEE) Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 954-960.

Kong, T. Yao, A. Chen, Y. & Sun, F. (2016). Hypernet: towards accurate region proposal generation and joint object detection. In CVPR.

Li, B. Wu, T. & Zhu, S. (2014). Integrating context and occlusion for car detection by hierarchical and-or model. In European Conference Computer Vision (ECCV).

Liu, W. Anguelov, D. Erhan, D. Szegedy, C. Reed, S. Fu, C. & Berg, A. (2016). Ssd: Single shot multibox detector. In ECCV.

Maddern, W. Pascoe, G. Linegar, C. & Newman, P (2017). 1 Year, 1000km: The Oxford Robot Car Dataset. The International Journal of Robotics Research 36 (1), pp. 3-15.

Manzoor, M. & Morgan, Y. & Bais, A. (2017). Vehicle Make and Model Classification System using Bag of SIFT Features. 7th IEEE Annual Conference on Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA. pp. 572 577.

Ohn-Bar, E. Trivedi, M. (2015). Learning to detect vehicles by clustering appearance patterns. International Conference on Computer Vision (ICCV), Transactions on Intelligent Transportation Systems 16(52511–2521.

Red;mon J. & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arxiv1804.02767.

Salamon, J. Jacoby, C. & Bello, J. (2014). A Dataset and Taxonomy for Urban Sound Research. In Proc. of the 2nd international conference of Association for Computing Machinery (ACM) on Multimedia, pp. 1041-1044.

Wen, L. Du, D. Cai, Z. Lei, Z. Chang, M. Lim, H. Yang, M. & Lyu, S. (2015). UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking. arXiv:1511.04136 [cs.CV].

Yebes, J. Bergasa, L. Arroyo, R. & L´azaro, A. (2014). Supervised learning and evaluation of KITTIs cars detector with dpm. In Intelligent Vehicles Symposium (IV) IEEE.

Zhang , F. Li, C. & Yang, F. (2019). Vehicle Detection in Urban Traffic Surveillance Images Based on Convolutional Neural Networks with Feature Concatenation. Published in MDPI.

Ontology Based Expansion of Vague Queries

Manal Anwer Khedr[1] Fatma El-Licy[2] Akram Salah[3]

Abstract

The Internet is a valuable and important source of information. One of the benefits of Internet is searching for resources usually achieved with the support of a search engine. The right keywords in a query are a challenge. The search is generally performed using a set of keywords, which, in most cases, could have more than one meaning that leads to a contextless query and irrelevant results. Expansion of Vague Queries **(EVQ)**, framework has been proposed to add contextual and semantic information layer on top of the Google search engine. It is as attempt to improve information retrieval by expanding the initial short query. The proposed model is implemented, evaluated and compared against the regular performance of Google search engine.

Keywords: Query Expansion, Ontology, semantic web search, short queries.

1. Introduction

There is a massive amount of data available on the Internet which is growing, freely, daily. This free information-growth has not been assessed by a reliable information retrieval tool. The large percentage of the results obtained from a web-search does not produce relevant results. The Lack of knowledge, short queries and complexity to formulate appropriate queries, are challenges that lead to ambiguous queries, thus, irrelevant results.

Surveys indicate that almost ***25%*** of Web searchers are unable to find useful results in the first set of URLs that were retrieved [Mukhopadhyay et al., 2011].

To obtain more relevant documents, queries must be disambiguated by looking at their context. Query expansion techniques range from relevance feedback mechanisms to the use of knowledge models such as ontologies to resolve ambiguities [Khan and Mustafa 2014].It is not, always easy to specify the needed information using exact query terms, because of the

[1] Department of Computer Sciences. Faculty of graduate studies for Statistical Research, Cairo University. Email: ma_anwer@yahoo.com
[2] Department of Computer Sciences. Faculty of graduate studies for Statistical Research, Cairo University. Email: why2fatma@yahoo.com
[3] Department of Computer Sciences. Faculty of Computer and Information Sciences, Cairo University. Email: akramsalah.21@gmail.com

vocabulary problem. This problem increased by polysemy and synonymy [Bhogal et.al, 2007]. Some search engines have employed query refinement method to address short query problem. The idea is that, whenever a query is ambiguous or misspelled, the search engine will indicate, to the user, some more refined, narrowed and less ambiguous queries, to try for better results. Although query refinement method may handle the short query problem, successfully, to some extent, it suffers from several limitations and shortcomings [Sun et al., 2008].

This paper proposed system aims to reduce end users' efforts and tolerate their errors, by expanding their query into a specified domain. The specified domain knowledge is extracted from knowledgebase using Web Ontology Language (OWL). This approach helps in disambiguated queries and related data retrieval by providing query formulation to generate semantic formulation. This system works as both ontology-based query convertor and text-based search engine.

The objective of the presented system is to expand short query using ontology. The ontology language used in the proposed system is based on the Resource Description Framework (RDF) structure which was written in eXtensible Markup Language (XML) which enables them to be machine interpretable [Munir et al., 2018]. In order to expand vague queries, new terms are to be extracted as a natural language from a predefined vocabulary of the specified domain.

This paper is structured as follow: section 2 discusses Query Expansion. Section 3 presents related work. Section 4 describes the proposed system *EVQ*. Section 5 introduces system implementation. Section 6 evaluates the system. Section 7 displays results and discussion. Finally, conclusions presented in section 8.

2. Query Expansion

Query expansion (QE) also known as query augmentation, is the task of adding new meaningful terms to the original query to increase the number of relevant retrieved documents. The significant problem of query expansion is the choice of the expansion keywords. Terms that are similar and relevant to query terms are usually considered as good terms for expansion. The process of adding terms can either be manual, automatic or user-assisted. Manual query expansion relies on user expertise to make decisions on which terms to include in the new query. In the case of automatic query expansion, weights are calculated for all terms and the terms which have the highest weight are added to the initial query. With

user-assisted query expansion, the system generates possible query expansion terms and the user selects which of these to include [Bhogal et al., 2007].

Several semantic QE approaches, including linguistic and ontology-based and mixed-mode are available for addressing the vocabulary mismatch problem [Raza et al., 2019].The main advantage of semantic QE using ontology domain is that, the knowledge structure is always available in the expansion mechanism, which clarifying the context of the query. Therefore, it was adopted in the proposed system to expand vague queries. In this semantic QE, the query term is expanded with super categories in the knowledge hierarchy, then expanded terms are processed by search engines in a Boolean way using AND operator.

The main shortcoming of using ontology as a QE approach is that it relies on exact matching between the query and the ontology vocabulary, which is a complex task for user queries that are written in natural language. To overcome this shortage, techniques are used to measure the similarity to decide which term or expression from ontology is the closest to user's query.

3. Related Work

There are a lot of work that have been accomplished by researchers in the field of query expansion. Most of them relay on external sources to provide supporting for queries expansion. Some of those resources are syntactic and others are semantic. QE often bridge the vocabulary gaps between queries and web resources [Fang, 2008].

[Sun et al., 2008] presented a method for refinement short queries. They used query retrieval models that construct related multiple derived queries for the user's query. They ranked each of the derived queries according to its similarity to the user's query. This method is useful for improving query refinement and constructing the final results.

[Imran and Sharan, 2009] constructed thesaurus as a help tool for, semantically, selecting related terms to expand the query.

[Sarmah et al., 2013] developed system for retrieving Assamase unstructured documents from digital library. They used Assamase WordNet and query expansion techniques, to extend user's original query.

[Khan and Jaafar, 2011] proposed a novel approach for the query expansion using WordNet and ConceptNet, which discarded the less important words from the query semantic space.

[Tran, 2016] provided a method to match a company's human resources with job assignments received from clients. He used ontologies as a solution to implement a query

augmentation that will improve defining the context by adding suggestions of relevant words.

[Rajasurya and Swathi, 2012] developed a semantic search engine- SIEU in the university domain. They used ontology as a knowledge base for the information retrieval. It was one layer above a search engine that retrieves more relevant results by analyzing just the keywords.

[Al-Khateeb et al., 2016] proposed enhanced system to expand the user query using WordNet and generates query list to obtain results, and then select the best and relevant one. Based on above related work, the usage of WordNet in query expansion, still poor for domain specific because it aims to cover general English words. It is obvious using ontology in expansion process, where semantic web provided tools and languages to facilitate machine readable access to ontology such as RDF and OWL.

4. The proposed system

The main functionality of ***EVQ*** system is handling a vague query by expanding it with context which generated from domain ontology.

EVQ system facilitates the searching tasks for internet user, by targeting the search onto specific domain, and expanding the search query to be significant and relevant to the domain of interest. ***EVQ*** system is not intended to change the current keyword- based search, but supplements keyword search as attempt to handle short queries.

4.1 Modeling the Ontology Domain

Figure 1 illustrates the main steps for ontology development.

Figure 1: Summary of development ontology

- **Concept identification:** Identifies the domain and Prepare a list of possible concepts to be included in the domain ontology. In this research project, a manual mechanism is applied to extract the domain terms from various online resources [Data, 1999], [JavatPoint, 2011], [Oracle, 1995] and [BeginnersBook 2012]. The Concepts that were used in ontology are the core concepts of the domain of interest.
- **Concept organization:** Determine the key concepts, such as the upper-level concept and the commonly used concepts in this field, in order to establish the core concept sets. Then Define the relations and assigned them to concepts, which are usually, “is-a” relations.

The ontology organized the domain items into a hierarchical tree, in which nodes of the ontology are concept words, and the edge is the relationship between ontology concepts. Then select the top down structure approach to be utilized to be consistent with the format in which information wasprovided in the textbooks/resources.

- **Representation & Evaluation:** The generated concepts are organized into the ontology representation using editor and be adopted for defining classes and class hierarchy. The ontology class hierarchy shows the relationships among upper-level and lower-level classes.
    - Using ontology to provide additional information for each concept.
    - The type of created Ontology is Light-weight Ontology which is hierarchical or classificatory characterized. It designed to represent relation (is-a hierarchy between concepts). It was selected because a light-weight Ontology does not include too many or too complicated relationships [Slimani, 2014].
    - Evaluate the consistency and check the correctness and validity of the generate ontology using the system reasoning to be sure that ontology does not allow any contradictions.

4.2 System Architecture

Figure 2 illustrates the architecture of the proposed system, each of its components are described in the following subsections.

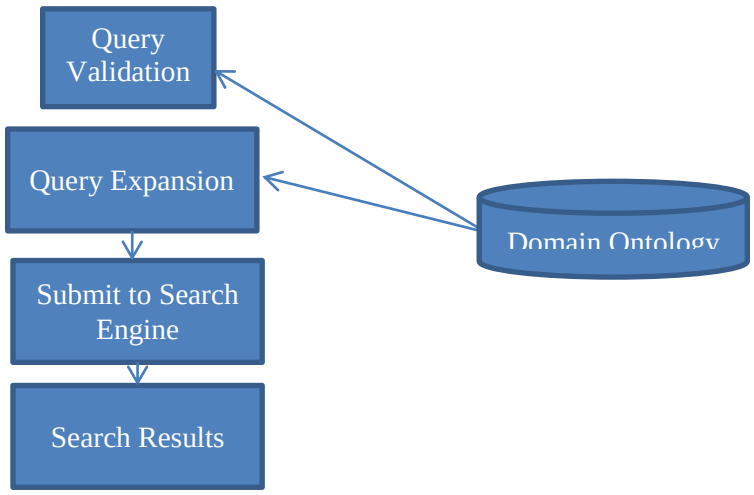


Figure 2: Architecture of the proposed system

4.2.1 Query Validation

In this stage, query is processed for validation and is utilized to extract a list of the matched ontology terms. This stage performs the following tasks:

- Be certain the query not empty.
- Retrieve all ontology candidate terms and keep it into a list.
- Validate the search query to match ontology classes prior to the expansion phase to guarantee that the query belongs to the domain of interest using similarity method.
- Keep the matched term.

4.2.2 Expanding the Query

Expanded Query layer is one of the basic layers of the system model. It provides the semantic for initial query concept by adding new terms from ontology. It utilizes the generated ontology list to perform two tasks, namely: Computing similarity matching and query expansions.

1- Computing Similarity

A significant step in the process of semantic query expansion is the similarity computing to decide how close user's query with ontology concepts is. Then return the related concept with the highest similarity to the system for expanding process.

Cosine method was adopted to measure the highest, syntactically similar. This step is to overcome the exact matching to ontology concepts and misspelling problem.

2- Expanding the Query

Expanding trail started with specific matched term as a bottom-up direction on one path to obtain all super direct and indirect classes to be concatenated together by *AND* operator. So the context of the searched Concept is determined by extracted set of directly and indirectly linked concepts in the ontology by explicit relations.

4.2.3 Applying search engine

The expanded query is to be utilized and redirected to Google search engine to obtain the results of its semantic search. Thereafter, the list of the obtained web pages is retrieved to the user.

5. System Implementation

The proposed system was implemented into the model prototype ***EVQ,*** as shown in

Figure 3. It illustrates the flow of the data through system components. In which, user query obtained from the user interface is validated to obtain a validated query. This query is expanded and delivered to the search engine that retrieves the search results.

This implementation focuses on create and execute a web application using JavaServer Faces (JSF) as a GUI. And used OWL API as a semantic web tool which can read, write and process data in domain ontology, via Java programming language within Eclipse editor. Finally, used the JavaServer Faces API, to capture and process the data from user interface.

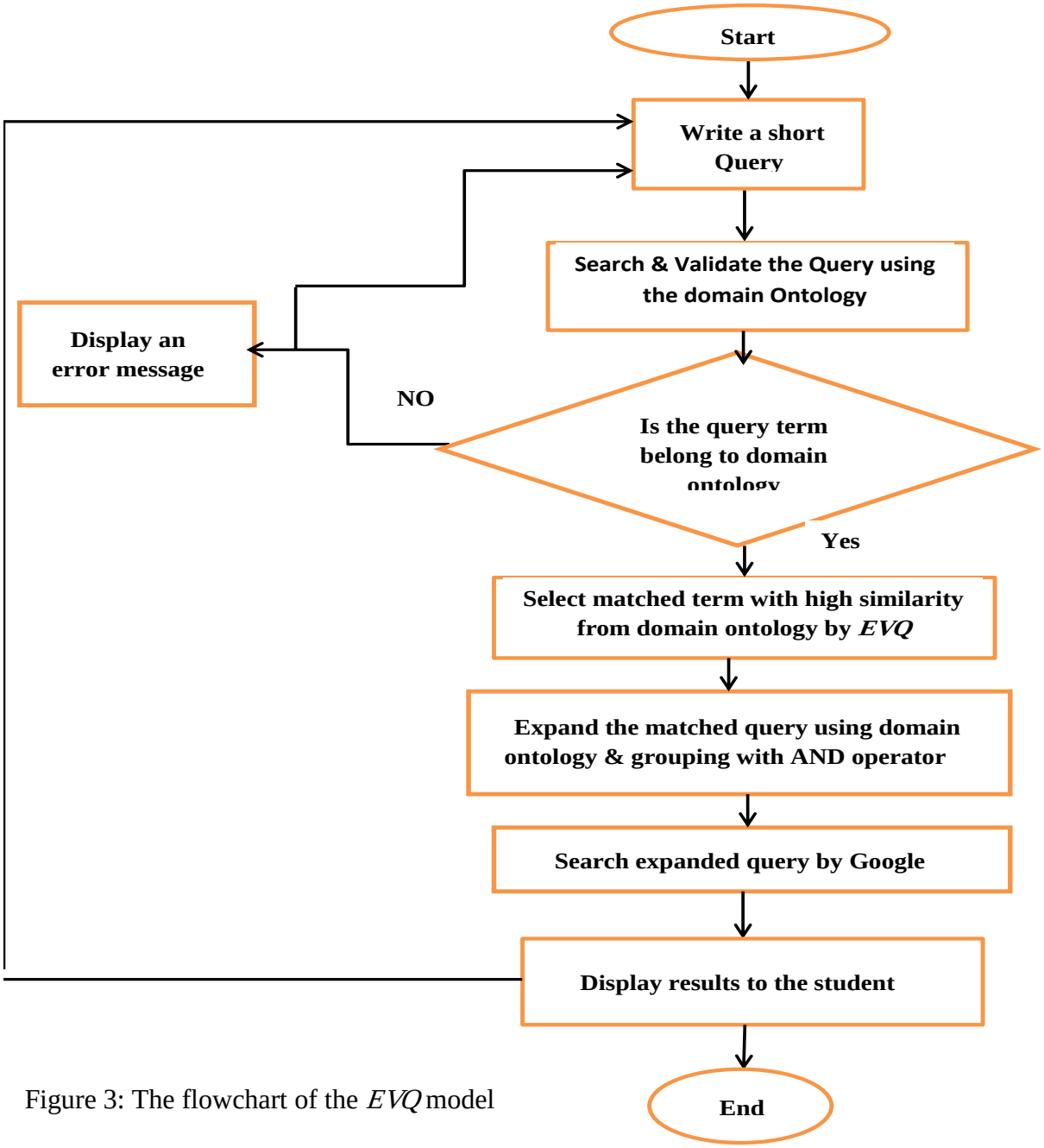


Figure 3: The flowchart of the *EVQ* model

5.1. Modeling Ontology for the Domain of Interest

The ontology was developed using Protégé editor in OWL/RDF to define classes and class hierarchy. The ontology covered ***Java 2*** Standard Edition (J2SE) to be as a context for learning Java Programming.

5.2. Validation Process

This process developed to insure that the user query belongs to JAVA domain. The main function here is the similarity method and the result from this phase is a matched concept or

display wrong message.

5.3. Expanding Process

This phase used the OWL's methods to generate direct and indirect super classes for the matched concept which is produced from previous phase. Then, refine the initial query by concatenating these super classes and matched concept together using AND operator [Karch, 2018] to compose the new (refine) query.

5.4. Searching Process

This phase will exploited the new (refined) query to be a parameter within Google URL, and then redirect it to the Google search engine as an external resource using JavaServer Faces API.

5.5. Display search results

5.6. Running Example

This section presents a running example to illustrate the flow of data through the system to obtain the, relevant search, for a given vague query. First of all, the Ontology of the interested domain was generated. That is accomplished by defining the concepts and the roles of the Java language domain and translating them into Ontology. Figure 4, illustrates the Semantic context of one of the Java programming language, namely: 'Public' Concept. The semantic context of the Concept' **Public'** in the Java ontology is formed by the concepts '**Modifiers'**, ' **Access Modifiers** ' and ' **Java Programing** ' (all of them are super classes for the concept ' **Public**').

The implemented system executed through the GUI interface. The following is a typical scenario for an execution session:

1- A user wrote the query " **Publik** ", through the GUI interface
2- The query is validated (using Validation and similarity test) to be **- Public**'
3- A selected matched ontology terms were generated for the verified query (using OWL's method) to be:
   "Modifiers", "Access Modifiers" and *"Java Programing"*
4- The verified query was expanded (using Expanding method):
   By combined the matched term with its context using ***AND*** operator to compose the logical formula "*Public AND Modifiers AND Access Modifiers AND Java Programing*." Meaningful terms (concepts) that are closer to the user requirement.
5- The Expanded query was applied to Google URL as a searching parameter then

redirected to **Google** search engine.

6- The matched relevant web sites were displayed for the user.

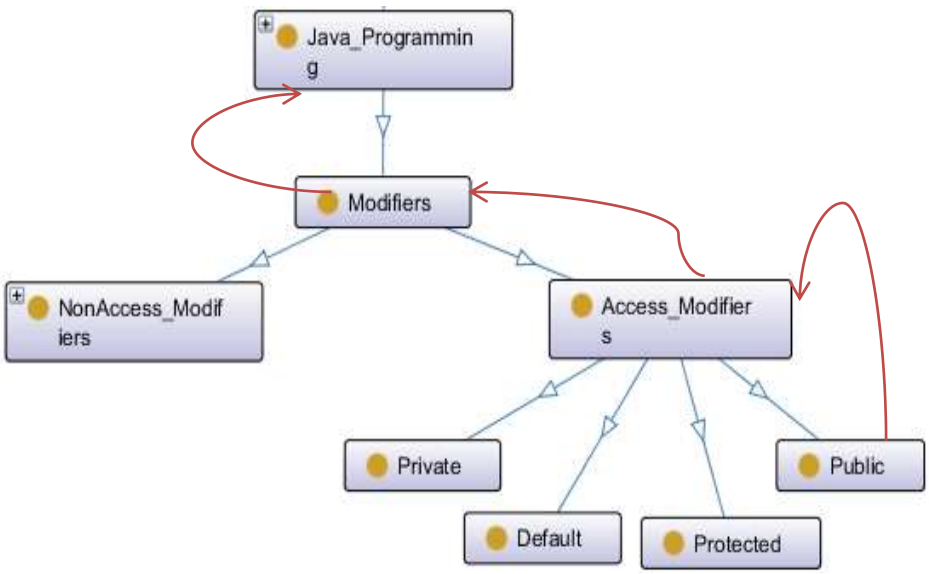


Figure 4. Semantic context of the Concept ' **Public**'

- **Similarity Method**

The similarity methods are available library on the web [Debatty, 2018], therefore, the algorithms of similarity was extracted from its code, as shown in Figure 5.

```
// Compute the cosine similarity between two strings.
Declare String s1, String s2;
Input: String s1, String s2;
Output: cosine ɵ similarity;

{
          For each (string) Do
1- Divide string into terms with length k characters.
2- Combining the both string terms into a set of unique
   terms, ignoring order of terms.
3- Count the number of frequency of each terms occurrences in
```

```
   each string.
4- Get the two vectors (a and b) of counts.
5- Compute the cosine similarity between those vectors using
   equation
        COS (ϴ) = ∑_{i=1}^{n} ai*bi / (√(∑_{i=0}^{n} a_i^2) √(∑_{i=0}^{n} b_i^2))

      Return ϴ.
}
```

Figure 5: Cos-Similarity algorithm

6. System Evaluation

The system was evaluated by comparing the fitness function and precision of the top ranked *K* results for the query results from *Google* and that of the expanded query resulted after applying *EVQ* system.

6.1. Fitness function

Fitness function is used to evaluate the quality of all expanded queries**.** Maximum fitness (score) is an indicator for optimal retrieval. The system computes the fitness like [Al-Khateeb et al., 2016] with few changes.

The fitness function was applied using *jaccard similarity* method between the query and the title, description and URL-name, for each of Google's retrieved results (with and without query expansions) using equation (1).

$$Fitness = A*Order + B*Sim1 + C*Sim2 + D*Sim3 \quad (1)$$

Where:

Sim1= Jaccard_similarity between ***query*** *and the* ***Title,***

Sim2= Jaccard_similarity between ***query*** *and the* ***Description,***

Sim3= Jaccard_similarity between ***query*** *and the* ***URL_Name,***

Order is the order of the web page in Google Results

A=0.1, B=0.2, C=0.3 and D=0.4;

Where (A, B, C and D) are factors to represent the degree of importance.

Jaccard_similarity(String1, String2)=$|String1 \cap String2| / |String1 \cup String2|$

Where String1, String2 are sets of data.

The average fitness was computed for top ranked ten results using equation (2).

$$Average\_Fitness = \frac{\sum_{i=1}^{10} Fitness}{n} \quad (2)$$

6.2. Precision@*K*

Precision (**P**): the proportion of retrieved documents that are relevant.
Precision@*K*: proportion of top-***K*** documents that are relevant [Craswell, 2009].

To compute Precision@*K*:
1- Set a rank threshold *K*
2- Compute the percentage of relevant in top *K*
3- Ignores documents ranked lower than *K*
4- Finally compute Average of *P@K* for every query.

Average of $P@K = = \frac{\sum_k p@k}{R}$

Where *k* is rank of each relevant document, *R* is a total number of relevant documents and *p@k* is the precision of the top-k retrieved documents [Zhang, 2009].

6.3. System Evaluation by external users

The *EVQ* system evaluated by two users with background about java programing and every user apply 45 queries with single word. Most of these queries processed an expanded by ontology and some of them are not, since those terms are not included into ontology domain. The below table1 as a feedback which presented the percentage of acceptance of *EVQ* system.

Table 1: Evaluation based on P@K of the two users, using Google and *EVQ* system, *k* = 10.

users	Number of queries	Satisfied queries	Unsatisfied queries	Average p@10 With ***EVQ*** %	Average p@10 With ***Google*** %
1	45	37	8	82.2	12.2
2	45	41	4	91.1	18.6

For every single query, the users tested first ranked 10 results before and after expansion.

The Satisfied queries imply that queries expanded from created ontology domain but Unsatisfied queries did not.

6. Results and Discussion

A test *EVQ* system by external testers with a sample of a single word query as table 1 was illustrated that expanded queries implies more relevant results than shortage ones. Some queries was contextless because the lack of ontology terms.

Also it's prepared a sample of queries to measure the performance and accuracy of the system. This randomly selected seven short queries; each query was applied as a search key into ***Google*** search engine, then into ***EVQ*** system. The precision and fitness function were

calculated for each of the queries as shown in Table 2.

The test sample had been designed to show how much the Ontology can improve retrieving results to be more relevant to the user needs. The results of Google alone and that after applying ***EVQ*** expansion system were evaluated using fitness function and Precision @10. Table 2 illustrates a comparison of the retrieval accuracy of queries before and after its expansion by the proposed system. Figure 6 shows the fitness function for the tested queries before and after the application of proposed system through Google search engine. Whereas, Figure 7 shows the precision of the tested queries, before and after, its expansion through the proposed system.

Table 2: Comparison of the retrieval accuracy between Google and the *EVQ* system, $k = 10$.

Query	**Initial Query + Google**			**Extended Query + Google**		
	Initial Query	**Average**		**Expanded Query**	**Average**	
		F.F	**P@K**		**F.F**	**P@K**
jva	Jva	61.39%	0%	Java Programming	74.16%	100%
Constructor	Constructor	61.43%	90.8%	Constructor AND Methods AND Java Programming	70.04%	100%
Loop	Loop	65.5%	0%	Loop Statements AND Statements AND Control Statements AND Java Programming	74.12%	100%
List	List	63.87%	20%	Lists AND Java Programming AND Collections	73.49%	100%
Catch	Catch	60%	0%	Try Catch AND Java Programming AND Exception Handling	78.52%	100%

Abstraction	Abstraction	67.94%	54.7%	Abstraction AND Java Programming AND Object Oriented Concepts	76.87%	100%
For	For	60%	0%	For AND Loop Statements AND Statements AND Java Programming AND Control Statements	75.04%	100%

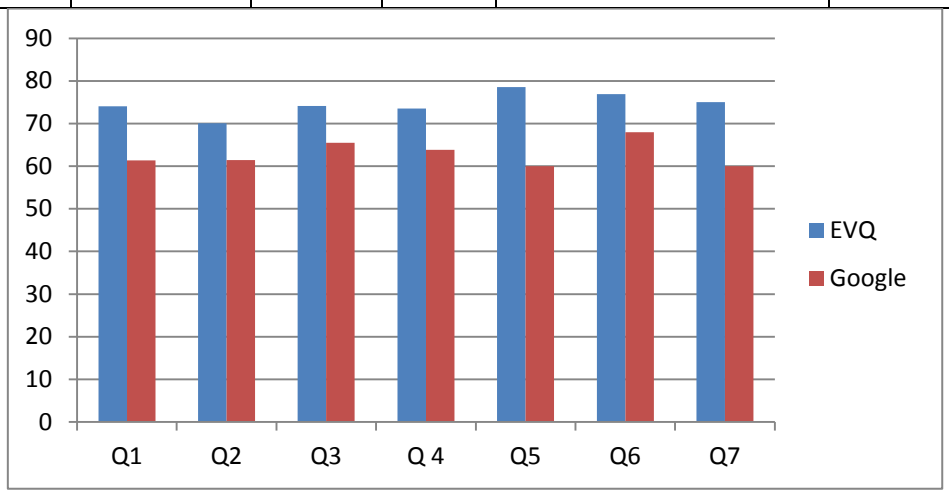


Figure 6: Fitness for tested queries in *EVQ* system and Google engine

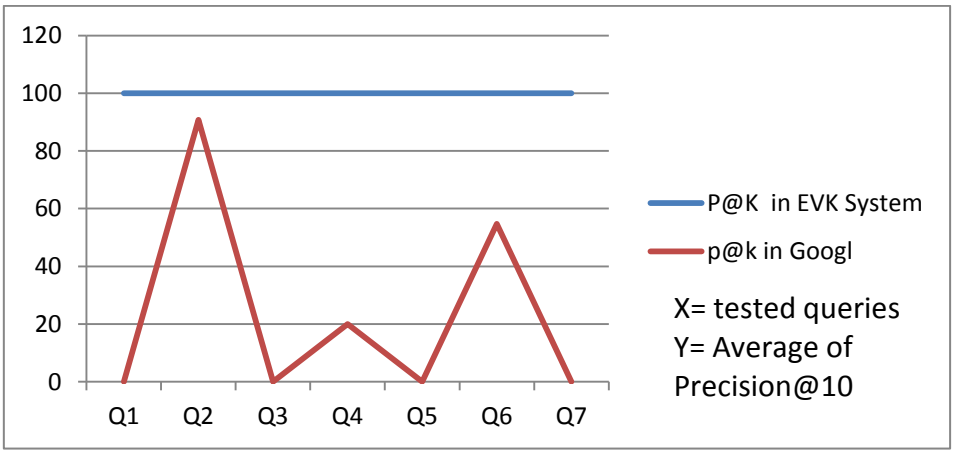


Figure 7: Precision for tested query in *EVQ* system and after Google engine

The evaluation of obtained results from proposed system, a semi-automatic process was used. Since there is no Google APIs support exact retrieval like Google search engine. And planning in future to be full-automatic.

7. Conclusions

The semantic search is convenient because of its richness and effective. This semantic search provides a wide set of related results regarding the user's requirements and thus, it was appropriate to be adopted. The WordNet is not directed to a particular domain. It is mainly language relation rather than semantic or context relation. This research paper employed users query from expansion by generating semantic from specific domain ontology to guarantee a better results from Google search engine.

Synonyms and polysemy are critical issue of vocabulary mismatch, which is forced by short queries. The developed domain ontology improved the search process by adding context to search query. Supply more related terms to ontology domain; will curing the lack of expansion. Derived ontology is sharable and reusable. The proposed framework can be portable for any given domain by generating the corresponding domain ontology.

Acknowledgment

Thanks and appreciation to Suzan Alsisi, Principle Advanced software Engineer for Fusion Middleware at Oracle (Egypt) and Abdllah Mohmed, .Net Team Leader at Infoglobe Company (Egypt), for their time and effort in testing the proposed system and provide their feedback.

References

Al-Khateeb, B., Hilal, A.J., and Al-Janabi, S. (2016). Web Search Enhancement Using WordNet Query Expansion Technique. Journal of University of Human development, pp. 542-547.

BeginnersBook (2012). Java tutorial Available at: https://beginnersbook.com/java-tutorial-for-beginners-with-examples/ last visited in 15/4/2018.

Bhogal, J., MacFarlane, A. and Smith, P. (2007). A Review of Ontology-Based Query Expansion," Information processing & management, vol. 43, pp. 866-886.

Craswell N. (2009) Precision at n. available at: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_484 last visited in 20/8/2019.

Data, R. (1999). Java tutorial. Available at: https://www.w3schools.com/java/ Last visited in 15/4/2018.

Debatty, T. (2018). java-string-similarity Available at:
https://cylab.be/research/java-string-similarity#shingle-n-gram-based-algorithms
last visited in 9/8/2019.
Fang, H. (2008). A Re-examination of Query Expansion Using Lexical Resources". Association for Computational Linguistics (ACL) pp 139–147 June.
Imran,H., and Sharan, A.(2009) Thesaurus And Query Expansion. international Journal of Computer science & Information Technology (IJCSIT), vol 1, No 2, pp 480-484.
JavatPoint (2011). Java tutorial. Available at:
https://www.javatpoint.com/java-tutorial last visited in 15/4/2018.
Karch, M. (2018). How to Do a Boolean Search in Google. Available at:
https://www.lifewire.com/boolean-search-terms-google-1616810 last visited in 29/3/2019.
Khan,R., and Jaafar, J. (2011). Semantic Query Expansion Using Knowledge Based for Images Search and Retrieval. International Journal of Computer Science & Emerging Technologies Volume 2.

Khan,S., and Mustafa,J. (2014). Effective semantic search using thematic similarity Journal of King Saud University – Computer and Information Sciences, vol. 26, pp.161–169.
Mukhopadhyay, D., Banik, A., Mukherjee, S., and Bhattacharya, J. (2011). A Domain Specific Ontology Based Semantic Web Search Engine.
Munir, K., Sheraz, M. A. (2018). The use of ontologies for effective knowledge modelling and information retrieval". Applied Computing and Informatics vol. 14, pp 116–126.
Oracle (1995). The Java Tutorials Available at:
https://docs.oracle.com/javase/tutorial/java/ last visited in 15/4/2018.
Rajasurya and Swathi (2012). Semantic Information Retrieval Using Ontology in University Domain. International journal of Web & Semantic Technology. vol 3, pp55-68.
Raza, M. A., Mokhtar, R., Ahmad, N., Pasha, M., and Pasha, U. (2019).
A Taxonomy and Survey of Semantic Approaches for Query Expansion, in IEEE, vol. 7, pp. 17823-17833.
Sarmah, J., Barman A. K., and Sarma, S. K. (2013). WordNet Based Information Retrieval System for Assamese" Proceedings of Computer Modelling and Simulation (UKSim), 15th International Conferenc, pp 89-97.
Slimani, T.(2014). A Study Investigating Typical Concepts and Guidelines for Ontology Building. Journal of Emerging Trends in Computing and Information Sciences.Vol. 5, pp. 886-893.
Sun, B., Liu, P., and Zheng, Y.(2008). Short Query Refinement with Query Derivation. AIRS'08 Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, pp. 620-625.
Tran,H. (2016). Human resource matching through query augmentation for improving search context. Master thesis.
Zhang, E. and Zhang, Y. (2009) Average Precision. Available at:
https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_482
last visited in 20/8/201

A Flexible Query Approach for Geographic Information Systems

Ahmed M. Rashad **Ahmed M. Gadallah** **Hesham A. Hefney**

Faculty of Graduate Studies for Statistical Research

{cs887, ahmgad10} @yahoo.com, hehefny@ieee.com

Abstract: Geographical information system has been use in many domains. The power of GIS comes from the ability to relate different information in a spatial context and to obtain details about this relationship. Searching for a specific address on a GIS system is a very challenging problem, especially when we search for an address using human-like text. Unfortunately, a human description of an address to reach them in a GIS system is a complicated problem due to the following causes: The human description of an address may be semi-structured or unstructured at all. Accordingly, there is a challenge to analyze such described address. This work attempts to recover such problem by analyzing a given address then put it in a formal structure that can easily be use to retrieve the desired address on the underlined GIS.

Key Words

GIS, Ontology, NLP, Unstructured Information Retrieval

1. Introduction

The use of geographical information systems (GIS) with the purpose of managing, analyzing and presenting data with reference to cartographic geography has grown rapidly during recent years. A key concern in GIS applications regards the storage, handling and presentation of data from various domains of activity at a common set of geospatial coordinates with the purpose of swiftly identifying and accessing the data available. The development and usage of geographical information systems (GIS) based on many disciplines such as knowledge representation. Ontologies, as specifications of conceptualizations, are possible tools to be use in GIS. They enable sharing terms used for information description and provide a basis for sharing, processing and integration of data. So, ontology can be used to access any address in GIS easily. Using Natural language processing tools to match various address, and to process expected mistakes in the text of addresses.

1.1 Geographic Information System

GIS is a computer program that connects geographic information (where things are) and descriptive information (what things are). It was develop to create thematic maps that support GIS data capture, data storage and data analysis. The power of GIS comes from the ability to relate different information in a spatial context and get details about this relationship. GIS can reveal important new information leading to better decisions. Also, convert existing digital information, which may not yet be in the form of a map, into forms that it can recognize and use. For example, digital satellite imagery can be analyze to produce a map of digital information about land use and land cover (Zhu et al. (2018)).

1.2 Natural language processing

Natural language processing is a set of computational techniques for the analysis and representation of naturally occurring texts at one or more levels of parsing for the purpose of achieving them as human language processing for a set of specific tasks or applications. The goal of pre-processing is to represent each document as a feature vector, that is, to separate the text into individual words. Keyword selection, which is the feature selection process, is the basic pre-processing step necessary to index documents (Srividhya and Anitha (2010)).

Stop Word Removal is remove Stop words, which has proven as very important. Examples of such words include 'the', 'of', 'and', 'to'.

Stemming techniques used to find out the root/stem of a word. For example, Search, Searching and Searches all have Search as the root stem.

1.3 Ontology

Ontology is a clear specification of common concepts, a key structure for the representation of knowledge, especially for the representation of complex or changing data. (Haruna and Ismail (2016)) Ontology consists of sets of axioms and facts. Assuming the open world, when lost information is treat as anonymous. No unique name assumption, individuals may have more than one name. Query answers reflect both schema and data. Can handle incomplete information (Partyka et al. (2008)).

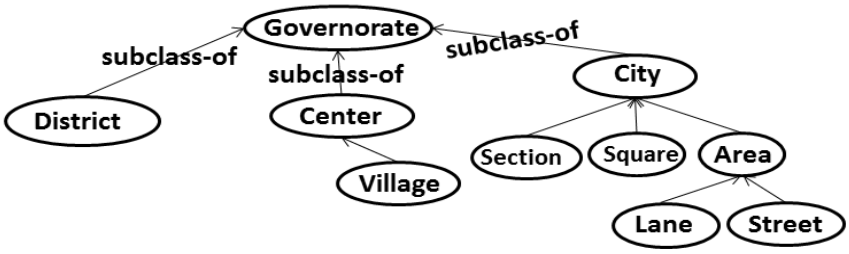


Figure 1: Example for GIS ontology

2. Related Work

Gao and Goodchild (2013) suggest a semantic framework for designing a question-based user interface that integrates different levels of ontology (ontology for spatial concept, field ontology and task science) to guide the process of extracting basic spatial concepts and translating them into a set of mathematical or equivalent equivalents geographic information system tasks. The semantic framework for the question-based GIS user at the vertical level can be divide into three layers: the question layer, the semantic layer, and the implementation layer .First, **Question Layer**: The interface allows users to ask spatial questions from their ideas or specify some examples of questions included in the user interface. Second, **Semantic layer**: natural language processing methods used to analyze questions on basic spatial concepts by referring to spatial existence. He also uses a specific field ontology to understand the descriptive structure of questions to capture a user's goal. Semantic output is a single task or a multi-step task that must be perform to answer user questions. Third, **Implementation layer**: resort to possible GIS tools or functions to get the job done on the GIS desktop program. Example of use, **Question 1:** How far is it from Los Angeles to Santa Barbara? This is a simple question. The semantic layer realizes that the word "how far" is the concept of distance and thus provides the distance function to the user. **Question 2:** What areas are affect by this wild fire? This top-level cognitive question requires some knowledge about spatial relationships and the environment (such as the speed of propagation and the direction of fire).

(Derbal et al. (2016)) suggest approach consists of three units: First, **Query Analysis Unit**: It performs a lexical analysis of the text query, such as removing special characters such as punctuation and symbols, canceling the list of stop words in the English language, and then exploiting the WordNet glossary. Moreover, define the relevant terms extracted from the query, which then used to explore geographical ontology. The terminology used to explore geographic scope retained to define the corresponding map layers. Secondly, **a user profile-learning module** aims to link a mapping user profile to a web user. User orientation for captured mapping Building an existential science called "CartoUserOntology" aims to represent different categories of mapping users. Thirdly, **filter unit**, which aims to define and standardize the recovered layers in the first stage that had better meet the requirements of the user query based on "CartoUserOntology". **An example of using** the approach "I am looking for some hotels in Babeloued" is equivalent to "some hotels in Babeloued" or "hotels in Babeloued". For example, the second query is "Some Hotels in

Babeloued". The language-processing unit takes inputs providing these two terms, Hotel and Babeloued. Babeloued is a location in the Administrative Boundaries layer, the "hotel" defined as the mapping layer, and all its instances will be retrieved and mark on the map. However, a "hotel" in the field of geography (with ties) is associated with many more specific classes such as restaurants, resort, breakfast and family hotels. All of these layers will be recover. The filter unit allows the merging of road and road name layers as master maps. Where required hotels are marked as descriptive information that integrates geographic location names and road maps; they are dynamically visualized at the bottom (left) of the visible map.

(Pillai and Karabatis (2018)) create a multi-geospatial model to help retrieve geographic information. It contains main elements, **GIS layer ontology**, which determines how different GIS layers interact with each other. How does a specific address relate to a park, fire station, or school district? How school districts related to zoning in the area? Ontology describes the relationships between different classes of organisms. There is a relationship between schools and streets in neighboring schools. **Inter-Layer Correlation Matrices**, when the relationship between two classes of organisms defined, the strength of the relationship between them is determined as a probability value**. Geographical relationships feature.** Its function is to identify distinct points for a particular class, for example, if a point of interest is identified, then streets are identified that are relevant to the inquiry .**Supra-Adjacency Matrix. I**t combines adjacency matrices and layer-link matrices (to determine the relationships of geographic objects to different classes). This matrix determines the relationships between a specific geographic object and all objects in all layers. **An example of usage,** point identification, searching for a related item, and setting the search frame around it in radius value. The greater the radius, the more results with less accuracy, the results can be reduced by using more than one layer in the search.

This paper proposes a new approach that improves the search for unstructured addresses in Arabic written in different formats. Spelling correction of words with an account of weights for each word according to the content of its letters. Remove unnecessary stop words in the address and its related words, such as “floor seventh” or “apartment No. 3”. The possibility of using another name for the individual. Can identify all parts of the text of the address of the research if needed and convert the query address search from an unstructured formal to a structured formal.

3. The Proposed Approach

The main objective of this work is to a human-like query approach for GIS systems. Using some natural language processing tools and ontology, we aim mainly to tackle the problem of searching about some locations in a GIS using human-like queries like a description of some addresses. Consequently, a preprocessing process takes place aiming to convert such information about addresses into an Ontology as shown in Figure 2.

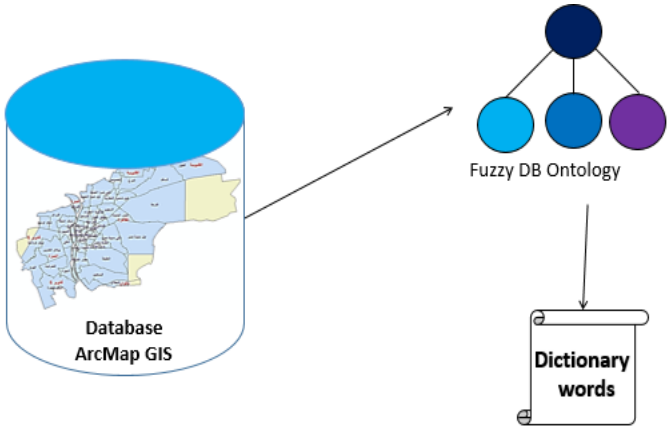


Figure 2: Pre-processing for Proposed Approach

We begin by dragging Database ArcMap GIS data into Database ontology to store data in ontology, and in ontology, the main class of different address and terminology sections recorded for each class .In addition, a dictionary of words and the weight of each word is prepared according to the sum of the weights of the letters formed.

3.1 The Architecture Of The Proposed Approach

Commonly, the proposed approach starts by receiving the search text from the user. Consequently, it makes the spelling correction of the words, and re-output the text after the correction. It deletes the stop words in the search for a location such as (the floor number or apartment number) and the preparation of the final sentence of the search. Then create an array to record the parts of the address that the program will recognize.

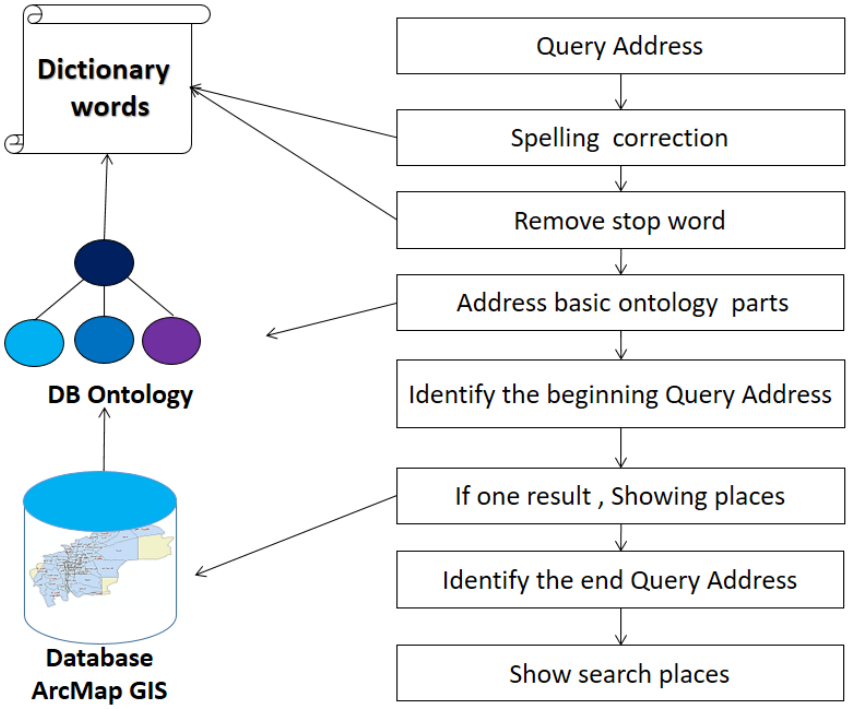


Figure 3: The architecture of the proposed approach

As shown in Figure 3, an ontology is used to search for major classes as (street, section, region,….) or other terms of the classes such as (the section is equal to the center) and if one of the words is found to be sure that the following words for this word follow this class and then record in its place in the array parts of the address. After reviewing the words with the classes of ontology in the address is to examine the words of the search address text from the beginning and identify the number of the house and the name of the street or area square . If they are found, the map is searched for and displayed, if the number of search results is more than one, the address text is searched from the end to search for the region or section, then search the map for the location.

4. An Illustrative Case Study

A random sample of how users wrote their addresses used to experiment with the proposed approach.

4.1 Schema and Data Sample

The research sample analyzed and found that the largest percentage begins writing the number of the house and then the name of the street or area square and then write the region or the name of the section and the lowest percentage after the name of the governorate and country. Figure 4 shows the result of the analysis of the sample of the study case.

Start of address		
Building No - Street	0.85	427
Building No - Area square	0.07	33
Other	0.08	40
Order full address		
Building No-Street-Region	0.53	266
Building No-Street-Section	0.27	134
Building No-Street-Area square	0.05	27
Building No-Area square-Region	0.04	19
Building No-Area square-Section	0.03	14
Other	0.08	40
Total sample		500

Figure 4: The ratios of retrieved addresses according to the search sample

4.2 Application Structure

The application consists of a set of tools have been added to the program ArcMap. Including a tool used to convert the data stored in the layers of the program to ontology and calculates the weights of words to use when correcting spelling errors for words search for addresses. Other tool to record other names of the individual stored from the layers of the program, Tool used to search for the address of using the proposed approach and the tool used to test the sample registered and determine the percentage of results found.

4.3 An Implementation respecting the Proposed Approach

- Using the frequency tool in ArcMap, extract the names of the streets in the building location layer and record them in a new table without duplicating the street names.
- In the same way, the names of the buildings and the names of the regions, divisions and governorates extracted and recorded in new tables without repetition.
- Record the values extracted as an ontology individuals, record the category of each individual and record the properties of each individual such as the governorate area and population.
- Record the other names of the individual from a file containing the other expected names of the individual contained within the ontology.

- Cut each individual into words, calculate the weight of each word, and recorded in the dictionary of words.
- Search for locations using the search tool where corrected spelling errors and remove stop words in the search and then use ontology in the search for class of parts of the address. If not found ontology, results are used search from the first address, search for the location, and if there are many locations searched from the end of the address, and then search the location layer for the parts found from the address structure.

4.4 Testing the proposed approach

The test transforms the content of the program layers into an ontology and then calculates the weights of the words as shown in Figure 5.

Figure 5: Convert to Ontology Tool and make Dictionary words

Figure 6: Test search 1

To start searching for a given address use the search form presented in Figure 6. Enter an address to search from. In consequence, remove the stop words in the address such as "شقة 8 الدور الرابع " , then has been addressing spelling correction such as "امبابة" equal "إمبابة" and "جيزة" equal "الجيزة" using ontology were identified street "31" and house number "15" and the name of use search last address was determined governorate "الجيزة" and section "إمبابة " and the region "مدينة التحرير" .

Figure 7: Test search 2

The identification of the building, "برج الجوهرة" and the street, "ميدان بن خلدون" and the region "مدينة الأعلام" by using ontology was finding the result, although the user writer section name "العجوزة" and the location administratively follows "المهندسين" section, but the user made a mistake because of the proximity places from each other.

Result sample

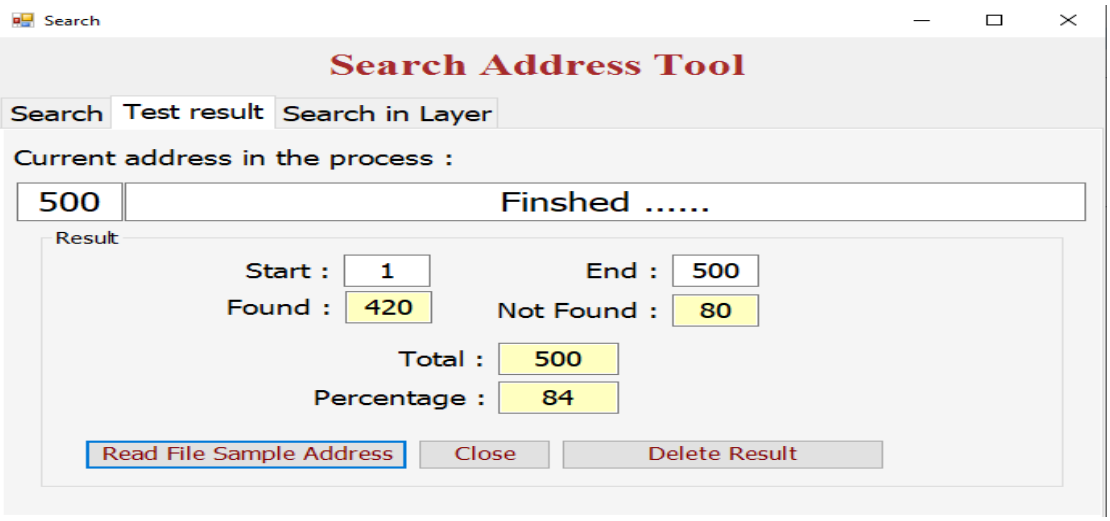


Figure 8: the total result of searching 500 misspelled addresses words

The research sample tested on the Excel file, the number of 500 addresses, and 420 of them identified by a resolution of percentage 84 %. The search results and the identified parts of the address parts recorded in a table showing the address text and the number of results that appeared in the search and the parts that have been identified and the date the address was last tested in the search.

A...	FullADDRESS	C...	NUM	STREET	REG	SEC	DIS	G...	C...	AREA	BULDNAME	A...	Last_Up
49	ش الوحدة المنيرة إمبابة 63	1	63	الوحدة	المنيرة	إمبابة							2019-11-21
50	ش دمشق المهندسين 18	1	18	دمشق		المهندسين							2019-11-21
51	آ ش عبدالمنعم رياض العجوزه 74	1	74	عبدالمنعم رياض								آ	2019-11-21
52	تاون هاوس فيلا 245 حى الاشجار 6 اكتو...	1	245	تاون هاوس	حى الأشجار								2019-11-21
53	ش عبدالمنعم رياض العجوزه 84	1	84	عبدالمنعم رياض		العجوزة							2019-11-21
54	ش خليفة وهبة البصراوى إمبابة 20	1	20	خليفة وهبة									2019-11-21
55	ش خان يونس من ش شهاب المهند 24...	0	24	شهاب		المهندسين							2019-11-21
56	ش المسجد الأقصى آرض اللواء المهند 9...	1	9	المسجد الأقصى		المهندسين							2019-11-21
57	ش طنطا العجوزة جيزة ج م ع 59	1	59	طنطا		العجوزة		الجيزة	مصر				2019-11-21
58	ش على سليمان الزاوية الحمراء الوايلى 26	1	26	على سليمان		الزاوية الحمراء							2019-11-21
59	ب ش امين الرافعى الدقى 31	1	31	امين الرافعى								ب	2019-11-21
60	ش رامز المهندسين 5	1	5	رامز		المهندسين							2019-11-21
61	ش السودان مدينة الطليه 405	1	405	السودان						مدينة الطلبة			2019-11-21
62	ش الغوث العجوزه 10	1	10	الغوث									2019-11-21
63	ش عدن المهندسين 42	1	42	عدن		المهندسين							2019-11-21
64	ش الحجاز المهندسين 7	1	7	الحجاز		المهندسين							2019-11-21
65	برج الجوهرة ميدان بن خلدون مدينة الأع...	1		ميدان بن خلدون	مدينة الأعلام						برج الجوهرة		2019-11-21
66	شارع السلام شقه 3 المرج 20	6	20	السلام		المرج							2019-11-21
67	مجاورة 9 عمارة 43 6 اكتوبر الحى الساد...	1	43	مجاورة 9									2019-11-21
68	ح الحنة من ش العامل العمرانية الغرب 7...	2	7	العامل	العمرانية ال...			الجيزة				ح	2019-11-21
69	آ ش سوريا المهندسين 25	1	25	سوريا								آ	2019-11-21
70	ش عبدالمنعم رياض العجوزة جيزة 84 ...	1	84	عبدالمنعم رياض		العجوزة		الجيزة	مصر				2019-11-21
71	ش المساحة الدقى 3	1	3	المساحة		الدقى							2019-11-21

Figure 9: Parts of test result table

5. Conclusion

This work tackles the problem of searching a GIS by unstructured text. It starts solving this problem analyzing a given address into Governorate, City, Street and house number as possible. Consequently, it maps such unstructured address into a formal structure that can easily be use to search and retrieve the desired address on the underlined GIS. It reaches around 77% accurate one matching address, around 81.5 % with up to two matching result(one of them is correct) and around 84 % with up to three matching addresses (one of them is correct), see Table 1.

Table 1: Detailed result sample

Matching addresses	Cases count	percentage
0	63	12.5 %
1	386	77.2 %
2	22	4.5 %
3	12	2.5 %
4	3	0.5 %
5	1	0.2 %
6	2	0.4 %
7	1	0.2 %
9	2	0.4 %
More than 9	8	1.6 %
total	500	100 %

The result indicates that, the count of cases of un-matching addresses is 63 cases. On the other hand the number of exact on matching address is 386 cases. In addition, the result shows that the number of cases with two candidate addresses (one of them is correct) is 22 cases. Finally, the count of cases with more than 9 candidate address (just one of them is correct) is 8 cases.

References

Derbal ,K. , Gloria ,B. , Gabriella,P. and Zaia,A. (2016),"Spatial Querying Supported by Domain and User Ontologies: An Approach for Web GIS Applications" , Proceedings of the 11th International Conference FQAS , 1 , 353-365.

Gao, S. and Goodchild, M.F. (2013). "Asking Spatial Questions to Identify GIS Functionality", Fourth International Conference on Computing for Geospatial Research and Application, 1, 106-110.

Haruna,K. and Ismail,M.A. (2016). "An Ontological Framework for Research Paper Recommendation", International Journal of Soft Computing, 11(2), 96-99.

Partyka,J. ,Alipanah,N., Khan,L., Thuraisingham,B. and Shekhar,S. (2008) , "Content-based ontology matching for GIS datasets" , 16th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems , 51 , 1-4.

Pillai, M. and Karabatis, G. (2018). "Enhancing Spatial Query Results Using Semantics and Multiplex Networks", Https://Mdsoar.org/Handle/11603/7900 , heteronam.org/2018/papers/HeteroNAM_2018_paper_2.pdf.

Srividhya, V. and Anitha R. (2010). "Evaluating Preprocessing Techniques in Text Categorization", International Journal of Computer Science and Application Issue 2010, 1, 49 - 51.

Zhu, J., Wright, G., Wang, J. and Wang, X. (2018). "A Critical Review of the Integration of Geographic Information System and Building Information Modelling at the Data Level", International Journal of Geo-Information, 7(2), 1-16.

A Survey on Multimedia over 5G Networks for D2D Communication Based Wi-Fi Direct Technology

Mohamed E. A. Ebrahim[1], Ammar Mohmmed[2]

Abstract

Smart-Device to *Smart-Device* (D2D) communications over Wi-Fi Direct (WFD) technologies in 5G network is still faces many dilemmas especially if mention some services available to the users such as social sharing of video streaming, audio and caching solutions, with massive number of mobile users, considering rate-distortion (RD) characteristics for QoS with finding the best way to calculate distance between multi-hop candidate devices and whether the device is static or moving with regard to characteristics of flow traffic speed such as High-Speed railways and subway with the mobility feature of communications according to the future demands in 5G. In this article, we present a survey of current methodologies and techniques related to neighboring discovery in the network area, interference management, and low energy consumption communication and its impact on multimedia delivery.

Keywords: D2D Communication, Proximity Services (ProSe), Wi-Fi Direct (WFD), Rate-Distortion (RD), 5G, Internet of Smart-Things (IoST), Smart-Device (SD).

1. Introduction

Electronic Smart-Devices (SD) are developing rapidly and significantly, and thus the volume of user requests will increase and consequently data traffic will grow in the future, monthly data traffic global will be 77 XB by 2022, and annual traffic will reach about zettabyte [Cisco, 2019][David Astely et al.., 2013]. Users concerns emphasize the importance of supporting various multimedia different with minimizing latency time [ACM, 2018], such as text, audio, graphics, images, animation, video, and interactive content, etc. [Special Issue, 2018][Ze-Nian Li et al.., 2004]. Therefore, based on the goals and requirements of 5G are smarter, more efficient and much faster [IMT, 2015]. Now and with D2D communications technology, we have become faced many challenges for multimedia delivery. There are many topics that researchers have tried to provide ideas and logical solutions to reach the requirements of the fifth generation of communications to support D2D communications especially Wi-Fi Direct because of a potential candidate for outband D2D communication [Rafay et al.., 2018] [Alexander et al.., 2014].

[1]Computer Sciences Department, Faculty of Graduate Studies for Statistical Research, Cairo University.

[2]Computer Sciences Department, Faculty of Graduate Studies for Statistical Research, Cairo University.

D2D is not user equipment's only but expanded to include for an example, personal computers, smart mobile phones, laptop, netbook, multimedia players, MP3 players and Smart-Machine to Smart-Machine M2M and also a Smart-Vehicle to Smart-Vehicle V2V, with 5G IoT [IY Shin et al.., 2018][L. Militano et al.., 2015] . In addition, Smart-Machine to Everything M2X and Smart-Vehicle to Everything V2X. The main contributions of this paper are to provide a Summary of Cases related to Smart-Device to Smart-Device (D2D) communications in 5G and discussion of techniques in previous works regarding Network Topology with the presentation of Analytical Tools used. This paper is organized as follows. Section 2 provides an overview of D2D. In Section 3, types of wireless communication. Licensed cellular spectrum and unlicensed cellular spectrum for D2D communication were briefly discussed in Section 4. In Section 5, categorize the challenges of D2D communication and view some of the previous work and how you faced those challenges, In Section 6, present research issues related to the future of D2D then display the advantages and disadvantages of ProSe technology in section 7.

2. Smart-Device to Smart-device

Smart-Device to Smart-Device Communications, also known as proximity services (ProSe) Communication, is part of the 3GPP (LTP) architecture [Alexander et al.., 2015] this means that two Smart-Devices (SDs) communicate directly without a base station BS and the underlying network, and include Long-Term Evolution LTE-Assisted D2D communications and direct communications as V2V. The ProSe communication is also expected to become an essential part of Internet objects and communications from Smart-Vehicle to Smart-Vehicle or Smart-Vehicle to Everything (V2X), where ProSe will be used to pass any information from the vehicle to any entity that may affect the vehicle and vice versa [Yang Yang et al.., 2018]. D2D cellular-assisted communication is a technique that allows terminals to multiplex the resources of direct-cell communication under cellular system control. D2D communications can solve the lack of spectrum resources in the wireless communications system to some extent while improving the spectrum efficiency of the cellular communications system CCS and decrease the transmission capacity of the terminal in [Sergey et al.., 2015]. In addition, reducing consumption of the mobile terminal battery power in [Bouchaib et al.., 2017], reducing the load of cellular network, improving network infrastructure, and increasing the transmission rate in [Ke Zhang et al.., 2018] and supporting P2P data services [Alexander et al.., 2014] [Alexander et al.., 2015].

3. Radio communication types for the Smart-Device to Smart-Device

There are many technologies that support a direct connection between Smart-Device to Smart-Device D2D. But we will review which are essentially for D2D communication and particular Wi-Fi Direct, Bluetooth, Zigbee, NFC, RFID [B. Zhou et al.., 2013] and briefly review the key features of available D2D technologies [L. Militano et al.., 2015].

3.1 Wi-Fi Direct

Recently some advances have been made to new Wi-Fi Direct WFD is the protocol provides new features that enable Wi-Fi to make discovery and connectivity faster and more efficient [Alexander et al.., 2014] [L. Militano et al.., 2015] [Shahid et al.., 2014]. In addition, the Smart-Device (SD) also allows you to host multiple APs access points to others during the connection, which means better flexibility for Smart-Devices (SDs) in setting up D2D communication. Thus, Wi-Fi Direct can create or accept as many connections of independent as possible, it is very convenient as a MAC layer for Smart-Device to Smart-Device Communication [Alexander et al.., 2014] [NGUYEN-SON et al.., 2018].

3.2 Bluetooth

Bluetooth, along with the Wi-Fi network, is the most popular D2D technology on the unlicensed band. Bluetooth provides Wireless connectivity in private domain networks. in order to enable short-range communication, one master of communication that serves up to seven users to form Bluetooth architecture. Bluetooth Low Energy was recently standardized for the Internet of SmartThings IoST and opens doors to new application scenarios by exploiting smartphone connectivity [L. Militano et al.., 2015] [Alexander et al.., 2014].

3.3 Zigbee

Zigbee is a set of protocols dedicated to wireless sensor networks limited resources. It is the MAC layers that determine the balanced trade-off between the data rate, the communication range, and energy efficiency. [L. Militano et al.., 2015] Several enhancements have been proposed, including IEEE 802.15.4E standard and IEEE 802.15.4G. The former MAC layer has been redesigned to support industry-specific applications specifically, by introducing time sync and switch channels. The latter handles ultra-wide sensor networks, such as smart utility networks [Alexander et al.., 2014].

3.4 NFC

Near-field communication NFC plays a prominent role in its widespread adoption, being locally incorporated into modern smart mobiles. In addition to interacting with tags, it expects Peer-to-Peer P2P mode through which Smart-Devices (SDs) can exchange any type of data directly. This technology promotes the widespread Internet of SmartThings IoST system [L. Militano et al.., 2015] [Sara et al.., 2014] .

3.5 RFID

Radio-frequency identification RFID refers to a set of radio techniques whose main purpose is to provide quick recognition of objects through interaction between transceivers, also known as tags, and readers. Different classifications of RFID systems can be provided according to the operating frequency, radio interface, communication group, tag independence, and various standards have been ratified.

For short-range communications [XiaolinJia et al.., 2012] [Vedat et al.., 2013]. In table 1 show a comparison of types of D2D communication technologies.

TABLE 1: Comparison of types D2D technologies [wi-fi.org] [Yanjun Hou et al.., 2017]

Parameter	Wi-Fi Direct (WFD)	Bluetooth	Zigbee	NFC	RFID (UHF)
Range	up to 200 m	10–100 m	10–100 m	4–10 cm	up to 100 m
Data Rate	up to 250 Mbps	0.8–2.1 Mbps	0.02–0.2 Mbps	0.02–0.4 Mbps	up to 640 kbps
Power consumption	High	High	Medium	Low	High
Spectrum	2.4 GHz	2.4 GHz	2.4 GHz	13.56 MHz	433 MHz
Security	High	Low	Low	High	High
Network topology	One to one one-to-many	Piconets, scatternets	Star, tree, mesh	One to one	One to one
Devices per network	up to 8 devices	8	2–65, 536	2	1

4. 3GPP LTE Overview for Smart-Device to Smart-Device Communications

At first, D2D was linked to LTE-A, which is part of 3GPP Rel.12, [Rafay et al.., 2018] [3GPP Release, 2017] [Andreas et al.., 2017] and expected to be D2D cellular communication technology is fully unified for proximity services ProSe in the next 3GPP Rel.13 and 14 and highlighted evolution of D2D concept of Smart-Vehicle to Everything V2X communication [L. Militano et al.., 2015] [Arash Asadi et al.., 2014]. Under the IEEE 802.11 protocol, the D2D concept was introduced on WFD. This allows WFD to play an active role from an access point, which is one of the main advantages of using Smart-Device to Smart-Device networks [Rafay et al.., 2018]. Recently, Wi-Fi Direct technology is available on the mobile phone, which will help deploy WFD based Smart-Device to Smart-Device networks [Alexander et al.., 2014]. ProSe communication is an appropriate technique for multimedia sharing services relies on proximity. Therefore, it is suitable for 5G networks in the future. Instead of cellular networks, direct connections between Smart-Devices (SDs) are make use of nearness. ProSe communications are categorized in Inband connections and Outband D2D [Rafay et al.., 2018] [Arash Asadi et al.., 2014].

4.1 Smart-Device to Smart-Device communication in licensed cellular spectrum

The licensed Cellular spectrum (Inband D2D Communications) provides spectral efficiency because of the spectrum sharing licensed between D2D Communications and cellular network users [Rafay et al.., 2018] [Nisha Panwar et al.., 2015]. The quality management mechanism is controlled by LTE eNBs evolved Node Base stations BSs to manage radio resource and mobility in the cell to optimize all the user equipments UE's communication in a radio network structure, which helps to address problems such as interference [Magri et al.., 2016]. Inband communications can be divided into underlay and overlay [B. Zhou et al.., 2013]. In underlay D2D mobile

phone users and D2D communications sharing the same spectrum resources, while in D2D communications overlay, dedicated spectrum resources are allocated to D2D users. D2D-based communications represent related challenges *interference management* and *resource allocation* between Smart-Device to Smart-Device D2D users and Long-Term Evolution LTE users. Overlay D2D communication overcomes on interference management and resource distribution problems but does not guarantee the efficient use of resources [L. Militano et al.., 2015] [B. Zhou et al.., 2013] [Vasilios A. et al.., 2019].

4.2 Smart-Device to Smart-Device communication in unlicensed cellular spectrum

The unlicensed Cellular spectrum (Outband D2D Communications) e.g. Industrial, scientific and medical domains (ISMs) are internationally radio bands for radio use. Wi-Fi Direct WFD is the perfect choice for Outband D2D to now. It is necessary to have an additional interface apply Wi-Fi Direct, Zigbee or Bluetooth to use an unlicensed spectrum[Rafay et al.., 2018] [B. Zhou et al.., 2013]. The main goal of exploiting Outband D2D communication is the ability to remove interference between D2D communications. D2D Outband communications face many challenges in coordination between different bands [Nisha Panwar et al.., 2015] [B. Zhou et al.., 2013].

5. Classification challenges of Smart-Device to Smart-Device Communications

D2D communications challenges in the cellular network can be categorized into both (*interference management* and *resource allocation*) based on Inband D2D and *(an extra interface* and *cluster communication*) based on Outband D2D. After what we reviewed D2D communications types and the challenges, in this section, we will mention the work of the researchers and their contribution in order to provide ideas and solutions to the challenges of Smart-Device to Smart-Device Communication with a focus on social sharing of multimedia streaming. The several authors like [Abbasi et al.., 2018][Mingyue Ji et al.., 2015][Ala'a et al.., 2018][Guangsheng et al.., 2017] [Taeho Kong et al.., 2016][Kaichuan et al.., 2018][Xin Song et al.., 2019] proposed the scheduling algorithms and scheduling scheme for sharing the multimedia using Smart-Device to Smart-Device D2D, the scenarios take considering interference by ranked the neighboring UEs. In [Abbasi et al.., 2018][Mingyue Ji et al.., 2015][J. Kim et al.., 2016][Ala'a et al.., 2018][Guangsheng et al.., 2017] [Charles T. et al.., 2018] authors provided optimal rate for caching files at random. Cached and segmented Video download techniques in [Guangsheng et al.., 2017] [Yunpeng et al.., 2016]. In [Taeho Kong et al.., 2016][Kaichuan et al.., 2018][Charles T. et al.., 2018] [Xingyan et al.., 2017][Arash et al.., 2017] [Arash et al.., 2016] developed framework probability to measure the video playback quality and intelligent emotion in 5G. The proposed system in [Zufan et al.., 2018] [Hafiz A. et al.., 2016][Changchun et al.., 2017] [Vasileios A. et al.., 2017] [Gagandeep et al.., 2019] [Caie Xu et al.., 2019] the video is split to data packet and users and using one cluster for same packet users. In [Kaichuan et al.., 2018][Charles T. et al.., 2018] [Caie Xu et

al.., 2019] proposed a D2D local caching to provide better video download services in a socially conscious environment. Presented in [Xingyan et al.., 2017] successful connection without human interaction based on Smart-Device to Smart-Device D2D Communication. In [Yunpeng et al.., 2016][Changchun et al.., 2017] [Jeevan et al.., 2019] [Jian Yang et al.., 2019] [Hussein et al.., 2019] [HUSSEIN et al.., 2012] the authors focus on video sharing and an energy in 5G networks for D2D. In [IY Shin et al.., 2018][Xin Song et al.., 2019] the authors proposed a resource-sharing scheme for multiple D2D groups and cellular users. For a better comparison of the use of caching techniques [Ke Zhang et al.., 2018][Yasser et al.., 2017] [Pablo et al.., 2017] [Jian Yang et al.., 2019] . In paper [Arash et al.., 2017] [Arash et al.., 2016] [Vasileios A. et al.., 2017] [Muhammad et al.., 2019] [Zeeshan et al.., 2019] [Alfonso et al.., 2019] [Jian Yang et al.., 2019] [Hussein et al.., 2019] presents framework based cross-layer for resources allocation of network for the transmission of (High-Efficiency Video Coding) encoded video streams in 5G. In [Farhan et al.., 2018] proposed (Ultra-High-Definition) 5G-UHD framework achieving adaptive for video streaming. The authors in [Yuhong Li et al.., 2017] analyzed network assisted D2D clustering in 5G cellular networks using Wi-Fi Direct and LTE-A. In [MEENAKSHI Li et al.., 2018] [Gagandeep et al.., 2019] proposed a cluster-based D2D communication in vehicular networks. The proposed methodology which uses distance as a parameter for radio resource allocation in Smart-Device to Smart-Device, in [Guangsheng et al.., 2017] [Ke Zhang et al.., 2018][Yasser et al.., 2017] [Essam et al.., 2018] [Rakshith K et al.., 2017] . Some authors using Classification algorithms based on features extracted from previous data like IoT module classification in [Jianguo Sun et al.., 2018] [Zhenggao et al.., 2019] and using Scheduling algorithms, Stochastic algorithm, Machine Learning algorithms. Finally, the summary of the literature proposed in table 2 and discussion in table 3.

TABLE 2: Summary of the literature proposed

Ref.	Proposal	Analytical tools	Platform	Use-case	Technique	Network Topology
[Abbasi et al.., 2018][Zeeshan et al.., 2019]	- Proposed a method to use Wi-Fi links efficiently. - watermarking scheme proposed for enhanced high-efficiency video coding standard.	-Prefetching algorithm -Extraction algorithm	LTE 5G 4G	-Video streaming -Video caching - video coding	- Resource allocation	-Wi-Fi Direct -Wi-Fi Hotspot -D2D
[Abbasi et al.., 2018][Mingyue Ji et al.., 2015][Ala'a et al.., 2018][Guangsheng et al.., 2017] [Taeho Kong et al.., 2016][Kaichuan et al.., 2018] [Xin Song et al.., 2019]	-Proposed the scheduling algorithms for sharing multimedia content effectively over Smart-Device to Smart-Device communication. - Proposed scheduling algorithms for quality video streaming and that can be used for Smart-Device to Smart-Device video delivery applications. -Proposed scheduling algorithm for seamless transmission of data packets over the network.	-Distribution Scheduling algorithms -Stochastic algorithm -PSNR - Fuzzy Round Robin Scheduling algorithm -keyed-hash message Authentication code algorithm - load balancing algorithm - streaming algorithm	5G LTE	-Video streaming sharing - data packets - multimedia transmission	- Resource allocation -Interference Management -M2M cluster	-Cellular -D2D -M2M -Wi-Max - IEEE - Cloud -Ad-hoc -P2P

[J. Kim et al., 2016][Ramona et al., 2015] [Yunpeng et al., 2016][Vasileios A. et al., 2017] [Martin et al., 2019] [Kaichuan et al., 2018]	-Proposed scheme as Random Popularity-based caching with colored indexing Coding. - Proposed edge computing framework for video processing groups based on Smart-Device to Smart-Device cluster.	-A greedy algorithm -A greedy coloring -Beyond scaling laws - brute force algorithm -formation algorithm -matching algorithm -heuristic algorithm	5G LTE	- Caching files -Video streaming	-D2D cluster -Interference Management	-D2D -Wi-Fi Direct -Cellular - Cloud
[Guangsheng et al., 2017] [Ke Zhang et al., 2018][Yasser et al., 2017]	-Proposed improving the throughput of video transport in LTE networks. -Caching techniques.	-Distance algorithms -shortest distance algorithm	5G LTE-A	-Cached and segmented video download -Video streaming -voice call	-Interference Management -Resource allocation - D2D cluster	-Cellular -Bluetooth -Wi-Fi Direct
[Taeho Kong et al., 2016][Ruyan et al., 2017] [Srinivas et al., 2016]	- Proposed development framework based on probability to measure the video quality. - Proposed predicting the QoE of video streaming service based on the non-linear regression function.	-Probability Density Function -PSNR -Scalable video coding -Block Error Rate - nonlinear regression - particle swarm optimization algorithm	LTE	-video playback quality - Video streaming	-Interference Management -Resource allocation	-Cellular -D2D
[Zufan et al., 2018] [Hafiz A. et al., 2016] [Shuxia et al., 2019]	- Proposed providing functional features not only the media but, also information usage of a user Smart-Device (SD). - Proposed split video into multiple data packets and users. -Proposed an image enhancement algorithm based on improved non-saturating game.	-signal-to-noise-ratio (SNR) -matching algorithm -Pareto optimality theory - Deep learning - peak signal-to-noise ratio - Adam optimization algorithm - Euclidean distance	LTE	-Video file - images	D2D cluster	-Cellular -D2D
[Kaichuan et al., 2018][Charles T. et al., 2018]	- Proposed D2D caching according to Smart-Devices (SDs) subscribers different.	-water-filling algorithm	5G	Cashed video	-D2D cluster -Resource allocation	-Cellular -D2D
[Xingyan et al., 2017][MEENAKSHI Li et al., 2018] [Arash et al., 2017] [Jeevan et al., 2019] [Gagandeep et al., 2019]	-Proposed Smart-Device to Smart-Device communication without human interaction. -optimization data delivery framework During communication over fast-moving vehicles. - Proposed Enhancement of multimedia streaming services for rider traveling in vehicles.	-Network simulator programs -fault-tolerant algorithm -bio-inspired routing algorithm -Machine Learning algorithms	5G	-vehicle videos -pictures files -multimedia delivery	-D2D cluster -V2V cluster -Interference Management	-Cellular -D2D -Wi-Fi Direct -C2V
[Yunpeng et al., 2016]	- Proposed scheme for multimedia streaming over 5G-D2D and an energy.	-A greedy algorithm -Fast Forwarding Algorithm - brute force algorithm	5G	-Video streaming -Video caching	-D2D cluster -Resource allocation	-Cellular -D2D
[IY Shin et al., 2018][Xin Song et al., 2019]	Proposed resource-sharing scheme for multiple D2D groups and cellular.	-interference alignment algorithm -Shannon's equation	5G	Video sharing	-Interference management -Resource allocation -D2D cluster	-Cellular -D2D
[Farhan et al., 2018]	-Proposed 5G-UHD Ultra-High-Definition framework	-scheduling algorithm -full search algorithm	5G	Video streaming	-Resource allocation	-Cellular -D2D

[Arash et al.., 2017]	achieving adaptive. -Proposed a cross-layer based framework for the transmission of High-Efficiency Video Coding.	-max-slope algorithm				
[Yuhong Li et al.., 2017]	-Proposed architecture D2D communication based the forthcoming D2D features of 5G networks. -analyzed network assisted D2D clustering.	-Cluster formation algorithm -merge and split algorithm -game theory -beyond theory	5G LTE-A	Data	D2D cluster	-Cellular -D2D -Wi-Fi Direct
[MEENAKSHI Li et al.., 2018]	-Proposed routing method based 5G D2D to vehicular networks and V2V cluster.	-clustering algorithm	5G LTE	Multimedia	-Interference management -Resource allocation -D2D cluster	-Cellular -D2D -IEEE 802.11
[Essam et al.., 2018]	-Proposed a simple and effective algorithm for cluster formation and distance-based resource allocation in Smart-Device to Smart-Device.	-cluster formation algorithm - Distance algorithm	5G LTE-A	Data	-Resource allocation	-Cellular -D2D
[Changchun et al.., 2017] [Vasileios A. et al.., 2017]	- Proposed cluster solution for multimedia delivery in an LTE D2D using Wi-Fi Direct to reduce an energy.	-Test-bed measurements -discovery algorithm	LTE	-Multimedia stream -Video file -Video caching	-Resource allocation - D2D cluster	-Cellular -D2D -Wi-Fi Direct
[Muhammad et al.., 2019] [Jianguo Sun et al.., 2018] [Fadi et al.., 2018] [Alberto et al.., 2019] [HUSSEIN et al.., 2012] [L. Militano et al.., 2015]	-Proposed faster route of multimedia sharing among the users of the smart city. -Presented an e-Learning system based on IoT for the synchronous and asynchronous communications. -multi-focus images techniques	-An Efficient algorithm -complexity aware algorithms -fusion algorithm - WEKA - classifiers, clustering algorithms - feature selection algorithms -SVM -consolidation algorithm	5G 4G	-Multimedia streaming -images -video Cameras - video conferences	-Resource allocation -Energy efficiency	-IoT - IoMT
[Jian Yang et al.., 2019] [Martin et al.., 2019] [Hussein et al.., 2019] [Zhenggao et al.., 2019] [Shadi et al.., 2019]	-an improved adaptive Huffman coding algorithm -Enhancing video quality and Quality of Experience - proposed a QoE-aware video quality takes considering scalable video coding.	-Huffman coding Algorithm -compression algorithm -stochastic optimization -SNR -Advanced Video Coding -SVM - Benford's law -Emotion vector	Cloud 5G	-Video Coding -JPEG compression -video clips -video compression	-Data compression methods	-Wireless -cloud -V2X

In table 3 will discuss some articles based on the requirements and objectives of the fifth generation 5G of communications.

TABLE 3: Results and Criticism.

Ref.	Results/Conclusions	Criticism
[Abbasi et al.., 2018][Kaichuan et al.., 2018][Charles T. et al.., 2018]	The offloading was optimized by fetching pre-fetched video data in the local hotspot cache and the D2D communication.	These caching solutions are cause a delay in case of live streaming but they efficient for the stored video.
[Mingyue Ji et al.., 2015][Ala'a et al.., 2018][Xin Song et al.., 2019]	- Classification User Equipment's for live streaming and on demand for the D2D scenario based on flow strength, delay and jitter. -Enhance video stream component based on random quality optimization approach.	applied the proposed approaches on the distance between the transmitting and receiving devices is 50m, but the fifth generation communication does cover the distance between 100m to 1000m and the number of devices expected is much greater than the mentioned when simulated Thus, these experiences lack the hopes of the fifth generation of communications. -The authors consider the implementation of caching D2D systems.
[J. Kim et al.., 2016][Ala'a et al.., 2018] [Guangsheng et al.., 2017] [Ke Zhang et al.., 2018][Yasser et al.., 2017] [Pablo et al.., 2017] [Kaichuan et al.., 2018][Charles T. et al.., 2018] [Yunpeng et al.., 2016]	-Enhance of the status of random demands has been discussed in advance through the applying of the optimal min-max demand plan. -improving caching techniques.	-These caching and Segmented Video solutions are effective for the stored video but do not present solutions for the case of live streaming. -The authors did not provide solutions for popular videos based on the real caching system.
[Taeho Kong et al.., 2016]	Improve the transmission times of the cellular downlink between the video layers in the cellular network.	The work did not take consider both size video and optimizing transmission durations.
[Zufan et al.., 2018] [Hafiz A. et al.., 2016]	The video is divided into multiple data packets and users who consent the same packet are classify into one cluster.	The scenario needs that be more realistic.
[Xin Song et al.., 2019]	succeed to reduce the interference from the D2D pairs to the cellular users based on the distance requirement	-scenarios are users usually static, some users or all users sometimes might be moving at pedestrian speed. But, the characteristics of high traffic volume density such as subway and high-speed railways, the high mobility feature of wireless communications will be prominent in 5G. - The average distance between the sender and receiver of a D2D pair is 50 m
[Xingyan et al.., 2017]	Improving D2D communication without human interaction in both 5G applications advanced within WFD.	-The work needs to analyze this type of ProSe communication with a greater number of Smart-Devices (SDs). -The Smart-Devices (SDs) were placed on a table at an average distance of 3 m, and movements were not taken into consideration. -black pictures and white pictures, and at low resolution.

6. Some research issues related to the future of 5G

There are some research issues related to the future of Smart-Device to Smart-Device D2D communications and with 5G can be listed in a set of questions for example, can 5G design support billions of connected Smart-Devices (SDs) and users worldwide, and can support variety of services and applications, will 5G ensures low latency to avoid the bottleneck of the radio interface in real time, does the 5G offer network features that are different from those of the traditional IoT scene. In Smart-Device to Smart-Device D2D scenarios, how to detection Smart-Devices (SDs) in area and how to calculate the distance between each Smart-Device (SD) taking into consideration some scenarios, the users be static or users may move at pedestrian speed or high traffic volume density such as the vehicle, subway and high speed railways, how

applications provide road safety systems and traffic updates for Smart-Vehicle to Smart-Vehicle V2V communications and what automated mechanism exploits social connections between Smart-Devices (SDs) to create D2D dynamic social networks and information gathered from social networks can help share content in general and Smart-Devices (SDs) of similar interest can collaborate with each other to share content and does the D2D capability allow for efficient spectrum utilization, enhance network life, and allow Smart-Devices (SDs) to operate independently in situations where immediate access to energy infrastructure is not possible as well as many problems such as interference management, resource allocation, and live multimedia streaming, now current solutions to the state of live streaming are not enough as they rely heavily on caching and Segmented Video. And many questions and ideas for discussion of the future of 5G.

7. Advantages and disadvantages of the different types of Smart-Device to Smart-Device Communications

A brief overview of the advantages and disadvantages of Inband D2D and Outband D2D communications. Overall the advantages of D2D Communication lead to link reliability between Smart-Device to Smart-Device D2D, higher data rate, instant communication, low End-to-End latency, easily share files from Peer-to-Peer P2P, improve voice and video streaming services, online games, improve throughput, low power consumption, IoT Enhancement [Rafay et al.., 2018] . Inband D2D, increases underlay D2D the spectral efficiency of cellular spectrum leading to high spectral efficiency, QoS management is easy because the cellular spectrum It can be completely controlled by the base station BS [Nisha Panwar et al.., 2015], can use the Inband D2D on any device. Outband D2D, non-interference between D2D and cellular subscribers, nothing is necessary to dedicate cellular resources to spectrum D2D [B. Zhou et al.., 2013]. The main disadvantage of Inband lies in the interference caused by Smart-Device to Smart-Device D2D users to cellular users. In underlay Inband D2D communications, D2D users use a common cellular frequency, it increases the probability of interference; D2D users, therefore, need efficient power control mechanisms. When a group of Smart-Devices (SDs) selects the cluster head and processes the requests of the bundled Smart-Devices (SDs) and therefore needs to improve Spectral efficiency. In the overlay mode, Smart-Device to Smart-Device D2D users use dedicated cellular frequency and therefore require a solution to reduce interference by allocating dedicated resources for Smart-Device to Smart-Device D2D communication. In D2D Outband communication exploits unlicensed spectrum, the use of unlicensed spectrum requires additional interfaces such as Wi-Fi direct, Zigbee, and Bluetooth because these technologies are usually different from cellular technologies available such as 3G and 4G [Rafay et al.., 2018] [Nisha Panwar et al.., 2015].

Conclusion

In this article, we have provided a wide survey about the literature available on Smart-Device to Smart-Device D2D communication within cellular network and Wi-Fi Direct WFD and provided solutions for video streaming, Cached and segmented Video Download, and make a comparison between the articles with the mentioned techniques used in each work and Analytical tools then presented critical analytical based on 5G requirements. The survey found that the number of articles that were related to Smart-Device to Smart-Device D2D with Wi-Fi Direct WFD is few compared cellular network, it is considered an open issue.

References

- Cisco Visual Networking Index (2019), Global Mobile Data Traffic Forecast Update, 2017–2022.
- ACM Multimedia Systems Conference (2018), Amsterdam.
- Special Issue on Interactive Multimedia in 5G Communications Multimedia Tools and Applications (2018), Springer Multimedia Tools and Applications an international journal.
- Rafay Iqbal Ansari, Chrysostomos Chrysostomou, Syed Ali Hassan, Mohsen Guizani, Shahid Mumtaz, Jonathan Rodriguez and Joel J.P.C. Rodrigues (2018), 5G D2D Networks: Techniques, Challenges, and Future Prospects, IEEE Systems Journal.
- Alexander Pyattaev, Olga Galinina, Kerstin Johnsson, Adam Surak, Roman Florea, Sergey Andreev and Yevgeni Koucheryavy (2014), Network-Assisted D2D Over WiFiDirect, Springer.
- IY Shin, SH Han, SD Kim (2018), Device for performing communication in wireless communication system and method thereof, US Patent App - Google Patents.
- L. Militano, G. Araniti, M. Condoluci, I. Farris and A. Iera (2015), Device-to-Device Communications for 5G Internet of Things, EAI: European Alliance for Innovation Endorsed Transactions on Internet of Things, Volume 1, Issue 1.
- Ze-Nian Li and Mark S. Drew (2004), Fundamentals of Multimedia, School of Computing Science Simon Fraser University.
- IMT-2020(5G) Promotion Group (2015), White Paper on 5G Wireless Technology Architecture. http://www.imt-2020.cn/en/documents/3.
- Alexander Pyattaev, Jiri Hosek, Kerstin Johnsson, Radko Krkos, Mikhail Gerasimenko, Pavel Masek, Aleksandr Ometov, Sergey Andreev, Jakub Sedy, Vit Novotny, and Yevgeni Koucheryavy (2015), 3GPP LTE-Assisted Wi-Fi-Direct:Trial Implementation of Live D2D Technology.
- Yang Yang, Jing Xu, Guang Shi, Cheng-Xiang Wang (2018), 5G Wireless Systems Simulation and Evaluation Techniques, Springer International Publishing.
- Daniel Camps-Mur, Nec Network Laboratories, Andres Garcia-Saavedra and Pablo Serrano, (2013), DEVICE-TO-DEVICE COMMUNICATIONS WITH WIFI DIRECT: OVERVIEW AND EXPERIMENTATION, IEEE Wireless Communications.
- Shahid Mumtaz, Jonathan Rodriguez (2014), Smart Device to Smart Device Communication, book, springer- Technology & Engineering.
- Nguyen-Son VO, Trung Q. Duong, Hoang Duong Tuan, and Ayse Kortun (2018), Optimal Video Streaming in Dense 5G Networks with D2D Communications.
- David Astely, Erik Dahlman , Gabor Fodor, Stefan Parkvall, Joachim Sachs (2013), LTE release 12 and beyond, IEEE Communications Magazine, Volume: 51 Issue: 7.
- 4G_Americas. (2015), 3GPP Release 12 Executive Summary.

- Andreas Hoglund, Xingqin Lin, OlofLiberg, Ali Behravan, Emre A. Yavuz, Martin Van Der Zee, Yutao (2017), Overview of 3GPP Release 14 Enhanced NB-IoT, IEEE Network Journal, Volume: 31 , Issue: 6.
- Arash Asadi, Qing Wang, and Vincenzo Mancuso (2014), A Survey on Device-to-Device Communication in Cellular Networks, IEEE Communications Surveys & Tutorials Volume: 16 , Issue: 4.
- Nisha Panwar, Shantanu Sharma, and Awadhesh Kumar Singh (2015), A Survey on 5G: The Next Generation of Mobile Communication, arxiv.org.
- Magri Hicham, NoreddineAbghour and Mohammed Ouzzif, (2016), DEVICE-TO-DEVICE (D2D) COMMUNICATION UNDER LTE-ADVANCED NETWORKS, International Journal of Wireless & Mobile Networks (IJWMN) Vol. 8, No. 1.
- B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, (2013), "Intracluster device-to-device relay algorithm with optimal resource utilization," IEEE Transactions on Vehicular Technology, vol. 62, no. 5.
- Vasilios A. Siris and Dimitrios Dimopoulos, (2019), Multi-Source Mobile Video Streaming with Proactive Caching and D2D Communication, IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks.
- Ubaid Abbasi and Halima Elbiaze, (2018), Multimedia Streaming using D2D in 5G Ultra Dense Networks, Conference: IEEE Annual Consumer Communications & Networking Conference (CCNC).
- Mingyue Ji, Antonia M. Tulino, Jaime Llorca, Giuseppe Caire, (2017), Order-Optimal Rate of Caching and Coded Multicasting with Random Demands, arXiv,10 Feb 2015, IEEE /ACM Transactions on Networking, Published Online Volume 63 Issue 6.
- J. Kim, G. Caire, and A.F. Molisch, (2016), "Quality-Aware Streaming and Scheduling for Device-to-Device Video Delivery," IEEE/ACM Transactions on Networking, Published Online Volume: 24 Issue: 4.
- Ala'a Al-Habashna1 and Gabriel Wainer, (2018), Improving Video Transmission in Cellular Networks with Cachedand Segmented Video Download Algorithms, Mobile Networks andApplications.
- Guangsheng Feng, Yue Li, Qian Zhao, HuiqiangWang, HongwuLv, Junyu Lin, (2017), Optimizing broadcast duration for layered video streamsin cellular networks, Peer-to-Peer Networking and Applications.
- Taeho Kong, Junghyun Bum, and Hyunseung Choo, (2016), Media, Screen, Input, and Context Sharing System for D2D Services in Smart TV 2.0, ICCSApp. 180–194.
- Zufan Zhang, Tian Zeng, Xiulan Yu, Shaohui Sun, (2018), Social-aware D2D Pairing for Cooperative Video Transmission Using Matching Theory, Mobile Networks and Applications.
- Hafiz A. Mustafa, Muhammad A. Imran, Muhammad Z. Shakir, Ali Imran, and Rahim Tafazolli, (2016), Separation Framework: An Enabler for Cooperativeand D2D Communication for Future 5G Networks, arXiv.
- Kaichuan Zhao, Yuezhi Zhou, Wenjuan Tang, Shuang Li, and Yaoxue Zhang, (2018), A Game Theoretic D2D Local Caching System under Heterogeneous Video Preferences and Social Reciprocity, Springer Nature Switzerland AG, pp. 417–431.
- Charles T. B. Garrocho, Maurı´cio J. da Silva, Ricardo A. R. Oliveira, (2018), D2D pervasive communication system with out-of-band control autonomous to 5G networks, Wireless Networks Springer Nature.
- Xingyan Chen, Mu Wang, ShijieJia, and Changqiao Xu, (2017), Energy-Aware Fast Interest Forwarding for Multimedia Streaming over ICN 5G-D2D, Springer International Publishing.
- Yunpeng Li, Zeeshan Kaleem, Kyung Hi Chang, (2016), Interference-Aware Resource-Sharing Scheme for Multiple D2D Group Communications Underlaying Cellular Networks, Wireless PersCommun.

- Xin Song, Xiuwei Han, Yue Ni, Li Dong and Lei Qin, (2019), Joint Uplink and Downlink Resource Allocation for D2D Communications System, Future Internet journal.
- Sergey Andreev, Dmitri Moltchanov, Olga Galinina, Alexander Pyattaev, Aleksandr Ometov, and Yevgeni Koucheryavy, (2015), Network-Assisted Device-to-Device Connectivity: Contemporary Vision and Open Challenges,Proceedings of European Wireless 2015; 21th European Wireless Conference.
- Bouchaib Assila, Abdellatif Kobbane, Mohammed El Koutbi, (2017), A Survey on Caching in 5G Mobile Network.
- Ke Zhang, Supeng Leng, Yejun He, Sabita Maharjan, and Yan Zhang, (2018), Cooperative Content Caching in 5G Networks with Mobile Edge Computing, IEEE Wireless Communications.
- Yasser Fadlallah, Antonia M. Tulino, Dario Barone, Giuseppe Vettigli, Jaime Llorca, and Jean-Marie Gorce, (2017), Coding for Caching in 5G Networks, IEEE Communications Magazine.
- Pablo Salva-Garcia, Jose M. Alcaraz-Calero, Ricardo Marco Alaez, Enrique Chirivella-Perez, James Nightingale, Qi Wang, (2017), 5G-UHD: Design, Prototyping and Empirical Evaluation of Adaptive Ultra-High-DeÞnition Video Streaming Based on Scalable H.265 in Virtualised 5G Networks, Computer Communications.
- Farhan Pervez, Abdul kareem Adinoyi, Halim Yanikomeroglu, (2018), Efficient Resource Allocation for Video Streaming for 5G Network-to-Vehicle Communications, 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC).
- Arash Asadi, Vincenzo Mancuso, (2017), Network-assisted Outband D2D-clustering in 5G Cellular Networks: Theory and Practice, IEEE Transactions on Mobile Computing, Volume: 16 , Issue: 8.
- Yuhong Li, Haoyue Xue Zhihui Gai Xirong Que Xiang Su Jukka Riekki, (2017), A cluster-based routing method for D2D communication oriented to vehicular networks, IEEE International Conference on Systems.
- Meenakshi Singh, Varun Gupta, Deepika Rani Sona, (2018), CLUSTER FORMATION USING LEACH PROTOCOL AND DISTANCE BASED RESOURCE ALLOCATION IN D2D COMMUNICATION, International Journal of Electrical, Electronics and Data Communication, Volume-6, Issue-6.
- Essam Abdelsalam Mahmoud Osman, (2018), DEVICE DISCOVERY METHODS IN D2D COMMUNICATIONS FOR 5G COMMUNICATIONS SYSTEMS.
- Rakshith K and Mahesh Rao, (2017) NEW TECHNOLOGY FOR MACHINE TO MACHINECOMMUNICATION IN SOFTNET TOWARDS 5G, International Journal of Wireless & Mobile Networks (IJWMN) Vol. 9, No. 2.
- Joongheon Kim, Giuseppe Caire, and Andreas F. Molisch, (2015), Quality-Aware Streaming and Scheduling forDevice-to-Device Video Delivery, IEEE/ACM TRANSACTIONS ON NETWORKING.
- Arash Asadi, Vincenzo Mancuso, Rohit Gupta, (2016), An SDR-based Experimental Study of Outband D2D Communications, IEEE International Conference on Computer .
- Ramona Trestian, Quoc-Tuan Vien, Huan X. Nguyen, Orhan Gemikonakli, (2015), ECO-M: Energy-Efficient Cluster-Oriented Multimedia Delivery in a LTE D2D Environment, IEEE International Conference on Communications (ICC).
- Changchun Long, Yang Cao,Tao Jiang,and Qian Zhang, (2017), Edge Computing Framework for Cooperative VideoProcessing in Multimedia IoT Systems, IEEE TRANSACTIONS ON MULTIMEDIA, VOL.2, NO. 6.
- Vasileios A. Memos, Kostas E. Psannis, Yutaka Ishibashi, ByungGyu Kim, and B.B Gupta, (2017), An Efficient Algorithm for Media-based Surveillance System (EAMSuS) in IoT Smart City Framework Future Generation Computer Systems.

- Muhammad Munwar Iqbal, Muhammad Farhan, SohailJabbar, Yasir Saleem, Shehzad Khalid, (2019), Multimedia based IoT-centric smart framework for eLearning paradigm, Multimedia Tools Applications.
- Jianguo Sun, WenshanWang, Kejia Zhang, Liguo Zhang, Yun Lin, Qilong Han, Qingan Da, Liang Kou, (2018), A multi-focus image fusion algorithm in 5GCommunications, Multimedia Tools Applications.
- Fadi Al-Turjman and SinemAlturjman, (2018), 5G/IoT-enabled UAVs for multimedia delivery in industry-oriented applications, Multimedia Tools Applications.
- Jeevan Kharel, Soo Young Shin, (2019), Multimedia service utilizing hierarchical fog computing for vehicular networks, Multimedia Tools Applications.
- Gagandeep Kaur1, Singara Singh Kasana1, M. K. Sharma, (2019), an efficient watermarking scheme for enhanced high efficiency video coding/h.265, Multimedia Tools Applications.
- Zeeshan Ahmed, Salima Hamma, Zafar Nasir, (2019), an optimal bandwidth allocation algorithm for improving QoS in WiMAX, Multimedia Tools Applications.
- Alfonso Panarello, Antonio Celesti, Maria Fazio, Antonio Puliafito, Massimo Villari, (2019), A big video data transcoding service for social media over federated clouds, Multimedia Tools Applications.
- Caie Xu, Yang Cui, Yunhui Zhang, Peng Gao, JiayiXu, (2019), Image enhancement algorithm based on generative adversarial network in combination of improved game adversarial loss mechanism, Multimedia Tools Applications.
- Shuxia Wang, (2019), Multimedia data compression storage of sensor network based on improved Huffman coding algorithm in cloud, Multimedia Tools Applications.
- Jian Yang, Zhen Xiang, Lisha Mou, Shumu Liu, (2019), Multimedia resource allocation strategy of wireless sensor networks using distributed heuristic algorithm in cloud computing environment, Multimedia Tools Applications.
- Martin Fleury, Dimitris Kanellopoulos, Nadia N. Qadri, (2019), Video streaming over MANETs: An over view of techniques, Multimedia Tools Applications.
- Hussein Al-Bandawi, Guang Deng, (2019), Classification of image distortion basedon the generalized Benford's law, Multimedia Tools Applications.
- Zhenggao Pan, Xianwei Li1, Lin Cui1, Zhiwei Zhang, (2019), Video clip recommendation model by sentimentanalysis of time-sync comments, Multimedia Tools Applications.
- Shadi AlZu'bi, Bilal Hawashin, Muhannad Mujahed, Yaser Jararweh, Brij B. Gupta, (2019), An efficient employment of internet of multimedia things in smart and future agriculture, Multimedia Tools Applications.
- Alberto Huertas Celdran, Manue Gil Perez, Felix J. Garcıa Clemente, FabrizioIppoliti, Gregorio Martınez Perez, (2019), Dynamic network slicing management of multimedia scenarios for future remote healthcare, Multimedia Tools Applications.
- Alessandro Carrega, Giancarlo Portomauro, Matteo Repetto, Giorgio Robino, (2019), Energy efficiency for edge multimedia elastic Applications, Multimedia Tools Applications.
- Hussein Ahmad Al-Ofeishat, Mohammad A.A. Al Rababah, (2012), Near Field Communication (NFC), IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.2.
- Sara Amendola, Rossella Lodato, Sabina Manzari, Cecilia Occhiuzzi, and Gaetano Marrocco, (2014), RFID Technology for IoT-Based Personal Healthcare in Smart Spaces, IEEE INTERNET OF THINGS JOURNAL, VOL. 1, NO. 2.
- XiaolinJia, Quanyuan Feng, Taihua Fan, Quanshui Lei, (2012), RFID Technology and Its Applications in Internet of Things (IOT), IEEE International Conference on Consumer Electronics, Communications and Networks (CECNet).

- Vedat Coskun, Busra Ozdenizci, Kerem Ok, (2013), A Survey on Near Field Communication (NFC)Technology, Wireless PersCommun.
- https://www.wi-fi.org/
- Yanjun Hou, Wen'an Zhou, Lijun Song, Mengyu Gao, (2017), A QoE Estimation Model for Video Streaming over 5G Millimeter Wave Network Advances on Broad-Band Wireless Computing, Communication and Applications.
- Ruyan Wang, Yan Yang, and Dapeng Wu, (2017), A QoE-Aware Video Quality Guarantee Mechanism in 5G Network, Springer International Publishing.
- Srinivas Bachu, Y. Olive Priyanka, V. Bhagya Raju, (2016), A Novel Approach for Detection of Motion Vector-Based Video Steganography by AOSO Motion Vector Value, Innovations in Computer Science and Engineering, Advances in Intelligent Systems and Computing.

Bluetooth Smart based Automated Attendance System

Aisha Mohamed Hussein[1], Asmaa Mustafa[2], Hanan Ibrahim Shawky[3], Hiba Ahmed Joudah[4], Nesrine A. Azim[5]

Abstract

Managing student attendance during lecture periods has become a difficult challenge. The ability to compute the attendance percentage becomes a major task as manual computation produces errors, and wastes a lot of time. For the previously mentioned reasons, a portable device is designed, where in every student can fed his/her attendance each lecture. The student verification is done using wireless RFID (Radio Frequency Identification) system where the information contain in the RFID tag is stated as the ID. The student attendance ID is sent through the serial interface and start matching with the information stored in one Atmega AVR32 microcontroller. The information "ID" displayed on LCD and transmitted to the second Atmega AVR32 that interfaced with Bluetooth terminal. Finally, information sent to mobile Android. In this paper we designed an application that takes the attendance electronically by smart Bluetooth. The application reduces human errors, saves time and provides secure attendance for further use in managerial decisions.

Keywords: Embedded system, Rapid Frequency Identification, Atmega AVR, Bluetooth.

[1,2,3,4] Department of Computer Science, Cairo University, Faculty of Graduate Studies for statistical Research, Cairo, Egypt.
[5] Department of Information Systems and Technology, FGSSR, Cairo, Egypt.
Nesrinealiazim79@hotmail.com

1- Introduction

Attendance is one of the work ethics valued by employers. Most of the educational institutions and government organizations in developing countries still use paper-based attendance method for maintaining the attendance records. There are various disadvantages to this approach such as data is not available for analysis because paper-based registers are not uploaded to a centralized system, time taken for data collection reduces the effective lecture time and fake attendance by students. There is a need to replace these traditional methods of attendance recording with smart wireless attendance system. In recent years, there has been rise for the Rapid Frequency Identification (RFID) systems, which are already active in many sectors, such as health-care, transportation, agriculture and more. As such, the Internet of Things (IOT) technology uses the RFID technology to access control of a system. IOT can monitor the actual in and out time of a person from any location at a given time. IOT managed using a microcontroller, like a sensor, which will interact with a internet based device (mobile). The system used in this work, is a wireless embedded system application developed for daily student attendance in departments within the university. Embedded stands for hardware controlled by software. Here, the software using a Microcontroller controls all the hardware components. RFID is a wireless technology, which uses electromagnetic waves for communication between RFID reader and RFID tag. Though better than paper-based systems, RFID based systems also have certain problems such as the system is complex and absent student's card can be swiped by other students. For that, communication with Bluetooth terminal data will send to android and can be saved in a database. Bluetooth smart systems can provide accurate and precise data about tagged items that will help in creation of database system not prone to errors. The system is not only improving the work efficiency, students' study and development, but also can save human and material resources.

There are two stages of working of these systems: 1) Interface of RFID and AVR microcontroller and 2) Interface with Bluetooth terminal and android.

2- Previous work

An automated attendance management system using RFID reader was implemented in mobile and electronic platform respectively. In Zainal et al, (2014), a portable fingerprint attendance system is designed using Arduino board based on ATmega1280 (see Zainal et al, 2014). Fingerprint scanner ZFM 20 is used having its own processor and memory. TFT touch screen provides user friendly interface. SD card is used for storage of student's records. RTC (Real time clock) provides the exact attendance date and time. Caeser Cipher cryptographic technique is used so that data cannot be manipulated by unauthorized user. In Arulogun et al, (2013), RFID based system identifies tags and is used to mark students' attendance (See Arulogun et al, 2013). A computer has been used as the medium to perform this task. The RFID reader detects the presence of a tag and the system processes this information on the computer according to the programmed instructions. The tractability, availability and receptiveness of the technology highly affect the ease with which RFID system can be integrated into current operations.

A different type of automatic attendance system was proposed that uses Fingerprint sensor/scanner along with other technologies. In Yadav et al, (2015), the system is designed using 8051 microcontroller and R305 optical fingerprint sensor and LabVIEW Microcontroller communicates with computer in which LabVIEW is installed (See Yadav et al, 2015). LabVIEW is used in the system for storing attendance records, maintaining it in a text file and displaying it to the user. Student ID is also displayed on the LCD screen after fingerprint matching.

A more secure hardware system was proposed by Wang and Jingli (2015), using Internet of things that includes ARM9 S3C2440 processor board and FPS200 solid state fingerprint sensor. Database is designed using SQLite database management tool. In addition to fingerprint biometric, Vein recognition is also used. It allows for real time attendance data and monitoring using a website but the complex software system and high cost were the main drawback.

A system comprises of transmitter section, receiver section element and attendance supervision terminal were implemented by Kamaraju and Kumar, (2015). Transmitter section consists of optical fingerprint sensor OP-100N, ADSP-BF532 & ZigBee transmitter. Digital signal processor consists of ZigBee receiver and microcontroller. Image enhancement is performed using MATLAB. MS-access and Visual basic are used for database implementation. The system was low

power, cost and wireless but requires three different supply voltages (3.3V, 5V, 12V).

Another system implemented by Ansari et al (2015), combines RFID and GSM technology with biometrics for attendance management. Students ID is tagged with their RFID tag that matched with the database and attendance is finalized after fingerprint is verified using fingerprint sensor. GSM Modem is used for sending SMS to parents regarding student's attendance. RFID transponders are installed in classrooms, laboratories and staffrooms through which location of the student and staff can be traced. A website is designed through which teacher, students and guardians can view the location of a student in the campus and also the attendance record of the student. vb.net is used for server application and asp.net for website. System using NFC (Near Field Communication) is implemented in Benyo B. et al, (2012). NFC based system has lower range than RFID based systems. The system is fully automated and secure but costly, the software design is difficult and the system should always keep ON.

This review incorporates the problems of attendance systems presently in use, working of a typical fingerprint-based attendance system, study of different systems, their advantages, disadvantages and comparison based upon important parameters.

3- Analysis and Design

3.1 - Proposed model

The AtmegaAVR32 microcontroller is pre-programmed with Embedded C language using software Atmel studio 6.2. For RFID based systems, after connecting all the components of attendance system, give the power supply to switch on the circuit. In case of scanning the tag, the tag must be close or in contact with RFID reader. The scanning time is approximately same as the time it takes to manually count the students in the class. The information contain in the RFID tag is stated as the ID and attendance of the student. When person placed card in front of RFID reader, it reads the information and start matching with the information stored in one AVR32 microcontroller. The information (ID) will be displayed on LCD after swipe the card. Next step, the information transmitted to the second AVR32 microcontroller and interface with a HC-05 Bluetooth module through UART serial communication protocol. Bluetooth HC05 is programmed and configured such that it works in connection with the Android application via Bluetooth. Here, the data will send from HC-05 Bluetooth to Android Mobile figure (1,2).

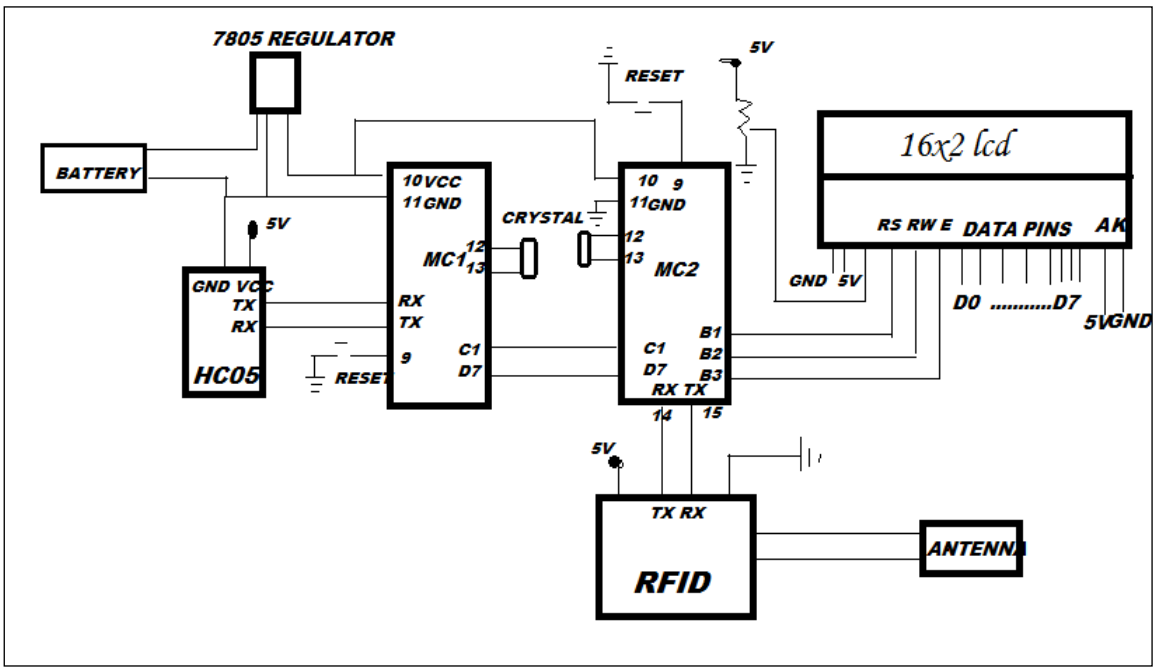


Figure (1): Block diagram of AVR32A wireless communication circuit

RFID has advantage that electronic tags can be embedded into the student ID card that require low power consumption, the tag can be read during motion as well no line of sight is required for wireless communication between the tag and the reader.

The application detects tags only within a particular range in order to avoid detection of tags outside the class.

Bluetooth chip operate in the range of (2.4GHz-2.4835 GHz ISM band) same as classic bluetooth that support high data transfer rate and low power consumption (See Lodhaa et al, 2015).

RFID tags are of 2 types: Passive tags contain 13-digit number inbuilt in it, whereas active tags are read/write tag. RFID reader contains a copper winding in it act as antenna. Due to interaction of tag with reader 12-byte (characters) ASCII string are sent to microcontroller via serial communication. Before this controller is loaded with a program. In microcontroller data of students are saved. In our case, data of two students are saved i.e. tag number and name. When we provide power supply to the circuit, the circuit switches on and RFID based attendance is displayed on LCD. When 12 characters are transferred to controller, the controller matches the characters with the saved characters. If the characters

are matched (figure 2) then controllers send ‘1’ and displayed student name. If not matched return and scan again.

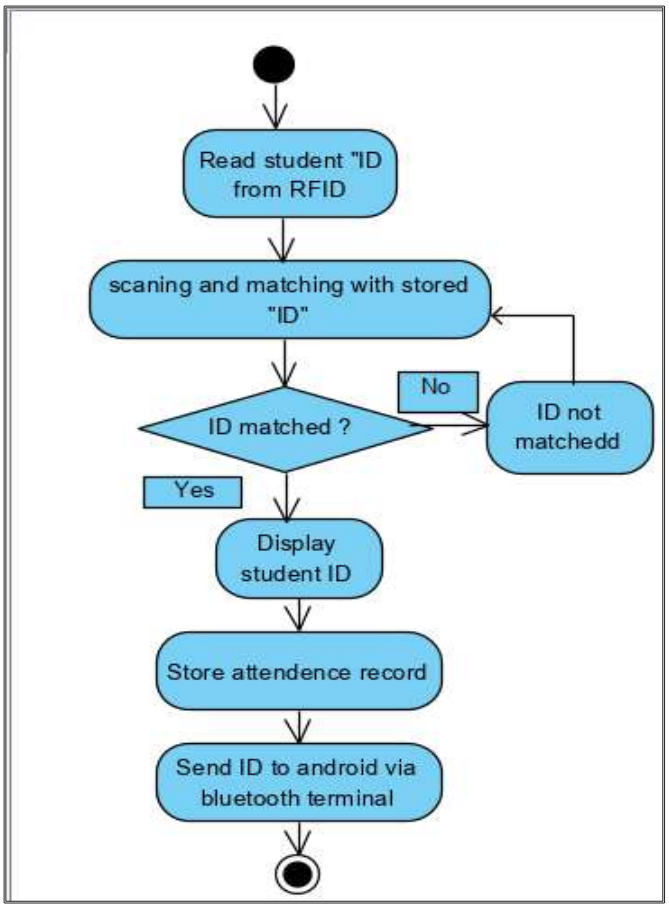


Figure (2): Activity diagram for attendance system

3.2 Interface of two ATmega AVR32 microcontroller

- The serial communication between two ATmega32 micronctrollers is UART (Universal Asynchronous Receiver Transmitter) (Figure.3). UDR or USART Data Register is meant for writing and receiving the data through the UART.
- The TXD pin (data receiving feature) of first controller must be enabled for transmitter and RXD pin of second controller must be enabled for receiver.

- Since the communication is serial we need to know whenever the data byte is received, so that we can stop the program until complete byte is received. This is done by enabling a data receive complete interrupt.
- DATA is transmitted and received to controller in 8bit mode. So two characters will be sent to the controller at a time.

- There are no parity bits, one stop bit in the data sent by the module.

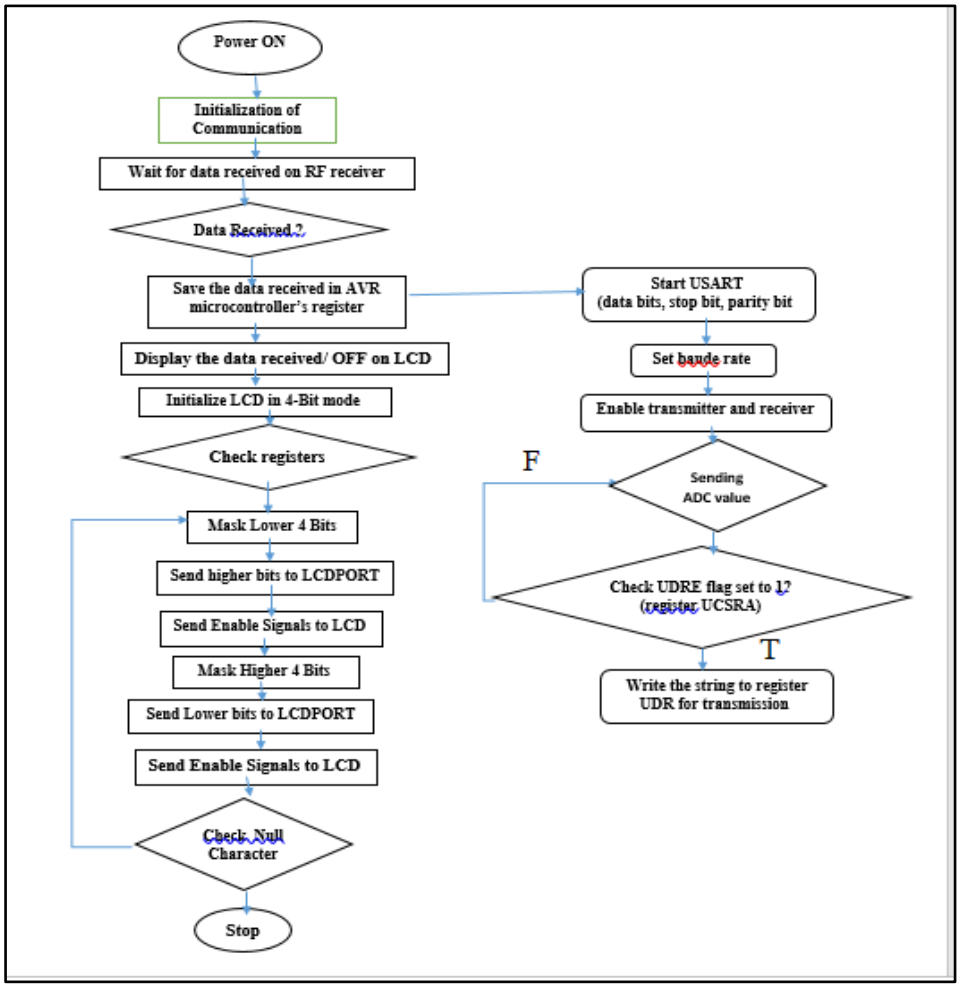


Figure (3): USART communication between two AVR microcontrollers and circuit

3.3. Interface of AVR with LCD

In this part of project, the 16x2 char LCD screen use 4 bit parallel interface with backlight. Using 4 bit mode take 4 lines (for data or command) + 3 control signal total 7 lines. The main objective are:

1- Displaying “Hello Word!!” message on the screen.
2- Interfacing the LCD to the AVR microcontroller using 4 Bit mode.
3- Generating and displaying “ID” and client no.

The connection of the LCD with the AVR microcontroller (ATmega32) is shown in the circuit diagram (figure1,7).

To send string on LCD:

i. Make a string pointer variable
ii. Pass the starting address of string to that pointer variable.
iii. Pass the string pointer value to the LCD_write function.
iv. Increment the pointer value and go to step (iii) till the value of pointer reaches NULL character.

```
int str_to_int(char*ptr)
 {              char base=0;
                int number=1;
                for(uint16 i=0;i<strlen(ptr);i++)
                {
                    number=(ptr[i]-48)+number*base;
                    base=10; }
                return number;}
 void stor_char(char*ptr,char loc)
 {              lcd_send_command(0x40+loc*8);
               for(int i=0;i<7;i++)
               {
                    lcd_send_data(ptr[i]);    } }
```

Figure (4): LCD display send_ string code

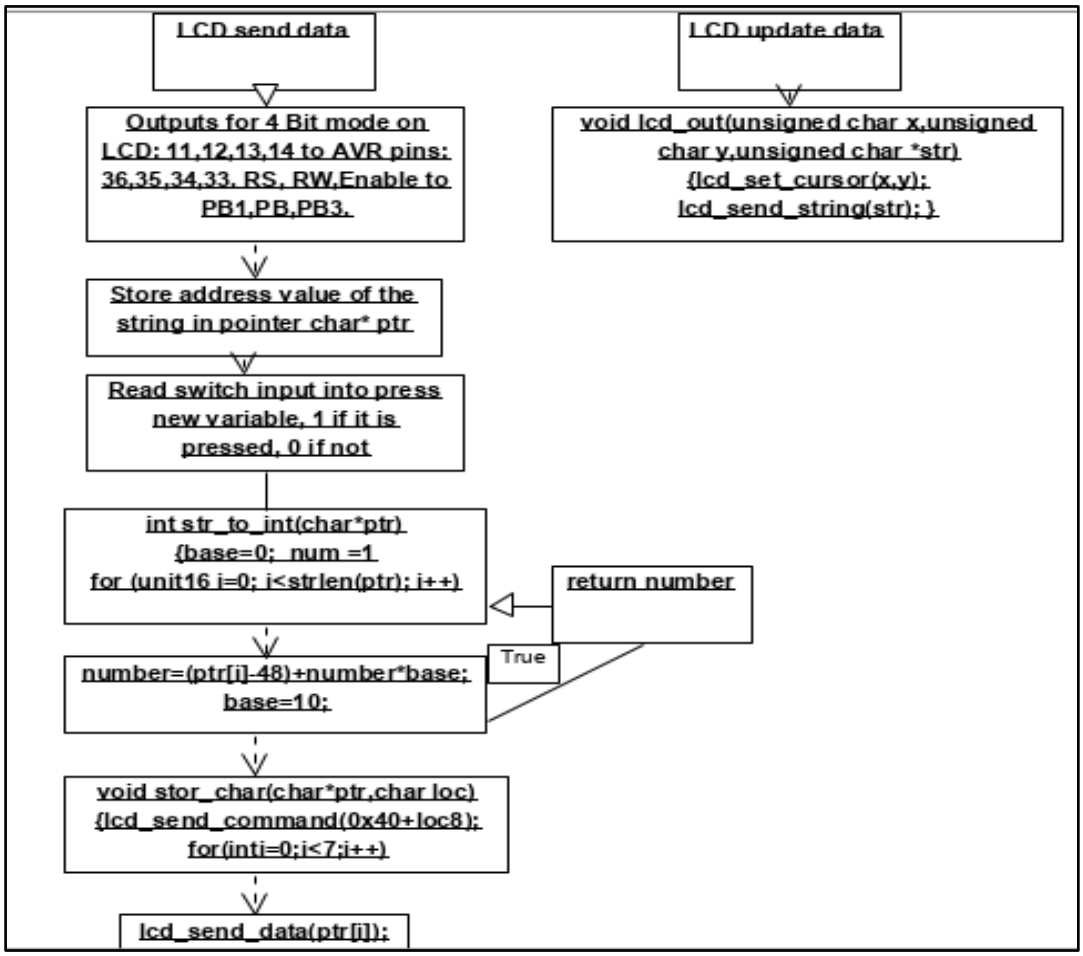


Figure (5): Implementation of LCD send and update data

3.4. Interface of AVR microcontroller and RFID

At the beginning RFID gets the input from the user (read card), each card has three serial numbers stored in an array [2][4] (Volatile unsigned variable X[2][4]), and the system is assigned to 4 bit modes .

Client 1 & client2 are defined to D7 & C1 with initialization =1 (high).Then convert each digit to string that the LCD accepts only stings, search the no. scanned from the card to compare it to the stored serials of each client (client1 & client 2) using flag increment (flag++), two flags are defined (flag 1 & flag 2) then generating a while loop including 2 "for loops" figure (6).

In case flag =8 then change D7 (client 1) value to zero (low) to display client 1 on LCD, (client1, 2, 2) which indicates that "client 1" will be written in the second row at the second column,

Else if flag2 =8 then change C1 (client 2) value to zero (low) to display client 2 on LCD, (client2, 2, 2) which indicates that "client 2" will be written in the second row from the second column (Benyo et al, 2012).

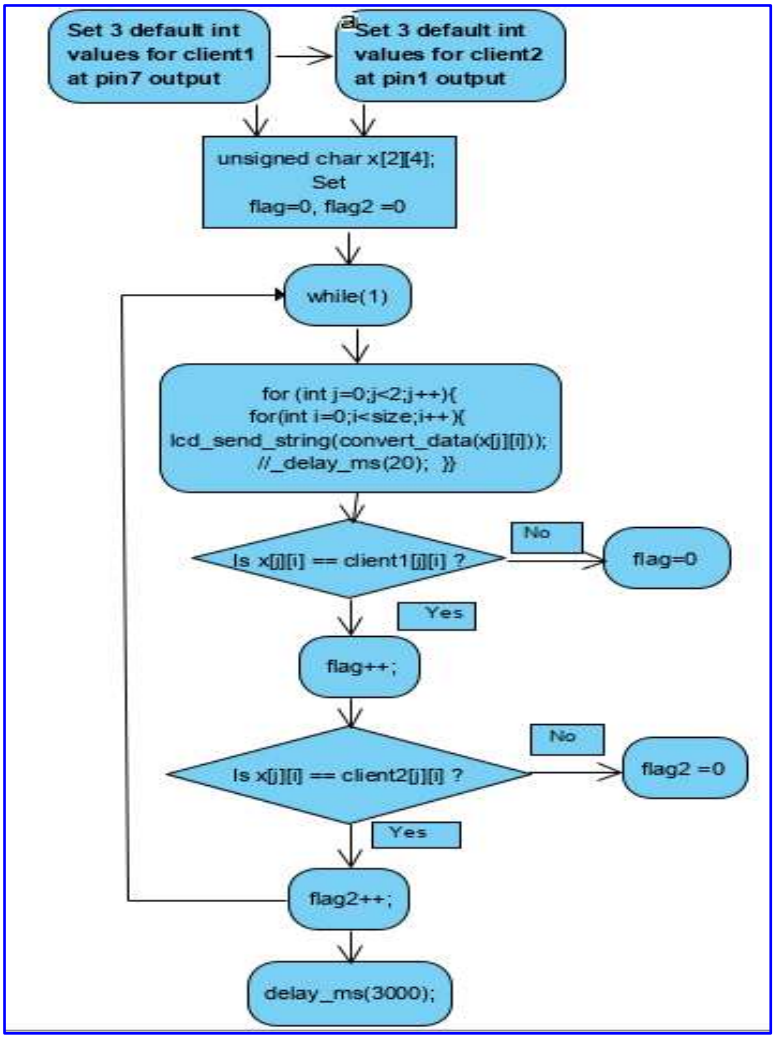


Figure (6): Display of RFID "ID's" on LCD

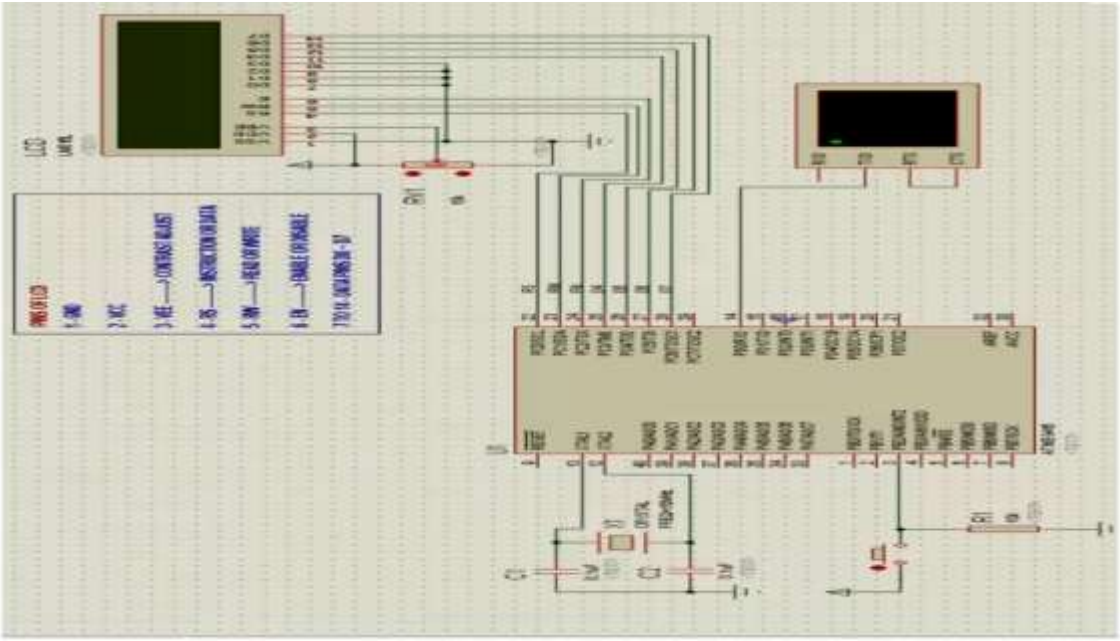


Figure.(7): Block diagram of interface between AVR32A, LCD and RFID [9].

3.5. HC-05 Bluetooth Module Interfacing with AVR ATmega32

The device can be used in 2 modes; data mode and command mode. The data mode is used for data transfer between devices whereas command mode is used for changing the settings of the Bluetooth module. The module works on 5V or 3.3V. As HC-05 Bluetooth module has 3.3 V level for RX/TX and the microcontroller can detect 3.3V level, so, no need to shift transmit level of the HC-05 module. To configure HC-05 to communicate with microcontrollers. The default baudrate of module was 38400. For direct communicate with microcontroller, the settings of HC-05 and its baudrate changed via Laptop and a USB to TTL converter, then to HyperTerminal program. The connections between HC05 module & the USB converter are:

HC – 05 MODULE		**USB-TTL MODULE**
VCC	——>	5.0V
GND	——>	GND
TXD	——>	RXD
RXD	——>	TXD
KEY	——>	5 V

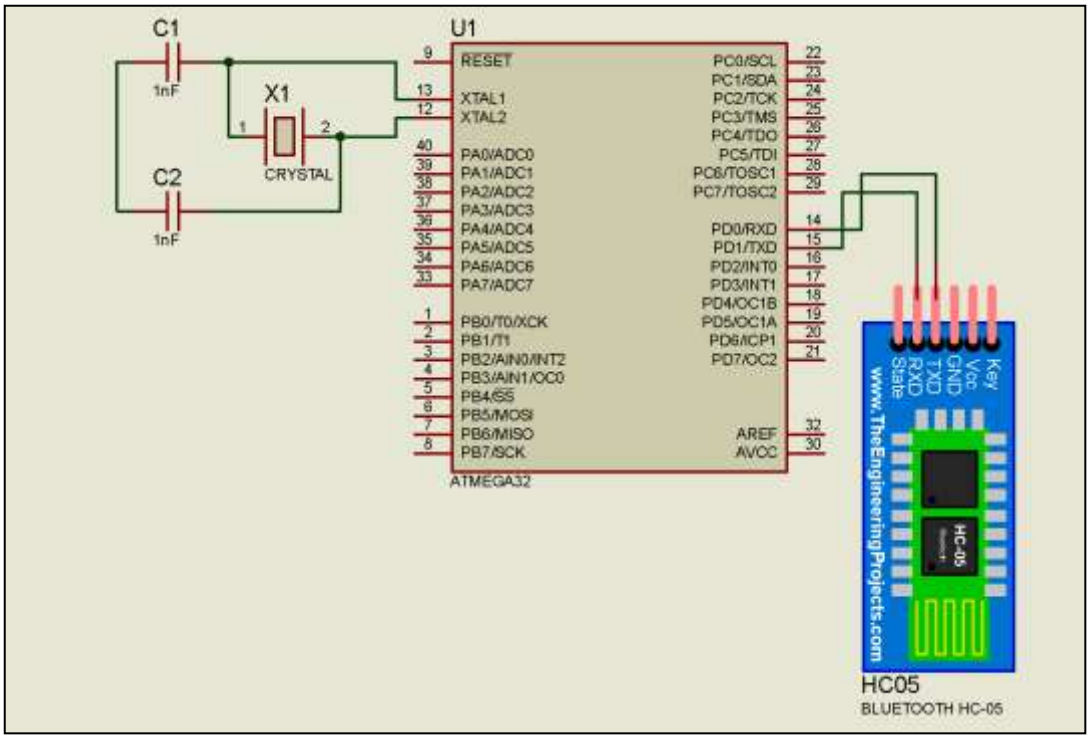


Figure (8) shows interface of Atmega32 µC with HC-05 Bluetooth terminal

3.6. Implementation and results

The system connected using HyperTerminal software. Figure (9) shows valid attendance displaying ID and name (Client1). Figure (10) shows invalid tag with no name displayed.

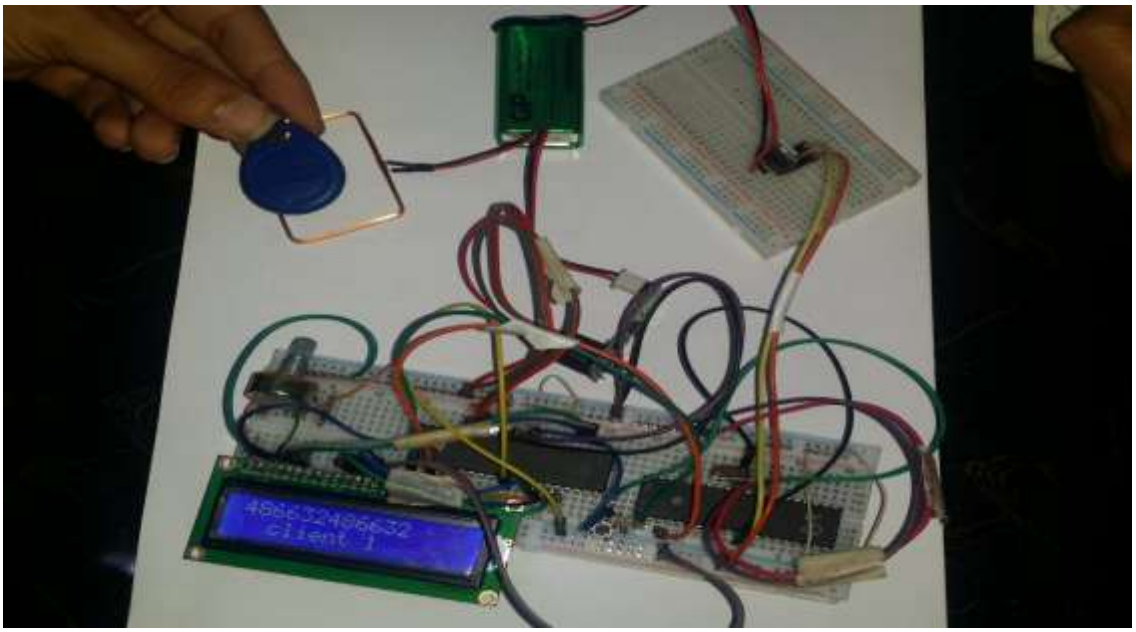


Figure (9): Embedded system project display "client1"

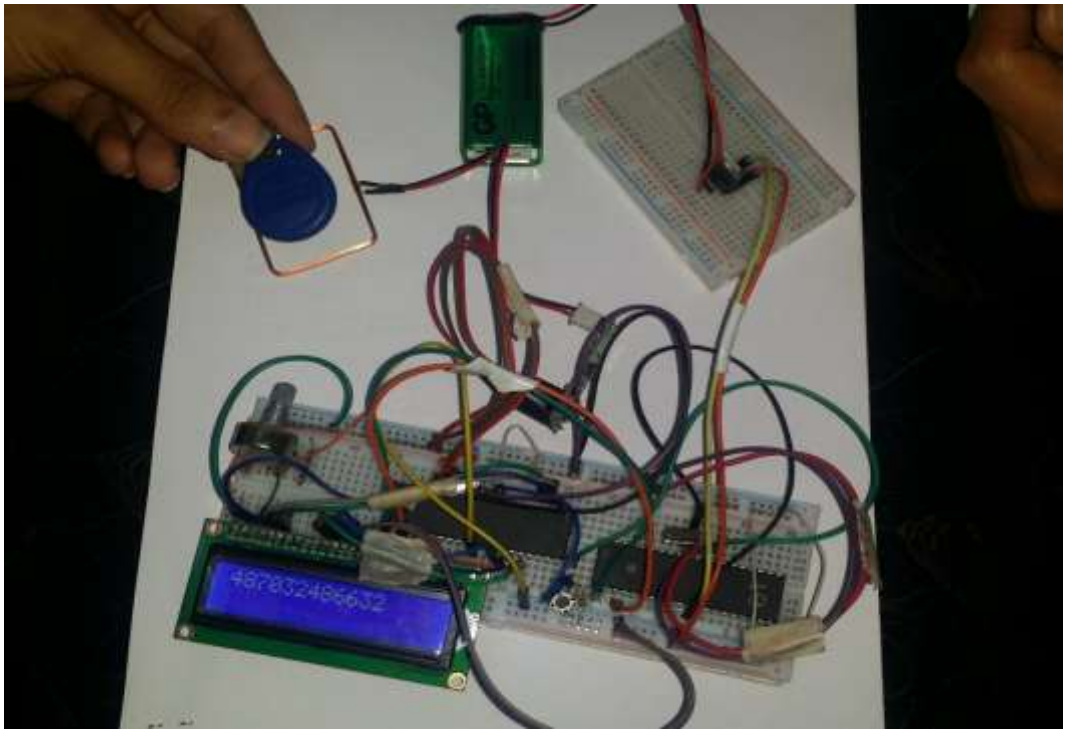


Figure (10): Embedded system project display client "ID"

4. Conclusion and future Work

This application of Bluetooth Smart to student attendance improves the time taken during manual attendance and human errors and provides administrators the statistics of attendance scores for use in further managerial decisions. The approach is to reduce manual work and to automate the attendance system that is basically an embedded one. The developed system implemented with embedded system framework using two atmega AVR. One AVR connected to RFID and LCD and the second AVR interfaced to bluetooth terminal connected to mobile android. There are many details in building process at Construction engineering field. Since our system is modular and can extend effortlessly, several future works will be done to develop our system. The system can use microcontroller with more than one USART for interfacing with other devices. The purpose of the future works is the creation of reliable cloud computing database that can accessed anywhere and cost efficient.

References

Ansari A.N., Navada A., Agarwal S., Patil S., and Sonkamble B. (2011). "Automation of Attendance system using RFID ,Biometrics, GSM modem with . Net framework", IEEE International conference on multimedia technology, pp. 2976-2979.

Arulogun O.T., Olatunbosun A., Fakolujo O.A., and Olaniyi O.M. (2013). "RFID-Based Students Attendance Management System", International Journal of Scientific & Engineering Research Volume 4, Issue 2, ISSN 2229-5518.

Benyo B., Sodor B., Doktor T., and Fordos G. (2012). "Student attendance monitoring at the university using NFC", IEEE Wireless Telecommunications Symposium (WTS).

Kamaraju M. and Kumar P.A. (2015). "Wireless Fingerprint Attendance Management System", IEEE International Conference on Electrical Computer and Communication Technologies (ICECCT).

Lodhaa R., Guptaa S., Jaina H., Narulaa H. (2015). "Bluetooth Smart based Attendance Management System", International Conference on Advanced Computing Technologies and Applications (ICACTA 2015), 45;24-27.

Wang and Jingli (2015), "The Design of Teaching Management System in Universities Based on Biometrics Identification and the Internet of Things Technology", IEEE 10th

International Conference on Computer Science & Education (ICCSE), Cambridge University, UK July 22-24, pp. 979-982.

Yadav D.K., Singh S., Pujari S., and Mishra P. (2015). "Fingerprint Based Attendance System Using Microcontroller and LabView", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 4, Issue 6, pp. 5111-5121.

Zainal N.I., Sidek K.A., Gunawan T.S., and Kartiwi H.M.M. (2014). "Design and development of portable classroom attendance system based on Arduino and fingerprint Biometric", IEEE International conference on information and communication Technology for the Muslim world, Nov. 17-18, 2014.

http://embeddedelectronicsforatmegaseries.blogspot.com/2011/01/tutorial-2-making-development-board-for.html

Enhancing Data Privacy in Heterogeneous Database Applications

Ahmed M. Elbatal, Ahmed M. Gadallah, Ammar Mohamed, Hesham A. Hefny

Department of Computer Science, Institute of Statistical Studies and Research, Cairo University

Abstract

In heterogeneous database applications, where more than one database management system is used, some features provided by one database management system may have different types of implementation in the others, or may even not exist at all. One of the most popular features is the Data Access Control which is used by the database administrator to ensure that each user has access rights to only the part of data he is allowed to see, edit, add, or delete. On the other hand, most of modern database applications are built using Object Relational Mapping (ORM) on top of the data repository as part of the Data Tier (e.g. in a 3-tier architecture). Therefore, data objects can be addressed, accessed and manipulated without considering how those objects relate to their data sources. This main objective of this paper is enhancing the approach used by heterogeneous database applications built using Object Relational Mapping to manipulate data with data access control feature on their higher layers than the database, so that data access policies can be shared across multiple databases with different types of database management system.

Keywords: Data Privacy, Heterogeneous Database, Information Retrieval

Introduction

Data Access Control is one of the main issues that should be tackled in any database application. One of the most useful data access control mechanisms is the Virtual Private Database provided by some database management systems (e.g. Oracle (Huey, 2017) and PostgreSQL (Row Security, n.d.)) to facilitate row-level data access control. However, this mechanism is not yet presented in other relational database management systems (e.g. SQL Server, MySQL, SQLite, etc.). On the other hand, most of modern database applications are built using the Object Relational Mapping that encapsulates the code needed to manipulate the data, so you don't use SQL anymore. Yet, they still depend on the mechanisms provided by database management systems to implement data access control in the lowest tier of their architecture, the database tier. This paper proposes an enhanced approach to facilitate virtual private database mechanism in a way that can be applied to any multi-tier application upon its data access layer regardless of the used database

management system. An illustrative case study shows the benefits of the proposed approach against other approaches.

Figure 1 describes the difference between the traditional heterogeneous database application in multitier architecture, and the proposed approach by which security policies are being defined and shared within the business logic layer then executed in the data access layer before sending data manipulation commands to the underlying data repository.

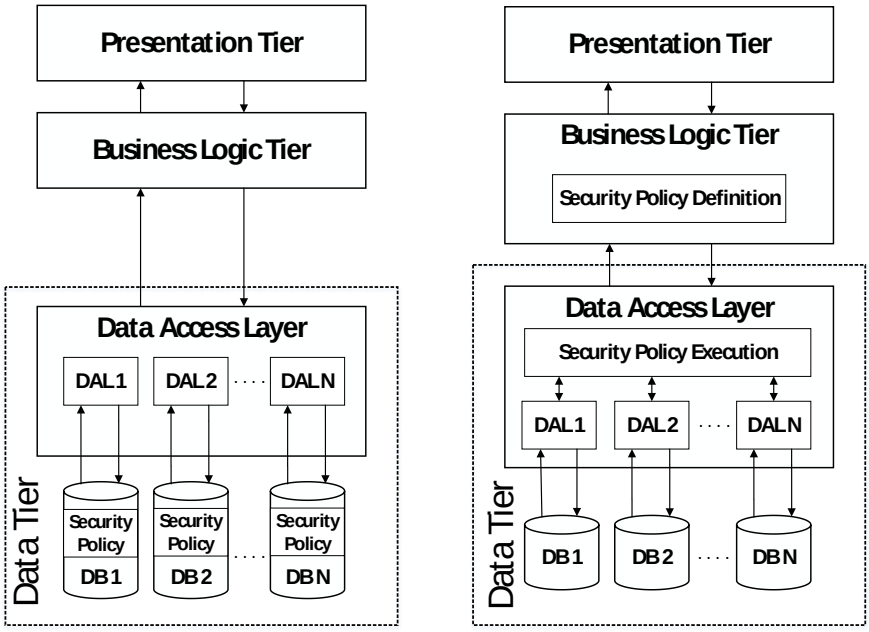


Figure 1. Traditional Heterogeneous Database Application vs The Proposed Approach

Problem Definition

While some vendors provide Row-Level Security feature to their database management systems, not all database management systems have this important feature. Also, not all database management systems that have this feature are implementing it in the same way. This leads to the following issues regard migration and integration processes:

- The cost of migrating the database of the application into a new database management system is very huge, as it will need to rewrite the data security policies in a way that the new database management system uses.
- The migration process could be canceled if the new database management system doesn't support Row-Level Security at all. Or at least it will cost

finding another custom technique to implement Row-Level Security in the database application regard the new database management system.

- In Integrated heterogeneous database applications, where more than one database management system is used simultaneously, Row-Level Security feature costs a lot to be implemented in different ways that the involved database management systems apply.
- It's painful to the application's administrator to maintain different non-shared security policies across all parts of the database application. He/she needs to work on different dashboards and handle redundant security policies that may produce the same condition but in different forms (i.e. depend on the underlying part of the database application).

Previous Work

Generally, data access control mechanisms provided by a database management system are not the only factor to put in mind when choosing the appropriate DBMS for a database application (Thomsen & Pedersen, 2005) (Angles, Prat-Pérez, Dominguez-Sal, & Larriba-Pey, 2013) (Kuhlenkamp, Klems, & Röss, 2014) (Difallah, Pavlo, Curino, & Cudre-Mauroux, 2013). Some customers choose a database management system that fits their business and system requirements such as license cost, scalability, and performance regardless of data access control. Such database management system may lack the appropriate data access control mechanism needed to fulfill the data access control issues. In such case, it becomes very important to build a custom data access control mechanism or using some other mechanisms as a workaround.

Some previous works uses VIEWS to implement row-level security as a workaround that is suitable for all kinds of existing relational database management system. However, this technique has some disadvantages:

- It is hard to administer, as the data access rules are embedded in the views.
- Its implementation is limited and requires careful design and development (United States Patent No. US7155612B2, 2006) (US Patent No. US9870483B2, 2018) (United States Patent No. US7103915B2, 2009).
- In addition, another crucial problem of using such approach takes place when attempting to migrate to another database management system. It would then cost a lot to transform the existing data access control implementation to the other DBMS.

- Moreover, this would cost more if the existing application uses some mechanism that does not even exist in the other database management system (e.g. like the Virtual Private Database in Oracle).
- Also, in a Multitier Application, it is more appropriate to put all the system's business rules including data access rules in the business layer; As putting them in a lower layer (i.e. Database DDL) would violate the multitier architecture of the application (Fowler, 2002).

Although, moving data access mechanism one layer up in a multi-layer application is a subject of contribution in many related works; Corcoran et al. proposed a new programming language called SELinks (Corcoran, Swamy, & Hicks, 2009) (Swamy, Corcoran, & Hicks, 2008). It provides a uniform programming model for building secure multi-tier web applications. Yet, it focused on label-based security (i.e. not as flexible as the Virtual Private Database mechanism). And, as a programming model, it is hard to be implemented and integrated with existing applications. This is due to the effort needed to transform the existing code to meet its equivalent syntax.

In 2014, Le et al. (Li & Xue, 2014) surveys the area of securing web applications from the server side, with the aim of systematizing the existing techniques into a big picture that promotes future research. They discussed SELinks in the perspective of application logic vulnerabilities within new web applications security construction. Other mechanisms mentioned in their survey are more about input validation and data flow tracking between layers rather than securing the data that resides on the database layer (i.e. attached to a database management system).

Recently, researchers are working on enhancing the use of SELinks for the purpose of controlling data flow over multi-tier applications. Balliu et al. toke SELinks as an inspiration to develop a new framework called JSLINQ as an extension of the WebSharper library to track data flow (Balliu, Liebe, Schoepe, & Sabelfeld, 2016).

In a previous work, A. Elbatal et al. (Elbatal, Gadallah, & Hefny, 2018) proposed an enhanced approach that simulates Oracle's virtual private database by implement row-level security using ORM mappers. Yet, it's incomplete simulation as it doesn't implement cell-level security that is also included in Oracle's virtual private database.

Proposed Approach

The proposed approach in this paper aims mainly to extend an ORM package that is popularly used in heterogeneous database application to allow attaching data security policies to its entity types. Accordingly, such defined data security policies can be used later to append the predefined filter expressions when executing data manipulation operations. An example of implementing data security polices in a heterogeneous database system based on the proposed approach is introduced in the case study section. The implementation uses the popularly used ORM package of Microsoft, the Entity Framework. However, the same approach can be implemented using any kind of code generated ORM.

Figure shows the architecture of the proposed approach in three-tier database applications. In such architecture, the security policy is moved away from the database dictionary. Accordingly, the functionality of modifying a SQL statement with the filtering condition is executed in the data access layer which resides in the application runtime environment rather than the database management system. Such a way makes it easy to work with heterogeneous database applications as it simplifies the task of administering the shared security policies instead of handle them in each involved database. Also, it facilitates migration from one database management system to another.

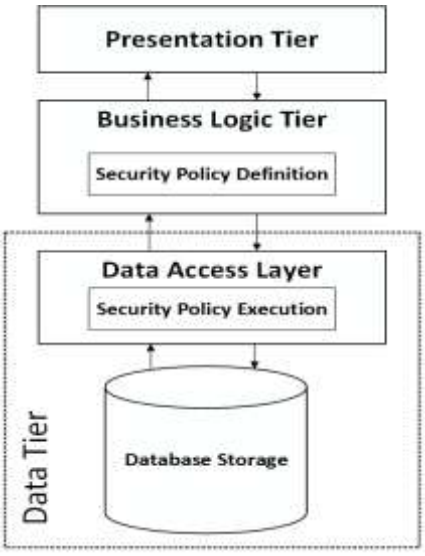


Figure 2. The proposed approach in a three-tier application architecture

Figure presents the architecture of the proposed approach in heterogeneous three-tier database applications. Consequently, the security policy definition becomes part of the business logic layer. Accordingly, the security policies are considered

as collection of business rules that are defined in the business logic layer in the form of predicate functions. On the other hand, the extended ORM package is shared among all DALs involved in the heterogeneous application. This extended package contains the necessary classes to do the functionality of executing security policies; which is abstracted away from being a part of each separated DAL. A simple implementation of the relationship between DALs and the shared ORM package can be done by using inheritance. Such that each separated DAL inherits its ORM model from the base classes that contain the security policy execution functionality.

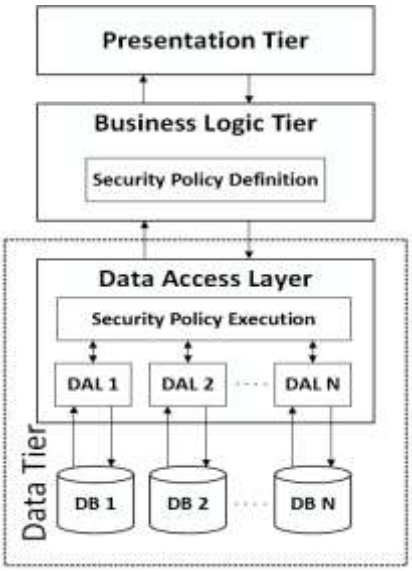


Figure 3. The proposed approach in a heterogeneous three-tier application architecture

An Illustrative Case Study

This case study represents an implementation of the proposed approach. It incorporates a set of most popular DBMS's (i.e. MSSQL, MySQL, and Oracle) using a simple HR database that existed on each DBMS with typical large amount of data.

Objectives

The main objective of this case study is to show the flexible applicability of the proposed approach with any type of DBMS that is supported by the Entity Framework. Another objective is to compare the performance of the proposed approach against Oracle VPD in the same environment (i.e. using Entity Framework as the DAL). Also, the case study shows the cost of processing different manipulation operations on a set of heterogeneous DBMSs that obey or violate a predefined data access security policy.

Schema and Data Sample

A simple HR schema is considered for the case study that includes two tables as follows:

- Employees (EmployeeId, FirstName, LastName, Email, Phone, Address, City, State, ZipCode, Country, SSN, Salary, BirthDate, DepartmentId) with 10 million records (about 2.5 GB).
- Departments (DepartmentId, DepartmentName), with 10 records.

Application Structure

The .NET application of the case study has been made up of a set of layers and sub-projects. The database layer includes three databases with the same database state on MSSQL, MySQL, and Oracle DBMSs. On the other hand, The DAL incorporates a library named EFRLS that contains the required classes that encapsulate the implementation of the data access security policies. Also, The DAL includes four entity framework sub-projects. Three of them are Hr.Model.MSSQL, Hr.Model.MySQL, and Hr.Model.Oracle that represent the entity framework DAL for each corresponding DBMS (MSSQL, MySQL, and Oracle DBMSs respectively). The fourth sub-project called Hr.Model.OracleVPD represents the normal entity framework DAL of the Oracle database without applying the proposed approach. The Hr.Model.OracleVPD is responsible for creating the required security policies respecting the Oracle supported VPD feature.

Testing the Proposed Approach

The implemented application starts with instantiating each of the four DALs. In consequence, it assigns their application context that will be used later in the security policy's predicate function. It then adds a simple security policy to all of those instances. Finally, it separately executes the operations stated in

Table 1, from operation Op01 to operation Op13, on HR databases existed on different database servers via their corresponding entity framework's DAL.

Table 1. A set of SQL operations that are considered in the experiment

Operation ID	SQL Statement	SQL Operation
Op01	SELECT	Retrieving the first 1 record that matches a given filter expression.
Op02	SELECT	Retrieving the first 10 records that match a given filter expression.
Op03	SELECT	Retrieving the first 100 records that match a given filter expression.
Op04	SELECT	Retrieving the first 1000 records that match a given filter expression.
Op05	UPDATE	Updating an existing entity with values that obey an assigned security policy.

Op06	UPDATE	Updating an existing entity with values that violate an assigned security policy.
Op07	UPDATE	Updating an existing entity with no security policy assigned.
Op08	INSERT	Adding a new entity with values that obey an assigned security policy.
Op09	INSERT	Adding a new entity with values that violate an assigned security policy.
Op10	INSERT	Adding a new entity with no security policy assigned.
Op11	DELETE	Removing an existing entity with values that obey an assigned security policy.
Op12	DELETE	Removing an existing entity with values that violate an assigned security policy.
Op13	DELETE	Removing an existing entity with no security policy assigned.

Results

The illustrative case study shows that the proposed approach can be easily applied with the entity framework DAL. It also shows good performance compared with the Oracle VPD especially on data manipulation operations (i.e. insert, update, and delete). This because the process of checking the assigned security policies on adding a new entity or updating or deleting an existing one is performed on the DAL runtime environment rather than passing it to the DBMS. Also, it can be noted that the currently used approaches in data manipulation operations in a DBMS that does not support row level security is to put the validation process (i.e. an IF condition) inline in a higher level than the DAL. So, it does not take more than one tick to proceed.

In this case study a set of different data manipulation operations are considered.

Table 1 shows a set of data manipulation operations that are applied in the case study. On the other hand, both of Table 2 and Table 3 show the result of the case study. The first column of each table shows the SQL operation that is being executed by the DAL. On the other hand, the second column shows the number of records affected/selected per each executed SQL operation. For each DBMS involved in the case study, "Normal" refers to the currently used approach in three tiers architecture (i.e. putting the required filter expression inline in a higher layer than the DAL), "VPD" refers to the using of virtual private database approach (i.e. in Oracle Database only), and "Proposed" refers to the implemented one respecting the proposed approach. Also, in this case study, the time cost of executing each SQL statement is measured by a number of CPU ticks (one tick = 10^{-4} millisecond) elapsed until processing that SQL statement. Accordingly, the numbers in the table show the CPU ticks of performing each operation. It has been calculated as an average of the elapsed ticks of 100 iterations per

process. For better visualization, Figures 4, 5,Figure , and 7 also show the same results in a bar chart format.

Table 2. The processing cost of executing different data retrieval operations respecting a predefined data access security policy using the proposed approach versus the used ones in Oracle, MSSQL and MySQL

Operation ID	Count of Records	Oracle			MSSQL		MySQL	
		Normal	VPD	Proposed Approach	Normal	Proposed Approach	Normal	Proposed Approach
Op01	1	5800	15872	6858	3001	3074	3398	3653
Op02	10	6342	16361	7096	8083	8468	8876	8934
Op03	100	7014	22818	7958	31050	31654	10425	11402
Op04	1000	7015	29048	12620	32038	34207	25206	15943

Table 3. The processing cost of executing different data manipulation operations respecting a predefined data access security policy using the proposed approach versus the used ones in Oracle, MSSQL and MySQL

Operation ID	SQL Statement	Rules	Oracle			MSSQL		MySQL	
			Normal	VPD	Proposed Approach	Normal	Proposed Approach	Normal	Proposed Approach
Op05	INSERT	Valid Rules	6190	16198	9945	4301	7144	4216	9432
Op06	INSERT	Invalid Rules	1	16206	3223	1	1790	1	1934
Op07	INSERT	No Rules	5919	14781	9661	7051	7482	4080	5607
Op08	UPDATE	Valid Rules	3153	12167	7669	2927	5119	10864	12706
Op09	UPDATE	Invalid Rules	1	15179	2393	1	1980	1	2082
Op10	UPDATE	No Rules	3073	12374	4169	3244	3908	10057	13604
Op11	DELETE	Valid Rules	3303	12100	5782	2673	4398	8057	10319
Op12	DELETE	Invalid Rules	1	11350	4406	1	3731	1	3894
Op13	DELETE	No Rules	4355	13475	5798	3908	4582	11196	12917

As shown in the results of the case study presented in Table 2 and Table 3, the processing cost of one DAP is almost the same when applied on different DBMS as it takes place in a same shared layer, the DAL runtime environment. Yet, there are different processing costs of the same data manipulation operation due to different processing scenario of each DBMS to manipulate such operation. Accordingly, the overall cost of performing a SQL operation respecting a defined data access security policy differs from one DBMS to another one.

Also, it can be noted that the time cost of executing a data security policy respecting the proposed approach is less than executing the same policy in Oracle DBMs using its supported virtual private data security VPD. As shown in Table 2, the proposed approach compared with Oracle VPD reduces the processing cost of data retrieval operations respecting a predefined data access security policy by around 59%. On the other hand, it reduces the processing cost of data manipulation operations compared with Oracle VPD respecting a predefined data access security policy by around 57% in average. This is reasonable as the process of generating a query statement in the proposed approach takes place in the application's runtime environment rather than the DBMS. Such generated query statement incorporates a *where* clause satisfying the assigned policies. Accordingly, the DBMS becomes responsible for processing a query statement without suffering from the cost of checking a security policy.

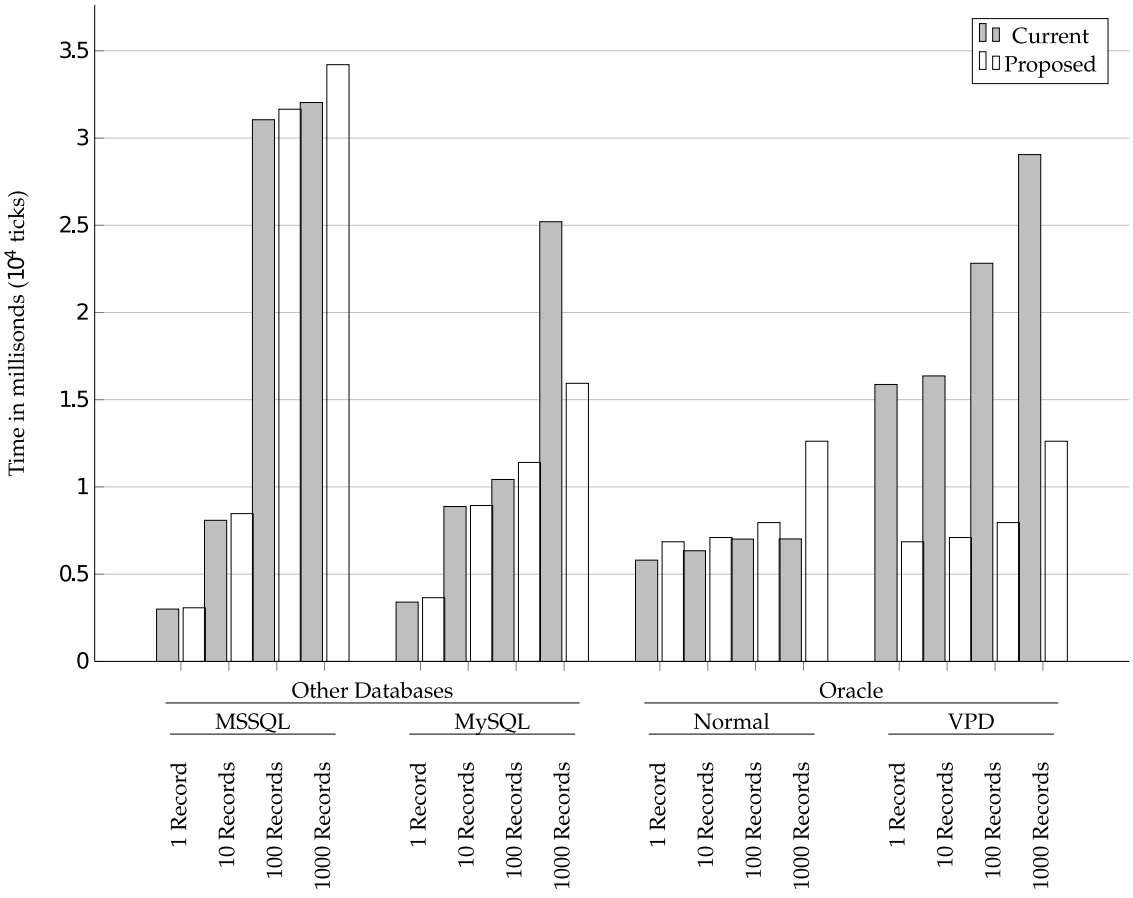


Figure 4. The case study results of the select statement

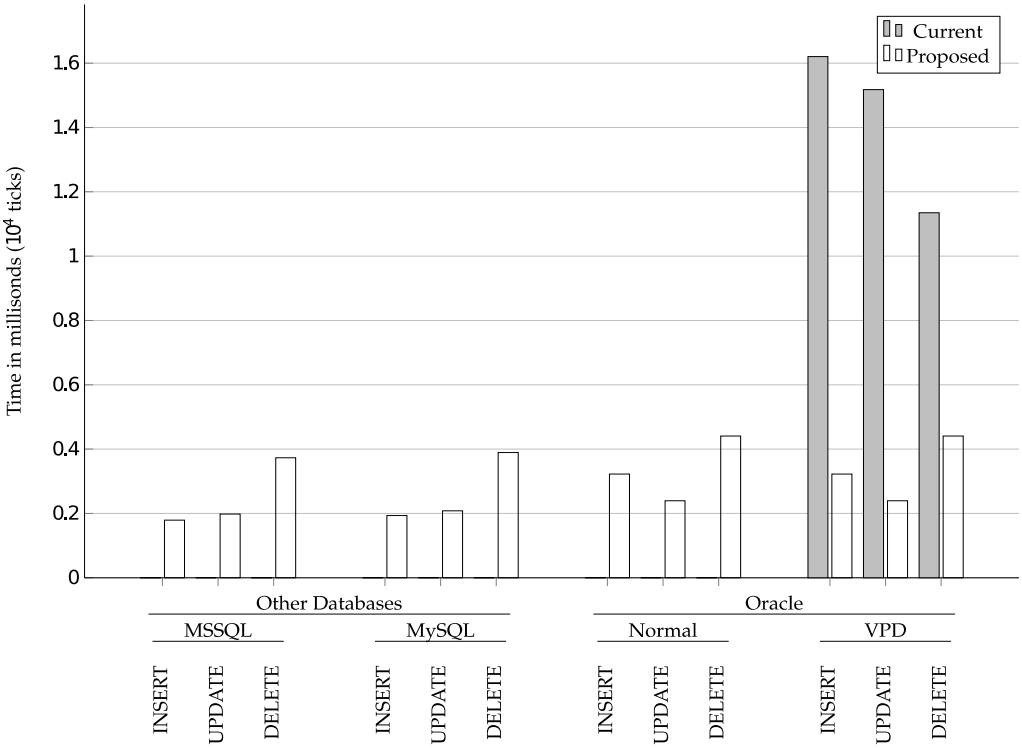


Figure 5. The case study results of the CRUD statements with invalid rules

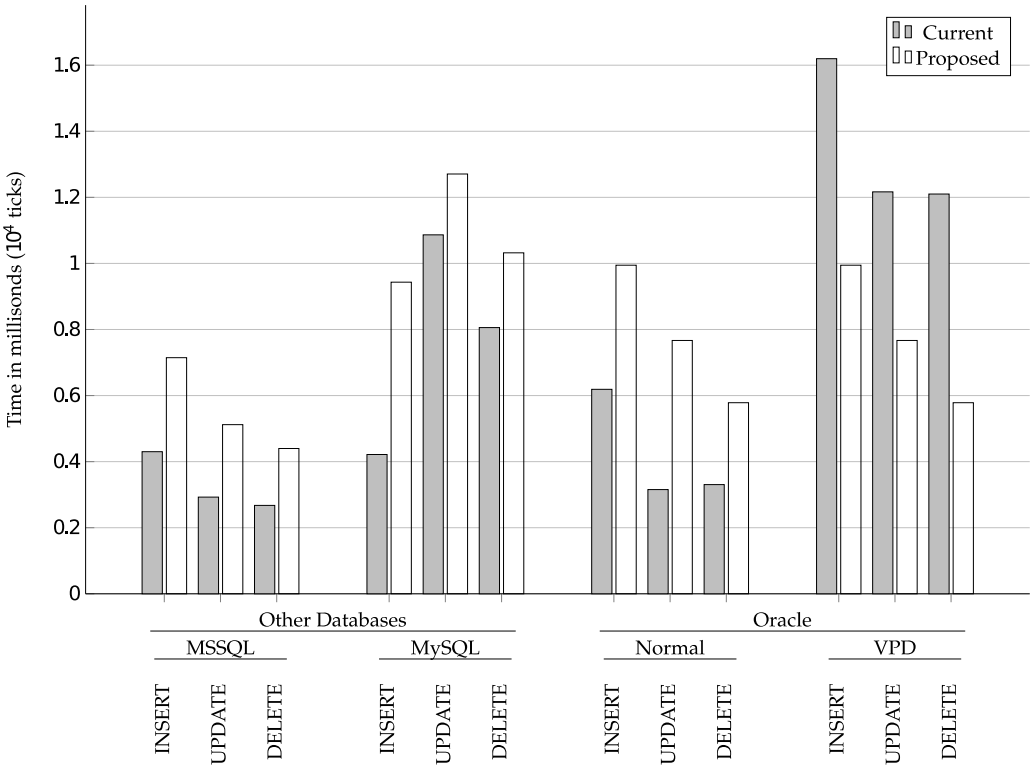


Figure 6. The case study results of the CRUD statement with valid rules

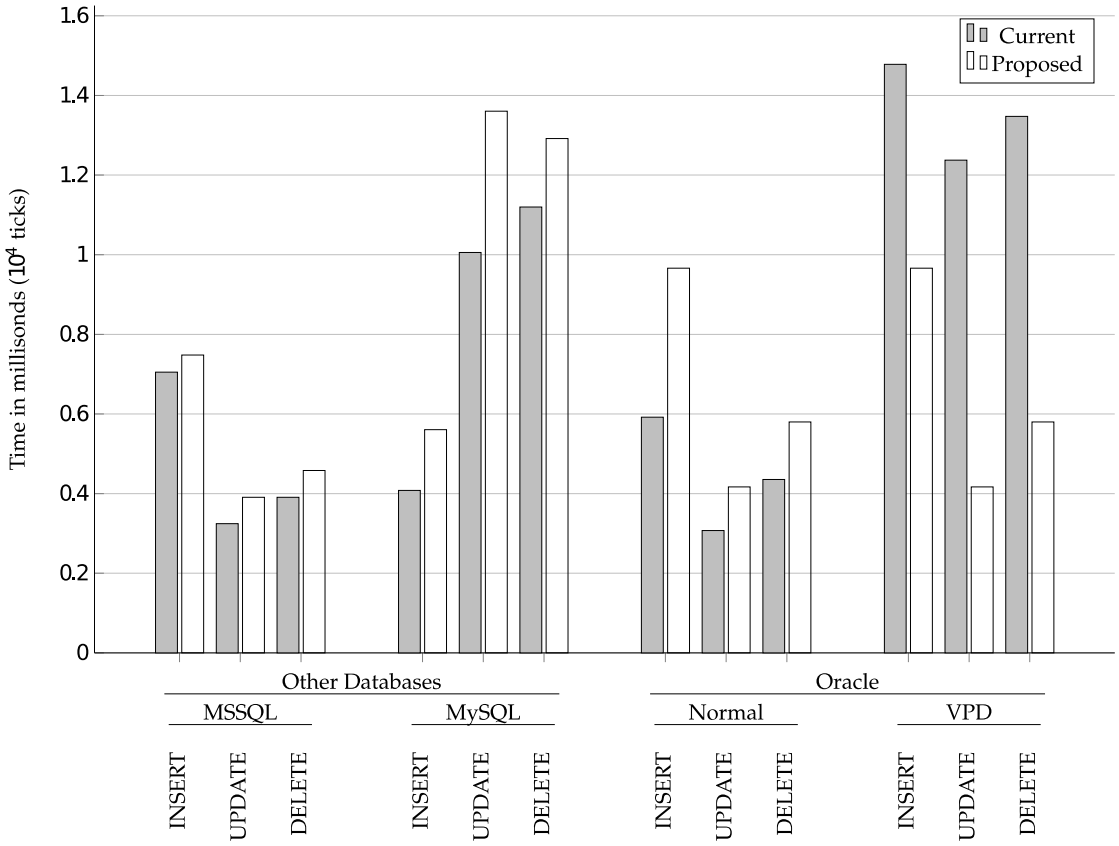


Figure 7. The case study results of the CRUD statements with no assigned rules

Conclusion

The proposed approach attempts to move the data access security one level up in a multitier application to the DAL instead of being a part from the DBMS itself. Accordingly, it can be used with all types of DBMS regardless of the security features supported by the selected DBMS for the database application. Also, the proposed approach is more flexible when migrating from one DBMS to another. The effort needed in case of such migration is just rebuilding the DAL (e.g. using the Entity Framework) to cope with the new DBMS with its appropriate database provider. On the other hand, the proposed approach fits better with the multitier architecture in that the business rules related to data privacy can be now represented in its appropriate position in a multitier application, the business logic layer. It modularizes the data access security needed in heterogeneous database applications.

Nevertheless, we must notice that it also comes with some shortages. It doesn't prevent data access to the DBMS directly. That is, if someone issues a SQL statement to the database directly using a database tool like SQL Plus, he would easily retrieve all data. Also, since the

implementation of the proposed approach depends on the framework used to build the DAL. So that, if needed to use another framework (e.g. other than the Entity Framework), a modification of the code generated from the new framework is required to perform the same mechanism. It also takes the task of maintaining data access security away from the database administrator.

References

Afyouni, H. A. (2005). In *Database Security and Auditing: Protecting Data Integrity and Accessibility* (pp. 207-256). Cengage Learning.

Ahmed, T., Ru, Y., Liang, C., & Pesati, V. R. (2019, May 28). *US Patent No. US10303894B2.*

Angles, R., Prat-Pérez, A., Dominguez-Sal, D., & Larriba-Pey, J.-L. (2013). Benchmarking Database Systems for Social Network Applications. *First International Workshop on Graph Data Management Experiences and Systems*, (pp. 1-7). New York. Retrieved from http://doi.acm.org/10.1145/2484425.2484440

Balliu, M., Liebe, B., Schoepe, D., & Sabelfeld, A. (2016). JSLINQ: Building Secure Applications across Tiers. *Proceedings of the Sixth ACM on Conference on Data and Application Security and Privacy*, (pp. 307-318).

Barry, D., & Stanienda, T. (1998, Nov). Solving the Java Object Storage Problem. *Computer, 31*(11), 33-40.

Berkus, J. (2009, Aug). Wrecking Your Database. *Computer*.

Block, M., Hohner, C., Schindewolf, M., & Zorn , S. (2019, Mar 19). *US Patent No. US10235531B2.*

Colombo, P., & Ferrari, E. (2019). *Access control technologies for Big Data management systems: literature review and future trends.* Springer.

Corcoran, B. J., Swamy, N., & Hicks, M. (2009). Cross-tier, label-based security enforcement for web applications. *Proceedings of SIGMOD*, (pp. 269-282).

Cotner, C., & Miller, R. L. (2018, Jan 1). *US Patent No. US9870483B2.*

Dettinger, R. D., & Kolz, D. P. (2018, Oct 23). *US Patent No. US10108813B2.*

Difallah, D. E., Pavlo, A., Curino, C., & Cudre-Mauroux, P. (2013). An extensible testbed for benchmarking relational databases. *Proceedings of the VLDB Endowment, 7*(4), 277-288.

Elbatal, A., Gadallah, A. M., & Hefny, H. (2018). An Approach for Enhancing Data Access Security in Heterogeneous Database Systems. *Indian Journal of Science and Technology, 11*(2).

Elmasri, R. (2011). Data Security. In *Fundamentals of database systems (6th edition)* (p. 847). Pearson Education.

Fowler, M. (2002). Organizing Domain Logic. In *Patterns of Enterprise Application Architecture* (pp. 25-29). Addison Wesley.

Gopi, A. P., Choegyen, T., V, S., Jois, S., & Herbst, A. (2017, May 11). *U.S. Patent No. 14/934,629.*

Huey, P. (2017). Using Oracle Virtual Private Database to Control Data Access. In *Oracle Database Security Guide 11g Release 2 (11.2)* (pp. 249-288). Oracle Corp.

Knox, D. (2004). In *Effective Oracle Database 10g Security by Design* (pp. 337-400). McGraw-Hill Education.

Kuhlenkamp, J., Klems, M., & Röss, O. (2014). Benchmarking scalability and elasticity of distributed database systems. *Proceedings of the VLDB Endowment, 7*(12), 1219-1230. Retrieved from http://dx.doi.org/10.14778/2732977.2732995

Lei, C. H. (2018, Feb 6). *US Patent No. US9886481B2.*

Li, X., & Xue, Y. (2014, April). A survey on server-side approaches to securing web applications. *ACM Computing Surveys (CSUR), 46*(4), 54. Retrieved from http://doi.acm.org/10.1145/2541315

Licis, N. D. (2006, Dec 26). *United States Patent No. US7155612B2.*

Mujumdar, P., Khanduja, R., & Verma, P. (2019, May 23). *US Patent No. 20190155794.*

Redlich, R., & Nemzow, M. (2009, Sep 5). *United States Patent No. US7103915B2.*

Row Security. (n.d.). Retrieved Feb 01, 2016, from https://wiki.postgresql.org/wiki/Row-security

Sheth, A. P., & Larson, J. A. (1990, Sep). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR), 22*(3), 183-236.

Swamy, N., Corcoran, B. J., & Hicks, M. (2008). Fable: A language for enforcing user-defined security policies. *Proceedings of the 29th IEEE Symposium on Security and Privacy*, (pp. 369-383).

Thomsen, C., & Pedersen, T. B. (2005). A Survey of Open Source Tools for Business Intelligence. *International Conference on Data Warehousing and Knowledge Discovery* (pp. 74-84). Springer Berlin Heidelberg.

Ontology with Deep Learning Method for sign language semantic translation system

Eman K. Elsayed[1], **Doaa Rezk**[2]
Mathematics and Computer Science,
Al-Azhar University (Girls Branch),
Cairo, Egypt

Abstract:

Ontology is a formal language used to refer to the elements of conceptualization. A Convolutional Neural Network (CNN) is a powerful deep learning technique used successfully in image classification problems. In this paper, we enhanced image classification via a hybrid method between deep learning and Ontology. We build an enhanced CNN architecture with new regularization techniques and adding semantic layer that maps the image to a formal representation of its semantic meanings. Merging deep learning and Ontology produced our first version from semantic deep learning (SDL). As a case study, we present a Multi Sign Language translation system. That is to interpret a user's signs into different languages. Some of the linguistic background has been taken into consideration, syntax and semantics. The system is useful for the deaf, mute community and can be used by common people who migrate to different regions and do not know the local language. Experimental results show that the training accuracy obtained using CNN model at 100 epochs was 98.66% and the semantic recognition accuracy of testing gestures was 95.45 %.

Keywords: Sign Language, Ontology, Convolutional Neural Network (CNN), Deep Learning (DL).

1. Introduction

Sign language considered the only way for communication between deaf-mute and normal people. According to the World Health Organization (WHO), there are around 466 million people around the world have disabling hearing loss, and more than 28 million of these are Americans, 13 million people within Egypt across all age groups (Reda et al., 2018). People use sign language gestures as a means of non-verbal speech to express their thoughts and emotions.

Sign language differs from one country to another. In Arabic countries, there are several sign languages like Jordanian, Saudi, Egyptian, and Yemeni (Al-Fityani and Padden 2010). Differences in sign language may be found among speakers in the same country and speakers in neighboring cities. This is because signs mostly created by hearing-impaired individuals themselves and are highly affected by the local environment (Shanableh et al., 2007).

Deaf and mute people have all the right to speak to each other even for different sign languages. Sign languages need an intelligent system to translate sign language to another based on natural languages. It is hard for most people who are not interested in sign language to communicate without an interpreter.

Human computer interaction (HCI) has proposed as an alternative to the traditional input devices such as a keyboard camera or a mouse. Humans mostly used their hands to communicate or to interact with the environment. So, hand gestures can play an important role in human-computer interaction. Automatic sign language translation systems are important for solving these problems.

Many researchers focused on translation sign language to text and spoken language and vice versa. But this is not easy to be done by machines since it depends on the natural language processing and image recognition. On the other hand, the traditional way of translation needs a translator that specializes in sign language may not be present in all situations. Translation of Arabic sign language (ArSL) is facing many challenges; the lack of linguistic studies on ArSL, sign language (SL) is assumed to be a universal language.

In the Artificial Intelligence field, there are many definitions of Ontology. One of these definitions is an explicit specification of a conceptualization. Ontology is used for analyzing and reusing knowledge. Ontology explicit description of individual, classes, attributes, relations, restrictions, rules, axioms, and events. Individuals represent objects or instances in the domain. The relation is binary on individuals; i.e., the relation links two individuals together or two classes together. The class is the concept or type of object. The attribute is property, feature, and characteristic. The function is a complex structure formed from relations between two individuals can have used them in a statement. Restrictions represented formally stated descriptions used in some assertions. The rule is a statement used in logical inference to draw from assertions. The previous Ontology definition and structure summarized by (Zhang et al., 2014).

WordNet (WN) is a linguistic resource consists of words interconnected by their meanings through lexical. A single word connected to another through semantic conceptual relations between them (Miller et al., 1990). These relations provide semantic information about concepts and their original words. These concepts and their relations are exploited to improve Arabic Information Retrieval, Text Classification, and Text Summarization Technically. Arabic WordNet lexicon provides a good semantic structure for computing the semantic similarity between words. The semantic structure of Arabic WordNet can be presented as a tree graph. The nodes of the tree graph are sets of synonyms and the edges of the tree graph are semantic relations.

In this paper, we used a deep learning architecture for Sign Language Recognition. In recent years; Deep Learning (LeCun et al., 2015) algorithms have been successfully applied on image recognition problems. Deep CNN‟s introduces many hidden layers, for reducing the dimensionality of the image and enabling the model to extract image features in low dimensional space.

The main objectives of our research are: to enhance Arabic sign language translation to Arabic text and English and French sign or text using the power of semantic web technologies (i.e. ontologies), also using deep learning in the image recognition process. To advance the research in the domain of automatic multi sign language translation, proposed a Multi-Sign Language semantic translation system. We have two sets of experiments: Arabic sign image classification, Arabic sign image semantic translation. To train our model for classification and recognition, we create an Arabic sign image dataset to train and test the semantic translation system.

The rest of this paper is organized as follows. In the next section, we present some related works. Then the proposed Multi Sign Language Ontology (MSLO) will present in section 3. The details of the proposed method "Semantic Deep Learning" are introduced in section 4 and its subsections. Section 5 presents the application of the proposed method on a case study in Arabic Sign Language, while section 6 presents the analysis of results. And, finally, conclusion and future work.

2. Related works

In recent years, several research projects in developing sign language translation systems have been developed. But Most of them worked only on translating text to sign or sign to text. They did not take care of the semantics of the translated text. As an example, paper in reference (Ibrahim et al., 2018) presented an automatic visual technique that translates individual Arabic sign to text word. Geometric features of the hands are employed to formulate the feature vector. Euclidean distance classifier is applied for the classification stage. A dataset of 30 isolated words of the hearing-impaired children was developed. The system has a recognition rate of 97%.

Firstly, some related works on sign language and ontology. The first work in using ontology to enhance sign language translation was ATLAS (Lesmo et al., 2011). ATLAS was a project for automatically translate from Italian text to Italian Sign Language (LIS). The translation system communicates with the user through a virtual signer: the system takes a text written in Italian language and translates it into a formal intermediate representation of a sign language sentence called ATLAS Written Italian Sign Language (AWLIS). AWLIS sentences then translated into a character's gestures Animation Language (AL) that describes the way the basic movements are produced and linked. Reference (Almasoud and Al-Khalifa, 2012) presents a proposed system for semantically translating Arabic text to Arabic Signwriting in the jurisprudence of prayer domain using ontology as a semantic web technology. The system designed to translate Arabic text by applying Arabic Sign Language (ArSL) grammatical rules as well as semantically looking up the words in Ontology. Reference (Lozynska and Davydov, 2015) focuses on the development of information technology for Ukrainian Sign Language translation based on grammatically augmented ontologies.

Secondly, some related works in sign language recognition based on CNN. Although a variety of methods have been proposed in recent years to recognize hand gestures, most of them focus on using deep learning. ElBadawy et al. (ElBadawy et al., 2017) developed the system recognition for Arabic sign language. In this system, Convolutional Neural Network (CNN) was used to recognize 25 gestures from the Arabic sign language dictionary. The system achieved 98% accuracy for observed data and 85% average accuracy for new data.

3. Multi Sign Language Ontology (MSLO)

We proposed the Multi Sign language Ontology (MSLO) which is a linguistic Ontology takes the advantage of WordNet and multilingual WordNet in sign language domain. But analysis of the nature of sign language leads us to create some different relations to solve many sign language challenges.

We proposed MSLO's first version to solve these challenges. It extends the WordNet relations to deal with sign language. The signs are created in correspondence with the WordNet synsets', and semantic relations are imported from the corresponding synsets. We assume that if there are two synsets WN and a relation holding between them, the same relation holds between the corresponding signs in MSLO.

Generally, using ontology enables signs to express words (concepts). So we can say that it is important to sign a concept rather than a word. Ontology individuals (words) might have more than one name.

MSLO solves some challenges of sign language. The following examples explain three challenges:

1) Each sign means a bag of words in each language synonyms. For example, Figure 1 presents the sign of “car" in sign language. The same sign could represent the English words: Vehicle, automobile, transportation, and auto. The same sign also used to represent the Arabic words "سيارة"، "مركبة"،"حافلة"،"عربة"".

Figure 1. *The word “CAR” in sign language*

2) The words in the same meaning may be expressed in different signs. These differences give the difficulty of communicating and dealing with deaf people. For example, the word “bad” in English has the sign (a) in figure 2. The word “سيئ”"meaning of word bad" in Arabic has the sign (b) in Figure 2, In MSLO, the two words related to each other but have different signs.

Figure 2. *(a)"سيئ" gesture sign (b) bad gesture sign*

3) One sign may be used to represent different words such as; sign in Figure 3, represents the word "لا", the Character "ب" and number "1" in the Arabic language. These cases solved in MSLO ontology where sign gesture interprets based on its class and domain.

Figure 3. *Gesture sign of the word "لا", Character "ب" and the number "1"in the Arabic language*

Also, Each person has his-self a way to sign any word in any language.

The main concepts of MSLO Ontology are English_WORDNET, French_WORDNET, and Arabic_WORDNET. Each language consists of sub-classes. Each subclass is a bag of words from one domain.

We proposed some refinements to enhance WordNet relations in MSLO to be suitable with sign language.

The properties we used to describe a class/concept or individuals:

a) Object properties used for the relationship between words:

- Arabic_ meaning
- English_ meaning
- French_ meaning

- Has_the_same_sign; to relate the concepts which have the same gestures signs.

b) Data properties that used to describe words are:

- Arabic_sign
- English_sign
- French_sign
- DL_label to assign the word to its label in the deep learning model

c) Annotation property (label) also used to add all possible meanings of the words.

The link of the sign gesture image is saved as a data property value. Also, MSLO takes into consideration personality signs. There is a class for personal information in sign language to represent personal data; like name, address and birth date.

Deaf, dumb and normal users can interact with the ontology via interaction system to develop and update his Personal information (My-Dictionary class) and its corresponding signs. Part of MSLO Ontology is shown in Figure 4.

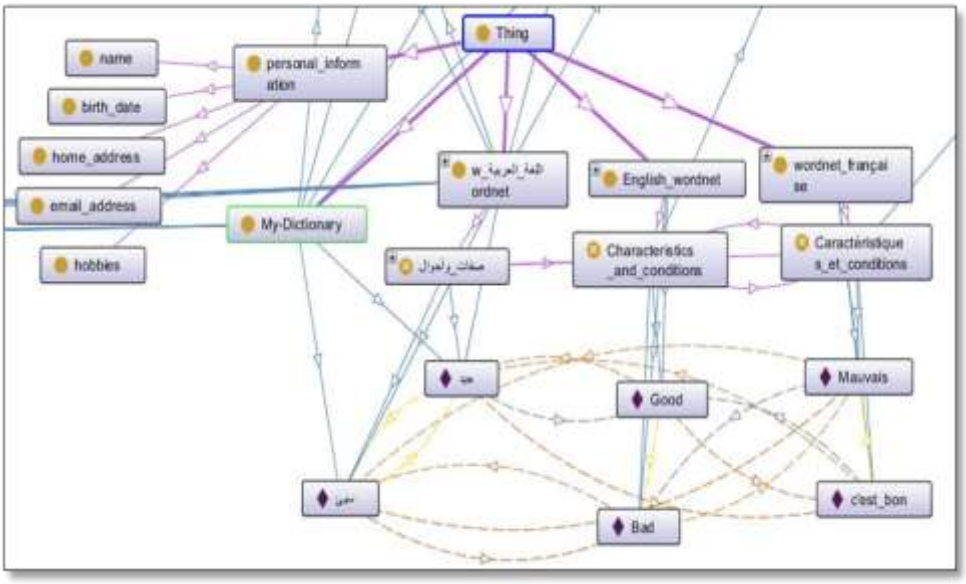


Figure 4. *Part of Proposed (MSLO)*

4. The Proposed Semantic with Deep Learning

Deep learning has been successfully applied in a variety of domains, such as image recognition and text mining. The link between ontologies and deep learning is being explored. (Petrucci et al,2016) addresses the extraction of OWL information from raw text with deep learning. (Hohenecker and Lukasiewicz, 2017) applies deep learning to Ontology extraction, obtaining encouraging results.

We propose semantic deep learning (SDL) technique, which refines the output of a deep learning technique. In our SDL technique ontology layer is inserted to add semantic information to deep learning output as shown in Figure 5. We need this semantic layer to enrich deep learning output with knowledge in Ontology. The "DL_label" name as data property value in ontology based so each class label has its corresponding data property value in ontology to retrieve all semantic data of this class.

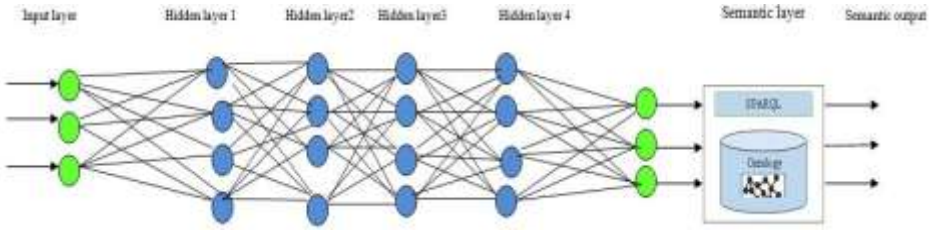


Figure 5. *Semantic with Deep Learning layers*

Generally, we used deep learning part in SDL as the following:

- Using a convolutional neural network (CNN) to identify and classify data into one of the predefined classes.
- Applying regularization techniques (data augmentation dropout, and batch normalization)

We start by Image augmentation technique to adapt the training set with such images. Data augmentation has a large impact on improving the DL accuracy by reducing overfitting and increase the generalize capacity of the network (Perez and Wang, 2017). Image data-augmentation increases the size of the training dataset by applying small geometric and photometric transformations.

Semantic and Deep Learning layers (SDL) are as the following:

- Convolutional layer composed of groups of neurons act as kernels. These kernels are associated with a small region of the image known as a receptive field, then convolving them (applying Convolution operation) with a specific set of weights. Where multiply each element of the filter (weights) with corresponding receptive field elements. Different sets of features extracted by sliding convolutional kernel on the image with the same set of weights. Convolution operation may categorize into different types based on the type and size of filters.
- Activation layer, also called non-linear layer, the convolution layer output is generally followed by a non-linear function called activation function to introduce non-linearity into the model. Thus, features generated by the convolution layer are transformed into another space by the activation function, and the data can be better classified. Here, we used (Rectified Linear Units) ReLU (Glorot et al., 2011) that applies the activation function max (0, x).
- Batch normalization (BN) layer, is the most widely used to enhance generalization (Ioffe and Szegedy, 2015) and avoid over. Also, is added after a convolutional layer and its ReLU activations. Batch Normalization achieves the same accuracy with fewer training steps thus accelerating the training process.
- Dropout layer also used to avoid over fitting. The key idea of Dropout (Srivastava et al., 2014) is to randomly drop units (along with their connections) during the training phase. The reduction in the number of parameters in each step of training has an effect of regularization.
- Pooling layer is used on one hand to reduce the spatial dimensions of the representation and to reduce the amount of computation done in the network. The most used pooling layer has filters of size 2 x 2 with a stride 2. This reduces the thoughts to a quarter of its original size. In our CNN max pooling is applied, which retains the maximum value within the local neighborhood of the sliding window.
- Fully connected layer, interpret these feature representations and perform high-level reasoning. Each neuron from a fully connected layer is linked to

each output of the previous layer. The operations behind a convolutional layer are the same as in a fully connected layer.

- Loss layer is used to penalize the network for deviating from the expected output. This is normally the last layer of the network. Various loss function exists: softmax is used for predicting a class from several different classes, where sigmoid cross-entropy is used to predict multiple independent probabilities (from the [0, 1] interval).
- Semantic layer, maps the class label to a formal representation of its meanings. When a set of words share the same fundamental meaning, they will have one DL_label.

As shown in Figure 6, CNN models are 3 groups of 2 convolutional layers each group followed by a max-pool layer with a filter of shape 2 x 2 with stride 2. Batch normalization and dropout layers are employed in all convolution layers. Fully-connected (FC) layer follows the convolutional layers. It has 1024 channels where is a soft-max layer to perform the classification over 10 classes. Dropout regularization and batch normalization are used also in the fully-connected layer. The dropout rates of 0.25 and 0.5 were set for the convolutional-pooling layers and the fully-connected layer, respectively.

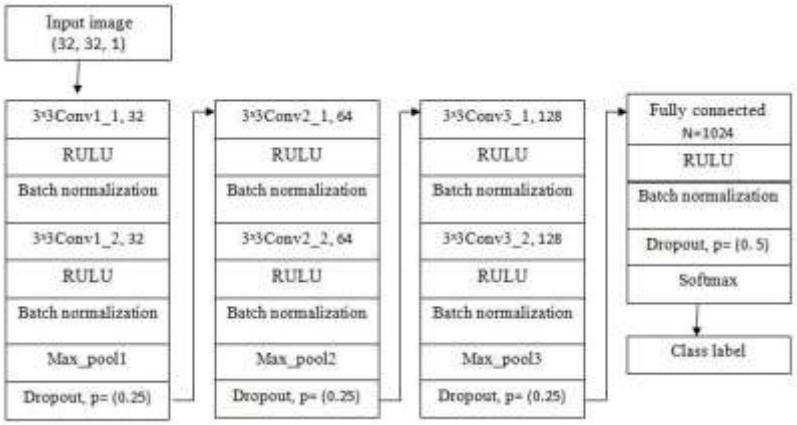


Figure 6.*The architecture of Deep Learning layers in SDL*

5. The Case Study on Arabic Sign Language Semantic Translation System

We apply our proposed system by Python using Keras (Chollet , 2015). We chose the backend TensorFlow (Google , 2015), where it is mainly used in classification. TensorFlow is an open-source software library for deep learning created and maintained by Google. We used Anaconda (Continuum Analytics) because it's flexibility to work with any operating system. Also, it has Spyder editor, a powerful python editor that has all the functionalities needed to write and run the codes. It also has the option to create an environment to install the packages. Also, we used Python OpenCV library for image processing, and view the results in the 'IPYTHON' console. The MSLO ontology created and merged with WordNet using Protégé (Protégé ,2000) as the Ontology editor.

5.1 Arabic Signs Dataset

All standard Arabic datasets based on letters and numbers. So, we collect samples of 10 Arabic words in Arabic sign language, their Arabic concepts, and their corresponding

English and French signs. Images are real, where we picked using a mobile camera. About 80% of the dataset used in the training model and 20%used in testing. A snapshot of a few images from the collected samples is given in Figure 7.

Figure 7: *Snapshot of the dataset*

Since our real dataset images were not captured in a controlled setting, it was especially prone to differences in light, skin color, and other differences in the environment.

5.2 Pre-process phase

The pre-processing phase consists of many main operations, which applied to the Arabic singe language images after it is taken from the camera. That is to enhance the image. We applied these operations on the training set, testing set and on any new images to be recognized. The pre-process operations are as the following:

- Cropping: cropped out the hand object and discarded all the other unnecessary elements in the image.
- Resizing: The images were in different dimensions; therefore, we resized each image to a common resolution of 32 × 32 pixels.
- Converting: All the images were converted into grayscale before feeding into the model.
- Denoising: Smoothing the images by convolving with a Gaussian filter and median blur to reduce noise in the images.

5.3 Training phase

The input that we used consists of standard RGB images of similar size 32 x 32 pixels. Then the dataset was augmented where the images are randomly rotated 0 to 30 degrees. Images were randomly sheared in a range of 0.2-degree, width and height shift in a range of 0.1 degree, zoomed in a range of 0.2 degree and images were horizontally flipped. We used 200 sign images of Arabic gestures in the training phase of the proposed CNN model. We used 88 sign images to measure the recognition rate of the model. After the model was trained the net parameters were saved. It can be used in the testing phase for testing the accuracy of the model and for recognition of the input sign.

Table 1 shows the classification accuracy on the training set for 100, 150 and 200 epochs. The table shows that the model has high train accuracy at 100 epochs.

Table 1. *Overall classification accuracy on the training Arabic signs dataset*

Epochs values	Accuracy
Epochs=40	95.29
Epochs=80	97.7
Epochs=100	98.66

Figure 8, shows training and validation loss/accuracy plot during training our dataset.

The model trained for 100 epochs and achieved low loss with limited over fitting. With additional training data, we could obtain higher accuracy as well.

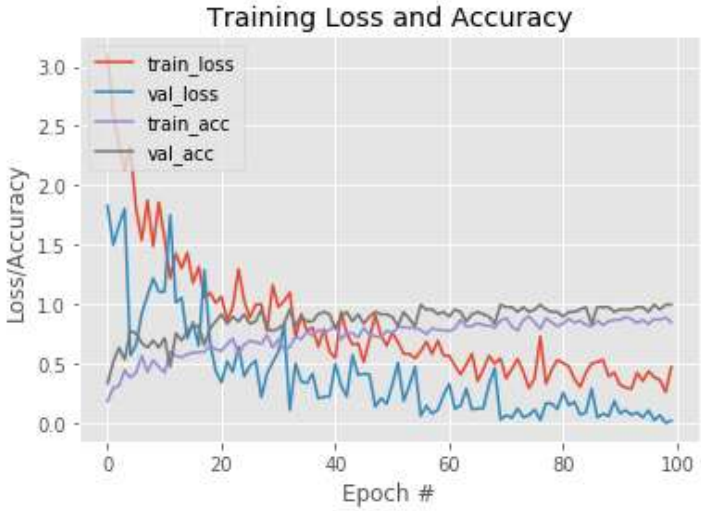


Figure 8. *Training and validation loss/accuracy plot during Arabic signs training*

Our network performed quite well when we used a premade numbers dataset (Barczak et al., 2011) to compare our dataset's performance, about 25% of samples used in testing the model. Training loss was 0.0446, Training accuracy was 99.0291%. Validation loss was 0.1210 and Validation accuracy was 97.8261%.

5.4 Testing phase

In the testing phase, the accuracy of our trained model is tested. The saved network parameter is loaded to test new signs. The input of the system is the image of the sign language static gestures that need to be recognized. The system used our trained CNN network to recognize these signs and produce their corresponding labels. The method used for all test cases and tabulated the obtained accuracies in section 6.

Also, the input of the system could be a text (Arabic, English or French word) which users type to select its respective sign.

5.5 Support Personality

There is a class in MSLO Ontology used as a user personal dictionary (My-Dictionary class). Deaf, dumb and normal users can develop and update their signs. That is solving the challenge of the individual differences between users.

5.6 Connection between DL and MSLO ontology for Semantic Translation

The output of deep learning represents the input of the semantic layer. In this layer the result of the previous layer used in searching in the Ontology to get all possible words that share the same fundamental meaning.

Perform searching on (MSLO) ontology using SPARQL to predict the corresponding meanings. SPARQL is used to access the RDFs inside the ontology by using SELECT statement. SPARQL query has a standard syntax and depends on using variables that contain the predicate, subject, and objects for RDF.

The process of retrieving semantic and displaying relevant images and meaning is based on the users' queries. SPARQL query used in searching ontology to get meaning that represents this word and gets the URL sign image of this meaning.

Some SPARQL queries used in searching and retrieving data from MSLO:

- SPARQL query to translate Arabic sign language using DL_label to its Arabic meanings.

```
SELECT ?z
WHERE { ?arb sign:المعني-باللغة-العربية ?z
FILTER (str(?arb) = """+lable+""")
bind( strafter(str(?arb),str(sign:)) as ?arb)}
```

- SPARQL query to translate Arabic meaning to its corresponding Arabic sign gesture.

```
SELECT ?arb ?o
WHERE {?arb rdfs:label ?x.
?arb sign:arabic_sign ?o
FILTER (regex(str(?x) , """+arabictext+"""))
bind( strafter(str(?arb),str(sign:)) as ?arb) }
```

- SPARQL query to translate Arabic sign to English sign gestures.

```
SELECT ?o  ?s ?z
WHERE {?o sign:label ?x.
?o sign:english_meaning ?s.
?s sign:english_sign ?z
FILTER (regex(str(?o) ,"""+label+"""))
bind( strafter(str(?s),str(sign:)) as ?s  ) }
```

6. Analysis of results

The performance of our system is evaluated by the rate of successful semantic sign language recognition. Without the use of semantic, it might be difficult to find meaningful connection among different languages words and synonyms words in one language. The recognition rate is defined by the ratio of the numbers of correct sign recognition to the total number of signs. The recognition rate in percentages in case of 10 Arabic signs is showed in Table 2.

Also, system performance is evaluated by the rate of successful semantic sign language recognition of new signs such that all possible meanings are listed. The semantic translation rate in percentages in case of our collected 10 Arabic signs is showed in Table 2.

Table 2. *Number of samples, Recognition Rate (%) and results per word*

Arabic words	DL_ Label	Training samples	Testing samples	Incorrect recognize	Recognition Rate (%)	Result with traditional DL	Results of proposed method SDL	
							Synonyms	Words in Same sign
سئ	Bad	16	5	0	100	Bad	الردِيء، 'السُّوء، السَّيِّء، الطالِح ،حرف الواو	
سيارة	Car	20	8	0	100	Car	سيّارة، سيْارة، عربة،حافلة، مرْكبة	
جيد	Good	30	22	1	95.45	Good	مُتْقَن ،جيد ،حسن	حرف الألف
مرحبا	Hello	19	7	0	100	Hello	مَرْحَباً ، مرحبا	الرقم خمسة
أحبك	I_love_you	17	9	0	100	I_love_you	اقبلك ،احبك	
الحب	Love	22	7	0	100	Love	حُبّ، وَجْد، قبول	الاعتذار
لا	No	20	9	2	77.7	No	لا ،نفي ،اعتراض	حرف الباء ،الرقم واحد
صورة	Photo	29	6	0	100	Photo	تِمْثال ، خَيَال ، شَبَه، شَبِيه ، صُورَة ، نموزج ، مَثِيل ، مَجَاز ، مِثَال ، نَظِير ،صورة فوتوغرافية ، تَصْوِيرَة	
شكرا	Thanks	11	8	1	85.7	Thanks	اِمْتِنان، عِرْفانُ بالإِحْسان، الثَّناءُ الجميلُ ، شكرا	
نعم	Yes	16	7	0	100	Yes	نعم ،موافقة ،قَبُول ،أجل ، بلى	
		200	88	4	95.45%			

Table 2 shows the number of training and testing samples for each word, Recognition Rate (%) of sign gestures, results with traditional deep learning model and result with adding a semantic layer.

Also, Table 2 shows the number of correctly recognized and misrecognized samples for each Arabic word and the corresponding accuracy. The results also show that while most of the words signs are recognized with 100% accuracy, it can be seen that the system recognition accuracy becomes less when gives the input sign for words "جيد"،"شكرا"and "لا".

Overall, out of the 88 Arabic sign images used for testing 84 images were recognized into correct meanings and the remaining 4 images were misrecognized resulting. The recognition accuracy was checked as per the correctness of testing gestures made was 95.45 %.

The result obtained when recognize using deep learning model was only the DL_label that represents the word, but while recognize using deep learning and Ontology the result was all available meanings of this sign gesture.

7. Conclusion and future works

In this paper, we propose a semantic with the deep learning method that bridges the gap between Ontology and deep learning thus taking advantages of both methods. We proposed Multi-Sign Language Ontology (MSLO) first version to solve some sign language challenges. It extends the WordNet relations to deal with sign language. Some refinements to enhance WordNet relations in MSLO were added to be compatible with multiple sign languages.

The proposed method was applied on a semantic translation system for translating Arabic static sign language to its meanings as text and sign language meaning in another language. We evaluate the system based on the rate of successful recognition. We used CNN trained model in the recognition process and adding the semantic layer.

Secondly, to check the feasibility of the proposed system, we collect and used signs of ten Arabic gestures and their meanings in English and French sign languages. Best results were obtained. Also, we compared the results when using deep learning only and when using semantic with deep learning models. The SDL presents all available meanings of this sign gesture.

As future work, we will enhance MSLO to detect the correct meaning, and the SDL to cover dynamic real video in real-time applications. We will convert the system to an Android application to produce a free long-life mobile application for deaf-mute people.

References

Al-Fityani, K., & Padden, C. (2010). Sign language geography in the Arab world. Sign languages: A Cambridge survey, 433-450.

Almasoud, A. M., & Al-Khalifa, H. S. (2012). Semsignwriting: A proposed semantic system for Arabic text-to-signwriting translation. Journal of Software Engineering and Applications, 5(08), 604.

Barczak, A. L. C., Reyes, N. H., Abastillas, M., Piccio, A., & Susnjak, T. (2011). A new 2D static hand gesture colour image dataset for ASL gestures.

Chollet, F.: Keras, (2015). https://keras.io/, [Online; accessed 15-october -2019].

Continuum Analytics, Anaconda Python. http://store.continuum.io/cshop/anaconda/.

ElBadawy, M., Elons, A. S., Shedeed, H. A., & Tolba, M. F. (2017, December). Arabic sign language recognition with 3d convolutional neural networks. In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 66-71). IEEE.

Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323).

Google.: TensorFlow,(2015). https://www.tensorflow.org/, [Online; accessed 15-october -2019].

Hohenecker, P., & Lukasiewicz, T. (2017). Deep learning for ontology reasoning. arXiv preprint arXiv:1705.10342.

Ibrahim, N. B., Selim, M. M., & Zayed, H. H. (2018). An automatic arabic sign language recognition system (arslrs). Journal of King Saud University-Computer and Information Sciences, 30(4), 470-477.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Latif, G., Mohammad, N., Alghazo, J., AlKhalaf, R., & AlKhalaf, R. (2019). ArASL: Arabic Alphabets Sign Language Dataset. Data in Brief, 23, 103777.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Lesmo, L., Mazzei, A., & Radicioni, D. (2011). Linguistic processing in the atlas project. Proceedings of SLTAT.

Lozynska, O., & Davydov, M. (2015). Information technology for Ukrainian Sign Language translation based on ontologies. ECONTECHMOD: an international quarterly journal on economics of technology and modelling processes, 4(2), 13-18.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis. Recognit.

Petrucci, G., Ghidini, C., & Rospocher, M. (2016, November). Ontology learning in the deep. In European Knowledge Acquisition Workshop (pp. 480-495). Springer, Cham.

Protégé (ed.), (2000). http://protege.stanford.edu/.

Reda, M. M., Mohammed, N. G., & Seoud, R. A. A. A. A. (2018, April). SVBiComm: Sign-Voice Bidirectional Communication System for Normal,"Deaf/Dumb" and Blind People based on Machine Learning. In 2018 1st International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-8). IEEE.

Shanableh, T., Assaleh, K., & Al-Rousan, M. (2007). Spatio-temporal feature-extraction techniques for isolated gesture recognition in Arabic sign language. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 37(3), 641-650.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Zhang, X., Hu, B., Ma, X., Moore, P., & Chen, J. (2014). Ontology driven decision support for the diagnosis of mild cognitive impairment. Computer methods and programs in biomedicine, 113(3), 781-791.

A Flexible hybrid Genetic-Based Fuzzy Rough Decision Model for accuracy enhancement

Mohamed.S.S.Basyoni **Ahmed M. Gadallah** **Hesham A. Hefney**

Faculty of Graduate Studies for Statistical Research

{cs887, ahmgad10} @yahoo.com, hehefny@ieee.com

Abstract: In this paper, we introduce a new more powerful hybrid algorithm which integrates the advantages of rough set theory and fuzzy set theory together with Genetic Algorithms (GAs). Our Genetic-Based Fuzzy Rough Decision Model algorithm consists of four phases: (1) automatic attributes fuzzification, (2) Eliminate redundant attributes using rough set theory, (3) Generating Fuzzy rough rules then calculate automatically the accuracy and a fitness value (Confidence) for each rule and , (4) Using the genetic algorithm for the Fuzzy rough rules to enhance their accuracy. In phase one, the user input the number of fuzzy sets of each attributes, our algorithm determine the maximum and minimum values of each attribute then calculates automatically the width (Δ) which divides the universe of discourse of each attribute into "n" intervals according to the number of fuzzy sets, also the algorithm calculates automatically the width (δ_i) according to the width (Δ). In phase two, we use the rough set techniques to reduce the number of attributes that comes from phase one and produce fuzzy-rough rules. In phase three, the algorithm calculates the accuracy and the confidence (fitness value) of each fuzzy rough rule then calculates the total accuracy of all linguistic rules. In phase four, we run the genetic algorithm on the fuzzy-rough rules from phase three then the algorithm calculates the accuracy and the confidence of each fuzzy rough rule again and calculates the total the accuracy of all rules. The accuracy of our algorithm that applied on Iris plants dataset before using our genetic algorithm from randomly 75 rows from 150 rows is 0.56 but after using our algorithm will be 0.95.

Keywords: Genetic algorithm, Fuzzy logic, Rough set, Accuracy, Automated Fuzzy - Based Rough Decision model.

1. Introduction

Fuzzy rough decision models have been appeared in various recent researches due to its efficacy in generating potential decision rules especially in the cases of incomplete quantitative data sets. Genetic algorithms (GAs) have been proposed by various researches for tuning the rules obtained by such decision making models. However, the accuracy and simplicity of such models remains a research problem of such models that need further research.

Rough sets theory was developed by Pawlak in the early 1980's [16, 18, 25, 26] and has been applied successfully in a lot of domains. One of the major limitations of the traditional rough sets model is that it assumes that all attribute values are discrete. A real world data set always contains

mixed types of data such as continuous valued, symbolic data, etc. Therefore all numerical or continuous data should convert to discretized data, here "Fuzzy Logic" can solve this problem to reduce information overload in a Fuzzy-Based Rough Decision Model [5, 7, 9, 11, 15, 19]. One drawback of traditional Fuzzy-Based Rough Decision Model is that the linguistic values (fuzzy sets) for numeric values of each attribute should determining by the membership functions of these linguistic terms which the user should define the parameters of those membership functions from his view which is different from one user to another. Therefore, we introduce a new automated Fuzzy-Based Rough Decision Model algorithm that can define those parameters automatically which the user determine only the number of fuzzy sets then the algorithm automatically determine the maximum and minimum values of each attribute and calculates automatically the width (Δ) that divides the universe of discourse of each attribute into "n" intervals according to the number of fuzzy sets then the algorithm calculates automatically the width (δi) according to the width (Δ). Another drawback of the traditional rough sets model in the real applications is the inefficiency in the computation of core and reduct, because all the intensive computational operations are performed in flat files [1, 4, 14, 23]. In order to improve the efficiency of computing core attributes and reducts, a New Rough Sets Model Based on Database Systems has been introduced for this purpose [Hu, X., Lin, T., 2004], which redefine the core attributes and reducts based on relational algebra to take advantages of the very efficient set-oriented database operations, such as *Cardinality* to denote the *count* and *Projection*.

The paper is organized as follows: We give an overview of the genetic algorithm in section 2, and give an overview of the rough set theory based on the model proposed by Pawlak [25, 26] with some examples, also give an overview of the fuzzy set theory. In section 3, we propose a new automated Fuzzy -Based Rough algorithm that can define the parameters of membership functions of linguistic values automatically. After that, we generate fuzzy rules. Finally, we use the genetic algorithm on the fuzzy rules that generate other efficient fuzzy rules in accuracy. In the same section, we explain the contributions of our model. Finally, we conclude with some discussions and our future works in Section 4 and section 5.

2. Basic Concepts

2.1. Genetic algorithm (GA)

Genetic algorithms are used as a machine learning tool for generating (evolving) a rule-based classification system [33]. A genetic algorithm randomly generates a pre-specified number of fuzzy rules [34, 35], which each rule set can be represented as a bit string. Thus a population of individuals corresponds to a single rule set. GA has a list of parameters that are included in the GA adaptation file as inputs and outputs parameters. The first input parameter of GA is fitness value, which be evaluated for any population of individuals. The fitness parameter assigned the value of the highest fitness value of any population member from the most recent generation. The GA uses two basic operations which are the mutation and the crossover. After implementing this stage it will yield a list of rules are generating [36, 37].

2.1.1. The Basic algorithm:

Initialize population Pi randomly
For i=1 to max generations
Evaluate fitness of individuals of population Pi
Repeat

Select parents from Pi for reproduction
Perform crossover on parents creating Pi+1
Perform mutation of Pi+1
Evaluate fitness of individuals of population Pi+1
Until best individual is good enough

2.1.2. Chromosome ("individual") representation

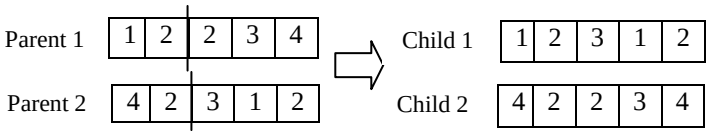
Each fuzzy set represented by integer number forming a gene on chromosome (individual) of fixed length that forms a population of fixed number of chromosomes

2.1.3. Selection Methods

The essential idea of a selection is that select the best individuals (candidate solutions) of the population to breed a new generation. The selection of the best individuals depends on their fitness

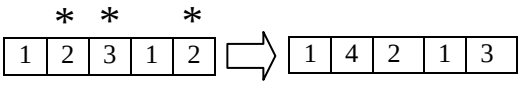
2.1.4. Crossover (recombination) operator

The crossover (or recombination) operator essentially swaps genetic material between two selected individuals from the population, called parent individuals



2.1.5. Mutation (bits flipped) operator

Mutation is an operator that acts on a single individual at a time. It usually applied with a small probability, typically much smaller than the crossover probability. Unlike crossover, which recombine genetic material between two or more parents, mutation replace (or flip) the value of a gene (a bit) with a randomly-generated value



2.1.6. Termination

Evaluate fitness of individuals of population, and then the above generational operators are repeated until a fitness condition has been reached.

2.2. Rough set theory (RST)

Rough set theory was developed by Pawlak in the early 1980's [8, 16, 18, 26] and has been successfully applied to knowledge acquisition as a powerful tool for data mining, decision analysis and forecast, knowledge discovery from database, and decision support system. The starting point of rough set is a dataset, which is usually organized into a table, and it is called information systems. In rough sets theory, the data is collected in a table, called decision table (or table in the database term) as shown in Table 1. We also assume that our data set is stored in a relational table with the form *Table (condition-attributes, decision-attributes)*. C is used to denote the condition attributes, D for decision attributes, where, $C \cap D = \Phi$. The main idea of rough set theory is based on the indiscernibility relation $[x]R$ (equivalence classes).

Based on the knowledge used in approximation, the universe U is divided into some element sets of the indiscernible objects that are based on conditional attribute sets, and the element sets are regarded as approximate "knowledge granularity" the knowledge used in approximation would divide the universe into the decision-making class derived from the decision attribute sets, on this basis, and one kind of knowledge can be used to approximate the other. Rough sets theory defines three regions based on the equivalent classes induced by the attribute values: lower approximation, upper approximation and boundary.

Table 1: Cars table with attributes *Door, Size, Cylinder* and *Mileage*

Tuple-id	***Door***	***Size***	***Cylinder***	***Mileage***
*t*1	2	compact	4	high
*t*2	4	sub	6	low
*t*3	4	compact	4	high
*t*4	2	compact	6	low
*t*5	4	compact	4	low
*t*6	4	compact	4	high
t7	4	Sub	6	low
t8	2	sub	6	low

Definition 2.2.1 Let *T = (U, A)* be an information table with any set $B \subseteq A$ and $X \subseteq U$ We can approximate *X* using only the information contained in *B* by constructing the *R-lower* approximations of *X* that contain all objects in equivalence classes that "surely belong to X" denoted by $\underline{R}X$ where:

$$\underline{R}X = \{x \mid [x]_R \subseteq X\}$$

Definition 2.2.2 Let *T = (U, A)* be an information table with any set $B \subseteq A$ and $X \subseteq U$ We can approximate *X* using only the information contained in *B* by constructing the *R-upper* approximations of *X* that contain all objects in equivalence classes that "may belong to X" denoted by $\overline{R}X$ where:

$$\overline{R}X = \{x \mid [x]_R \cap X \neq \phi\}$$

Definition 2.2.3 The boundary area is the difference between the upper approximation and the lower approximation where:

$$BND(X) = \overline{R}X - \underline{R}X$$

The set X is said to be *rough (inexact)* if its *boundary region* is non-empty, otherwise the set is *crisp or exact* [16, 18]

Example 2.2.1

In table 1, as a dataset in (X.Hu, T.Lin, and J.Han.2004) [6] we have a collection of 8 cars (*t*1,*t*2,…*t*8) with information about the attributes *Door*, *Size*, *Cylinder* and *Mileage*, where *Door*, *Size* and *Cylinder* are the condition attributes and *Mileage* is the decision attribute. The attribute

Tuple-id is just for explanation purpose and can be ignored. We can calculate the elementary sets, lower and upper approximations as follows:
The indiscernibility Relation of *C*:
IND (*R*) = *IND* ({*Door, Size, Cylinder*}) So, *R* = {*X1, X2, X3, X4, X5*} where,
[*X*]*R* = {{*t1*}, {*t2, t7*}, {*t3, t5, t6*}, {*t4*}, {*t8*}}

Let *D1* = {*x* | *Mileage* (*x*) = high}, then: *Y1* = {*t1, t3, t6*}
Lower approximation ={*X1*} = {*t1*}, *Upper* approximation = {*X1* ∪ *X3*} = {*t1, t3, t5, t6*}
Let *D2* = {*x* | *Mileage* (*x*) = low}, then: *Y2* = {*t2, t4, t5, t7, t8*}
Lower approximation = {*X2* ∪ *X4* ∪ X5} = {*t2, t4, t7, t8*}, *Upper* approximation = {*X2* ∪ *X2* ∪ *X3*} = {*t2, t3, t4, t5 t6, t7, t8*}

Boundary *C*/*D* = {*t3, t5, t6*}

This table is called "inconsistent table" that contains entities with the same condition attribute values but with different decision attribute values. Therefore, only 5 of 8 cars *t1*, *t2*, *t4*, *t7 t8* belong to the lower approximation of *D* based on *C*, while 3 of 8 cars fall in the boundary area. This fact indicates that the information of *Door*, *Size* and *Cylinder* collected so far is not consistent, for it is only good enough to make a classification model for the above five cars, but not enough to classify other three. In order to classify *t3*, *t5* and *t6*, more information is needed.

Example 2.2.2

In table 2, a new condition attribute, which is a *Weight* of cars, is added and collect information on for each car [6]. To classify the tuples in the boundary area, with the condition attributes set *C* = {*Weight, Door, Size, Cylinder*}, we can calculate the lower, upper approximations and boundary area as below:

Table 2: Cars table with attributes *Weight*, *Door, Size, Cylinder* and *Mileage*

Tuple-id	***Weight***	***Door***	***Size***	***Cylinder***	***Mileage***
*t*1	low	2	compact	4	high
*t*2	low	4	sub	6	low
*t*3	medium	4	compact	4	high
*t*4	high	2	compact	6	low
*t*5	high	4	compact	4	low
*t*6	low	4	compact	4	high
t7	high	4	Sub	6	low
t8	low	2	sub	6	low

The indiscernibility Relation of *C*:
IND (*R*) = *IND* ({*Weight, Door, Size, Cylinder*}), *R* = {*X1, X2, X3, X4, X5, X6, X7, X8*} where,
[*X*]*R* = {{*t1*}, {*t2*}, {*t3*}, {*t4*}, {*t5*}, {*t6*}, {*t7*}, {*t8*}}
Let *D1* = {*x* | *Mileage* (*x*) = high}, then: *Y1* = {*t1, t3, t6*}
Lower approximation = {*t1, t3, t6*},

Upper approximation = *{t1, t3, t6}*
Let *D2* = {*x* | *Mileage* (*x*) = low}, then: *Y2* = {*t2, t4, t5, t7, t8*}
Lower approximation = {*t2, t4, t5, t7, t8*},
Upper approximation = *{t2, t4, t5, t7, t8}*

Boundary *C*/*D* = ϕ
From above, one can see, this table is "consistent table".

2.3. Fuzzy Logic (FL)

Fuzzy logic was developed by Lotfi A. Zadah in the 1960's [15, 17, 19, 30], which is an extension of conventional Boolean logic constructed to deal with imprecise information i.e., statements that are neither completely true nor completely false. Fuzzy logic works through the use of fuzzy sets (FS). A fuzzy set is a set whose boundaries are not clearly defined. We can define a subset A of a given universal set U (the universe of discourse) as containing all elements x of U for which the element x is member of a fuzzy set A. Each element belongs to a set A to a certain degree, called the degree of membership μA(x). Typically represented by a real-valued number in the interval [0...1]. This is contrast with conventional, crisp sets, whose boundaries are clearly defined. Each element either belongs or does not belong to a crisp set. So in term of degree of membership, we can say that an object can belong to a crisp set with only one of two possible degree of membership: 0 or 1. Fuzzy logic can provide a great way to convert numerical and real-valued data into categorical data by using linguistic values for numeric values and determining of membership functions of these linguistic terms.
Linguistic Variables (terms) can be defined over the domain of each attribute based on characteristics of that domain which take words as values like "hot, cold, small, medium, etc"
Each Linguistic Variable has a range of states called linguistic values, each of which is a fuzzy (linguistic) set defined over the same domain represented by its membership function this domain is called the *universe of discourse* of that linguistic variable.

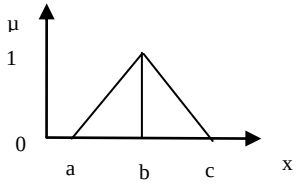
3. Proposed model

3.1. Phase 1: Automated Attributes Fuzzification

Fuzzy logic can provide a great way to convert numerical and real-valued data into categorical data by using linguistic values for numeric values and determining of membership functions of these linguistic terms which, every element in the universe of discourse is a member of the fuzzy set with some grade (degree of membership functions)
The traditional Fuzzy-Based Rough Decision Model has a major limitation that linguistic values (fuzzy sets) for numeric values of each attribute should determining by the membership functions of these linguistic terms which the user should define the parameters of membership functions of these linguistic values from his view which is different from one user to another. Therefore, we propose a new automated Fuzzy-Based Rough Decision Model algorithm that can define the parameters of membership functions of these linguistic values automatically that the user determine only the number of fuzzy sets (linguistic values) then the maximum and minimum values of each attribute are determined automatically then the algorithm calculates the width (Δ) that divides the universe of discourse "u" of each attribute into "n" intervals according to the number of fuzzy sets then the algorithm calculates automatically the width (δi) according to the width (Δ).

3.1.1. Triangular Membership function

$$\text{Triangular (X: a, b, c)} = \begin{cases} 0 & x \le a \\ (x - a) / (b - a) & a \le x \le b \\ (c - x) / (c - b) & b \le x \le c \\ 0 & x > c \end{cases}$$



For "n" linguistic values (fuzzy sets) then the universe of discourse "*U*" which is the attribute divided into "n" intervals with the width between center bi and bi+1 (Δ) where:

$$\Delta = \frac{Xmax - Xmin}{n - 1} \quad \text{where,} \quad 1 \le x \le n+1 \quad \text{with:} \quad Xmin = Xmin - D1, \quad Xmax = Xmax + D2, \quad \text{and} \quad \delta i = (\Delta+1) / 2$$

We can also calculate the parameters a, b and c by the following equations:

Ai = (ai, bi, ci)

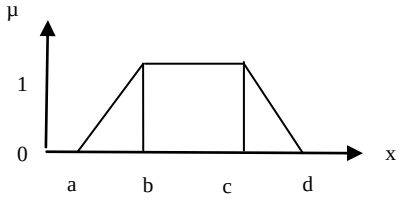
$bi = Xmin + (i - 1) * \Delta$, $ai = bi - \delta i$, $ci = bi + \delta i$

Where:

Δ: the width between center bi and bi+1 — Xmax: the maximum value of the attribute
Xmin: the minimum value of the attribute — n: the number of linguistic values (fuzzy sets)
D1: the value subtracted from Xmin to make it integer value — D2: the value added to Xmax to make it integer value
δi: the width between bi and ai, ci

3.1.2. Trapezoidal Membership function

$$\text{Trapezoidal (X: a, b, c, d)} = \begin{cases} 0 & x \le a \\ (x - a) / (b - a) & a \le x \le b \\ 1 & b \le x \le c \\ (d - x) / (d - c) & c \le x \le d \\ 0 & x > d \end{cases}$$



For "n" linguistic values (fuzzy sets) then the universe of discourse "u" which is the attribute divided into "n" intervals with the width (Δ) where:

$$\Delta = \frac{Xmax - Xmin}{2n - 1} \quad \text{where,} \quad 1 \le x \le n+1 \quad \text{with:} \quad Xmin = Xmin - D1, \quad Xmax = Xmax + D2, \quad \text{and} \quad \delta i = (\Delta+1) / 2$$

We can also calculate the parameters a, b, c and d by the following equations:

Ai = (ai, bi, ci, di)

$bi = Xmin + 2 (i - 1) * \Delta$, $ai = bi - \delta i$, $ci = bi + \Delta$, $di = ci + \delta i$ Where:

Δ: the width between bi and ci, and between ci and bi+1 — Xmax: the maximum value of the attribute
Xmin: the minimum value of the attribute — n: the number of linguistic values (fuzzy sets)
D1: the value subtracted from Xmin to make it integer value — D2: the value added to Xmax to make it integer value
δi: the width between bi and ai, and between ci and di

Example 3.1

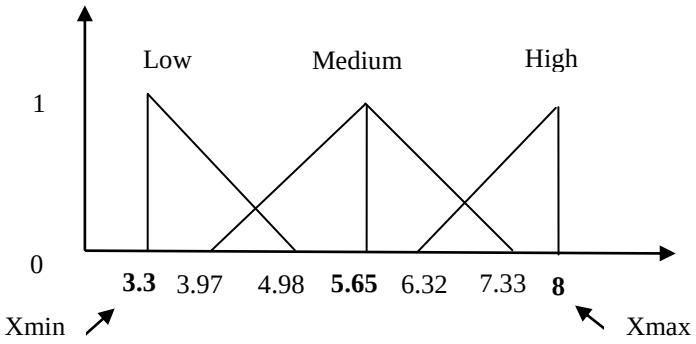
Suppose we have the Iris plants dataset (150 records), with information about the attributes *Sepal length (S_L)* in cm, *Sepal width (S_W)* in cm, *Petal length (P_L)* in cm, *Petal width (P_W)* in cm and *Class (RS)* which contain three classes: iris *Setosa*, iris *Versicolour* and iris *Virginica* , where *Sepal length, Sepal width, Petal length* and *Petal width* are the condition attributes and *Class* is the decision attribute.

Suppose we choose "Triangular Membership function" to linguistic terms and define "3" fuzzy sets:

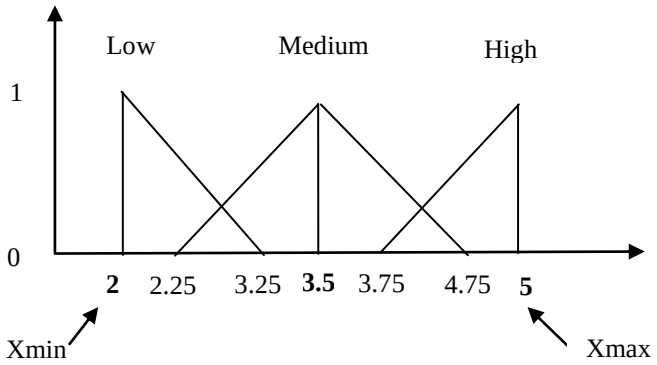
Low, Medium and High. Therefore, to fuzzifing numerical data of attribute *Sepal length (S_L)*, then:

Xmin = 4.3 and Xmax = 7.9 So, D1 = 1 and D2 = 0.1 then:
Xmin = 3.3 and Xmax = 8

The width between bi and bi+1 is:
Δ = 8-3.3/ (3-1) = 2.35
Then, δi = (2.35+1) / 2 = 1.68, therefore:
1- Fuzzy set A1 ⟶ Low, $\delta 1$ =1.68
b1 = a1 = 3.3, c1 = 4.98
2- Fuzzy set A2 ⟶ Medium, $\delta 2$ =1.68
b2 = 5.65, a2 = 3.97, c2 = 7.33
3- Fuzzy set A3 ⟶ High, $\delta 3$ = 1.68
b3 = c3 = 8, a3 = 6.32



Suppose we also choose "Triangular Membership function" to linguistic terms and define "3" fuzzy sets: Low, Medium and High. Therefore, to fuzzifing numerical data of attribute *Sepal width (S_W)*, then:
Xmin = 2.2 and Xmax = 4.4 So, D1 = 0.2 and D2 = 0.6 then:
Xmin = 2 and Xmax = 5
The width between bi and bi+1 is:
Δ = 5-2/ (3-1) = 1.5
Then, δi = (1.5+1) / 2 = 1.25, therefore:
1- Fuzzy set A1 ⟶ Low, $\delta 1$ =1.25
b1 = a1 = 2, c1 = 3.25
2- Fuzzy set A2 ⟶ Medium, $\delta 2$ =1.25
b2 = 3.5, a2 = 2.25, c2 = 4.75
3- Fuzzy set A3 ⟶ High, $\delta 3$ = 1.25
b3 = c3 = 5, a3 = 3.75



Suppose we choose also choose "Triangular Membership function" and define "3" fuzzy sets: Low, Medium and High. Therefore, the algorithm fuzzifing numerical data of two attributes *Petal length (P_L), Petal width (P_W)* in the same manner.
Thus, after fuzzifing numerical data of the all four attributes, the result will be as shown in table 3.

Table 3: Iris data table after fuzzifing the four numerical attributes

Sno	*S_L*	*S_W*	*P_L*	*P_W*	*RS*
1	medium	medium	low	low	setosa
2	medium	medium	low	low	setosa
3	medium	medium	low	low	setosa
4	medium	medium	low	low	setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	high	medium	medium	medium	versicolor
52	medium	medium	medium	medium	versicolor
53	high	medium	medium	medium	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
101	medium	medium	high	high	verginica
102	medium	low	medium	medium	verginica
103	high	medium	high	medium	verginica
⋮	⋮	⋮	⋮	⋮	⋮
150	medium	medium	medium	medium	verginica

Contribution 3.1

Before using our automated Fuzzy-Based Rough Decision Model algorithm on Iris plants dataset, the user should define the parameters of membership functions of these linguistic values from his view which is different from one user to another [11, 21, 31, 32], but by using the proposed automated algorithm the parameters of membership functions of these linguistic values defined automatically that the user determine only the number of fuzzy sets (linguistic values) which is three fuzzy sets in the example then the maximum and minimum values of each attribute are determined automatically then the algorithm calculates the width (Δ) that divides the universe of discourse "u" of each attribute into "n" intervals according to the number of fuzzy sets then the algorithm calculates automatically the width (δi) according to the width (Δ).

3.2. Phase 2: Eliminate Redundant Attributes using Rough Set Theory

There exist some limitations of traditional rough sets theory which restricts its suitability in practice [8, 14, 27, 28], one of which is the inefficiency in computation, which limits its suitability for large data sets in real-world applications. In order to find the reducts, core and dispensable attributes, the rough sets model needs to construct all the equivalent classes based on the attribute values of the condition and decision attributes. This process is very time-consuming, and thus the model is very inefficient and infeasible, and doesn't scale for large data set, which is very common in data mining applications [8, 14]. Further investigation of the problem reveals that most existing rough set models [1, 14, 31, 32] do not integrate with the relational database systems and a lot of computational intensive operations are performed in flat files rather than utilizing the high performance database set operations.

To overcome this problem, a New Rough Sets Model Based on Database Systems has been introduced [6, 14] for this purpose to redefine some concepts of rough set theory such as core attributes and reducts by using relational algebra so that the computation of core attributes and reducts can be performed with very efficient set-oriented database operations, such as the following relational algebra: *Cardinality* (*Card*) to denote the *Count*, and Π for *Projection* operation.

Definition 3.2.1. An attribute $Cj \in C$ is a *core* attribute [6] if it satisfies the following condition:

Card (Π (*C – Cj + D)*) ≠ Card (Π (*C – Cj)*) Where:

Card: the *cardinality* to denote the *count* of attribute, Π: *Projection* operation.

For example, in Table 2, it can be shown that

Card (Π (*C – Cj + D)*) =*Card* (Π (*Door, Size, Cylinder, Mileage*) = 6, and *Card* (Π (*C – Cj)*) = *Card* (Π (*Door, Size, Cylinder*) = 5

Therefore, the attribute *Weight* is a *core* attribute in *C* with respect to attribute *Mileage.*

Definition 3.2.2. An attribute *Cj* ∈ *C* is a *dispensable* attribute [6] in *C* with respect to *D*, if the classification result of each tuple is not affected without using *Cj*, that is,

Card (Π (*C – Cj + D)*) = Card (Π (*C – Cj)*)

For example, in Table 2, it can be shown that

Card (Π (*C – Cj + D)*) = *Card* (Π (*Weight, Size, Cylinder, Mileage*) = 6, and

Card (Π (*C – Cj)*) = *Card* (Π (*Weight, Size, Cylinder*) = 6

Thus, *Door* is a dispensable attribute in *C* with respect to attribute *Mileage*.

3.3. Phase 3: Generating Fuzzy Rough Rules and Computation of Accuracy and Confidence values

3.3.1. Generating Fuzzy Rough Rules

After Fuzzifing original information system and determining the reduct attributes we can get decision rules which help decision maker to take the proper decision. We use a new algorithm for extracting fuzzy rough rules from fuzzy table using SQL statements as following:

1- Create temp table

```
SELECT    CURR_REDUCT, D
INTO      TMP_TBL
FROM      T
```

2- Get equivalence classes for current reduct

```
SELECT    CURR_REDUCT
FROM      TMP_TBL
GROUP BY  CURR_REDUCT
```

3- Get decision rule for each equivalence classes

```
SELECT    DISTINCT  D
FROM      TMP_TBL
WHERE     X1 = Y1  AND  X2 = Y2  AND …. Xn = Yn
```

We choose a number of records that the algorithm *randomly* generates the fuzzy rules from them. So that, the remaining records will be the test records that calculate the accuracy of fuzzy rules from

them. After that the algorithm generates a list of Reducts for table 3, we can easily select *Class (RS)* attribute as a decision attribute then the reduct attributes in table 3 will be same four condition attributes *S_L*, *S_W*, *P_L* and *P_W*.
Suppose we choose randomly 75 rows from 150 rows then the algorithm generates 12 fuzzy rules from those records using triangular membership functions as following:

S_L is 'High' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor
S_L is 'Low' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'High' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Low' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor OR virginica
S_L is 'Medium' And S_W is 'Low' And P_L is 'Medium' And P_W is 'Medium' ==> RS is virginica

3.3.2. Confidence (fitness) and support of each rule

In the field of data mining, two measures are often used to evaluate association rules, which are Confidence and Support

1- Confidence $(Aq \rightarrow Cq) = \frac{|D(Aq) \cap D(Cq)|}{|D(Aq)|} = \frac{\sum \mu Aq(x) \cap \mu Cq(x)}{\sum \mu Aq(x)}$

2- Support $(Aq \rightarrow Cq) = \frac{|D(Aq) \cap D(Cq)|}{|D|} = \frac{\sum \mu Aq(x) \cap \mu Cq(x)}{\text{No. of all tuples}}$ Where,

Aq: the antecedent part of the rule
Cq: the consequent part of the rule
$|D(Aq)| = \Sigma \mu Aq(x)$: the cardinality of a fuzzy set
|D|: no of all patterns

For the pervious fuzzy rough rules that generated from table 3, the algorithm will automatically calculate the confidence and the accuracy of each fuzzy rough rule which the confidence calculated using the Average operator that takes the average confidence value of the fuzzy sets of each rule then it calculates the total accuracy of all rules as the following results in the table 4.

Table 4: *Confidence* and *Accuracy* of each Fuzzy rough Rule before running Genetic algorithm

Fuzzy rules	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Total	Accuracy %
frequency	1	9	1	2	3	2	1	5	28	1	15	7	**75**	
Confidence	0.82	0.6	0.76	0.66	0.84	0.83	0.63	0.71	0.7	0.66	0.55	0.58		
Accuracy	2	0	0	0	1	0	1	1	15	2	19	1	**42**	**0.56**

3.4. Phase 4: Using the genetic algorithm for the Fuzzy rough rules

The last phase is running the Genetic algorithm on the pervious fuzzy rough rules. We represent the three Fuzzy sets in each rule as a one chromosome with a number from 0 to 9. Suppose we

represent each fuzzy set of the pervious fuzzy rough rules by integer number forming a gene on chromosome (individual) of fixed length that forms a population of fixed number of chromosomes as following:

1: Low, 2: Medium, 3: High, 0: Do not care

After that, we set the accuracy value condition that we need to reach. Suppose we set the accuracy value to be 95%

Then the algorithm generates another new 12 fuzzy rules using triangular membership functions as following:

S_L is 'High' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is ' High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor
S_L is 'Low' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'High' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Low' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor OR virginica
S_L is 'Medium' And S_W is 'Low' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor OR virginica

- After that, the accuracy of those new 12 rules from randomly 75 rows from 150 rows will automatically calculated to be 0.95 instead of 0.56
- Suppose we choose randomly 100 rows from 150 rows, then the accuracy of those new 13 rules will automatically calculate to be 0.96 instead of 0.68
- Suppose we choose randomly 130 rows from 150 rows, then the accuracy of those new 14 rules will automatically calculate to be 0.95 instead of 0.7

The results of the three cases of Iris Dataset after running Genetic will be as following in the table 5:

Table 5: *Confidence* and *Accuracy* after running Genetic algorithm of three cases of Iris Data

	Fuzzy rules	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	Total	Accuracy %
Case 1 75 rows	**frequency**	1	***7***	1	2	3	2	1	5	28	1	15	***9***			***75***	
	Confidence	0.82	***0.71***	0.76	0.66	0.84	0.83	0.63	0.71	0.7	0.66	0.55	***0.58***				
	Accuracy	2	***6***	0	0	1	0	1	1	15	2	19	***24***			**71**	**0.95**
Case 2 100 rows	**frequency**	1	5	1	2	3	2	1	***19***	5	4	28	1	***28***		***100***	
	Confidence	0.82	0.7	0.8	0.7	0.84	0.8	0.63	***0.58***	0.71	0.6	0.7	0.66	***0.55***			
	Accuracy	2	5	0	0	1	0	1	***10***	1	5	15	2	***6***		**48**	**0.96**
Case 3 130 rows	**frequency**	1	3	8	1	2	2	2	2	***23***	6	8	41	3	***28***	***130***	
	Confidence	0.85	0.82	0.71	0.76	0.66	0.84	0.83	0.63	***0.58***	0.71	0.6	0.7	0.66	***0.55***		
	Accuracy	0	0	2	0	0	2	0	0	***6***	0	1	2	0	***6***	**19**	**0.95**

Another example is Data_User_Modeling_Dataset (145 records) from UCI Machine learning, which has 5 condition attributes STG, SCG, STR, LPR, PEG and one decision attributes UNS.

- Suppose we choose randomly 50 rows from 145 rows, then the accuracy of those new 27 rules will automatically calculate to be 0.46 instead of 0.18
- Suppose we choose randomly 75 rows from 145 rows, then the accuracy of those new 41 rules will automatically calculate to be 0.57 instead of 0.21
- Suppose we choose randomly 100 rows from 145 rows, then the accuracy of those new 52 rules will automatically calculate to be 0.62 instead of 0.29

- Suppose we choose randomly 125 rows from 145 rows, then the accuracy of those new 64 rules will automatically calculate to be 0.75 instead of 0.33

The results of the four cases of Data_User_Modeling_Dataset after running Genetic will be as following in the table 6:

Table 6: *Confidence* and *Accuracy* after running Genetic algorithm of four cases of DUM Data

	Fuzzy rules	R1	R2	R3	….	R27	R28	….	R41	R42	….	R52	R53	….	R64	Total	Accuracy %
Case 1 50 rows	frequency	1	***1***	1	….	5										***50***	
	Confidence	0.66	***0.46***	0.59	….	0.38											
	Accuracy	0	***5***	1	….	2										**44**	**0.46**
Case 2 75 rows	frequency	1	1	1	….	1	3	….	2							***75***	
	Confidence	0.69	0.61	0.66	….	0.48	0.43	….	0.43								
	Accuracy	0	0	0	….	0	3	….	3							**43**	**0.57**
Case 3 100 rows	frequency	1	1	1	….	1	1	….	4	***1***	….	4				***100***	
	Confidence	0.7	0.6	0.7	….	0.43	0.48	….	0.46	***0.46***	….	0.42					
	Accuracy	0	0	0	….	0	0	….	2	***5***	….	2				**28**	**0.62**
Case 4 125 rows	frequency	1	1	1	….	5	3	….	2	2	….	1	***1***	….	4	***125***	
	Confidence	0.69	0.65	0.7	….	0.41	0.51	….	0.51	0.47	….	0.56	***0.48***	….	0.42		
	Accuracy	0	0	0	….	0	0	….	0	0	….	1	***4***	….	2	**15**	**0.75**

3.5 The performance curve

The performance curve of values for maximum **Accuracy** for each generation of the running program can be represented by Figure 3.5 and 3.6

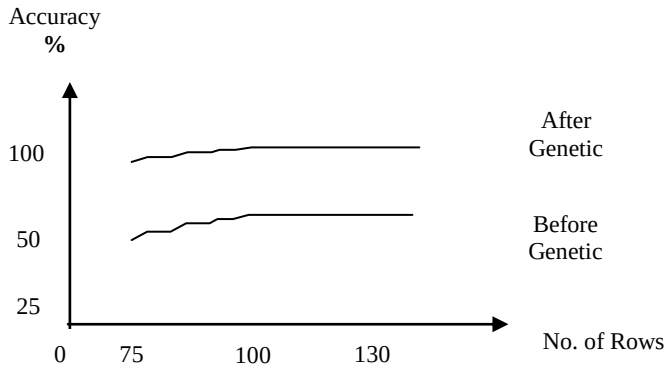


Figure 3.5: The performance curve of the Iris plants Dataset

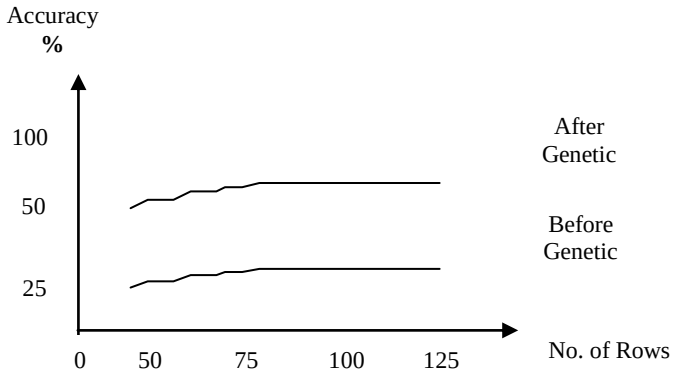


Figure 3.6: The performance curve of the Data_User_Modeling_Dataset

Contribution 3.2

Our Genetic-Based Fuzzy Rough Decision Model algorithm is more efficient and powerful than the Traditional Genetic fuzzy and rough models [33, 35, 36, 37], because our algorithm calculates automatically the confidence (fitness) of each rule by using the Average operator that takes the average confidence value of the fuzzy sets of each rule. Then the algorithm also calculates automatically the accuracy of each fuzzy rule. Then, the rules that that have a high accuracy are called *Strong rules*. After that, the algorithm using Genetic algorithm technique to generates another new fuzzy rules then it automatically calculates the accuracy of those rules which will be higher than the old rules before using Genetic algorithm.

4. Conclusion

Rough sets theory has been applied successfully in many disciplines. One of the major limitations of the traditional rough sets model is that it assumes that all attribute values are discrete. A real world data always contains mixed types of data such as continuous valued, symbolic data, etc. therefore all numerical or continuous data should converted to discretized data, here "Fuzzy Logic" can solve this problem to reduce information overload in a Fuzzy-Based Rough Decision Model.

One drawback of traditional Fuzzy-Based Rough Decision Model is that the linguistic values (fuzzy sets) for numerical values of each attribute has to be determining by the membership functions of these linguistic terms which, the user should define the parameters of membership functions of these linguistic values from his view which is different from one user to another. Therefore, we propose a new automated Fuzzy-Based Rough Decision Model algorithm that can define the parameters of membership functions of these linguistic values automatically by determine the number of fuzzy sets and the maximum and minimum values of each attribute then the algorithm finds the width (Δ) that divides the universe of discourse "*U*" of each attribute into "n" intervals according to the number of fuzzy sets then the algorithm calculates automatically the width (δi) according to the width (Δ).

Another drawback of the traditional rough sets model in the real applications is the inefficiency in the computation of core attributes and the generation of reducts. Most existing rough set models, do not integrate with database systems and a lot of computational intensive operations such as discernibility relations computation, core attributes search, reduct generation, and rule induction are performed on flat files, which limits their applicability for large data sets in data mining applications.

In order to improve the efficiency of computing core attributes and reducts, a New Rough Sets Model Based on Database Systems has been introduced for this purpose [Hu, X., Lin, T., 2004], which redefine the core attributes and reducts based on relational algebra such as *Cardinality*, *Projection*, and *Selection* and so on to take advantages of the very efficient set-oriented database operations.

The algorithm then will automatically calculate the confidence (fitness) and the accuracy of each fuzzy rough rule then it calculates the total accuracy value of all linguistic rules.

We build a new more powerful hybrid algorithm which integrates the advantages of our rough set and fuzzy set algorithms together with Genetic Algorithms (GAs). The proposed hybrid algorithm is hopefully to be more efficient and limited number of fuzzy rules for decision-making in terms of accuracy, simplicity and interpretability.

5. Future work

Depending on the choice of parameters, association fuzzy rule algorithms can generate an extremely large number of rules which lead algorithms to suffer from long execution time and huge memory consumption, So we will propose a Top k Rules algorithm to discover the *top-k* rules before using Genetic algorithm which having the highest support and confidence, where *k* is set by the user. This algorithm discovers all rules that have a support and confidence respectively higher or equal to user-defined thresholds minimum support *minsup* and minimum confidence *minconf*.

References

Bazan, J., Nguyen, H., Nguyen, S., Synak, P., Wroblewski, J., Rough set algorithms in classification problems, *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, L. Polkowski, T. Y. Lin, and S. Tsumoto (eds), 49-88, Physica-Verlag, Heidelberg, Germany, 2000

Cercone N., Ziarko W., Hu X., Rule Discovery from Databases: A Decision Matrix Approach, *Proc. Of ISMIS*, Zakopane, Poland, 653-662, 1996

Q. Shen, J. Richard**,** Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37(2004) 1351–1363.

Fernandez-Baizan, A., Ruiz, E., Sanchez, J., Integrating RDMS and Data Mining Capabilities Using Rough Sets, *Proc. IMPU*, Granada, Spain, 1996

Bhatt R B, Gopal M. On fuzzy-rough sets approach to feature selection. Pattern Recognition Letters, 2005, 26(7): 965–975.

Hu, X., Lin, T., Han, J, " A New Rough Sets Model Based on Database Systems" , Fundamenta Information 59 no. 2-3 pp.135-152 ., 2004

Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems, 1990, 17(2–3): 191.

Hu, X., Cercone N., Han, J., Ziarko, W, GRS: A Generalized Rough Sets Model, in Data Mining, *Data Mining, Rough Sets and Granular Computing*, T.Y. Lin, Y.Y.Yao and L. Zadeh (eds), Physica-Verlag, 447-460, 2002

Drwal G., Sikora M.: Fuzzy Decision Support System with Rough Set Based Rules Generation Method. In:Rough Sets and Current Trends in Computing, LNAI 3006. Springer-Verlag (2004) 727 733.

John,G., Kohavi, R., Pfleger,K., Irrelevant Features and the Subset Selection Problem, *Proc. ICML*, 121-129, 1994

Qiang He a, Congxin Wua, Degang Chen b, Suyun Zhao.: Fuzzy rough set based attribute reduction for information systems with fuzzy decisions, *Knowledge-based systems* 24 (2011) 689-696.

Kira,K., Rendell,L.A. The feature Selection Problem: Traditional Methods and a New Algorithm, *Proc. AAAI*, MIT Press, 129-134, 1992

Zhi, W.W, J.S. Mi and W.X. Zhang "Generalized fuzzy Rough Sets", Information Sciences, vol.151, p.p. 263–282, 2003.

Kumar A., New Techniques for Data Reduction in Database Systems for Knowledge Discovery Applications, *Journal of Intelligent Information Systems*, 10(1), 31-48, 1998

Wang, X.Z., E.C.C. Tsangb, S. Zhao, D. Chen and D.S. Yeung, "Learning fuzzy rules from fuzzy samples based on rough set technique", Information Sciences, vol. 177, p.p. 4493–4514, 2007.
Lin T.Y., Cercone, N. (eds), Rough Sets and Data Mining: Analysis of Imprecise Data, Kluwer Academic Publisher, 1997

A.M. Radzikowska, E.E. Kerre, A comparative study of rough sets, Fuzzy Sets and Systems 126 (2002) 137–155.

Lin T.Y., Yao Y.Y. Zadeh L. (eds), Data Mining, Rough Sets and Granular Computing, Physica-Verlag, 2002

Q. Shen, A. Chouchoulas, A rough fuzzy approach for generating classification rules, Pattern Recognition 35 (2002) 2425–2438.

Liu, H., Motoda., H. (eds), Feature Extraction Construction and Selection: A Data Mining Perspective. Kluwer Academic Publisher, 1998

Jiuping Xu, Lihui Zhao.: A multi-objective decision-making with fuzzy rough coefficients to the inventory problem. *Information sciences*,180(2010) 679-696

Modrzejewski, M., Feature Selection Using Rough Sets Theory, *Proc. ECML*, 213-226, 1993

Nguyen, H., Nguyen, S., Some efficient algorithms for rough set methods, *Proc. IPMU* Granada, Spain, 1451-1456, 1996

Lirong J, Liu S. Extension of rough sets model for probabilistic decision analysis from rough fuzzy decision tables. Journal of Southeast University, 2006b, 2(5): 246–250**.**

Pawlak Z., Rough Sets, *International Journal of Information and Computer Science*, 11(5), 341-356, 1982

Pawlak Z., Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1992
Polkowski, L., Skowron, A., Rough mereology, *Proc. ISMIS*, Charlotte, NC, 85-94, 1994

Polkowski, L., Skowron, A., Rough mereology: A new paradigm for approximate reasoning, *J. of Approximate Reasoning*, 15(4), 333-365, 1996

Skowron, A., Rauszer C., The Discernibility Matrices and Functions in Information Systems, *Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory*, K. Slowinski (ed), Kluwer, Dordrecht, 331-362, 1992

Yifan Wang, Mining stock price using fuzzy rough set system, Expert Systems with Applications 24 (2003) 13–23.

Tzung-Pei Hong a, Li-Huei Tseng b, Been-Chian Chien c.: Mining from Incomplete quantitative data by fuzzy rough sets, Expert System with Application 37 (2009) 2644-2653.

Yan-Qing Yao, Ju-Sheng Mi, Zhou-Jun Li. Attribute reduction based on Generalized fuzzy evidence theory in fuzzy sets and systems. Fuzzy Sets and Systems, (2011) 509-517.

Cordon, O., Herrera, F., Hoffmann, F., Magdalena, L. (2001): Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases. World Scientific Publishers, Singapore.

I.H. Toroslu, Y. Arslanoglu, Genetic algorithm for the personnel assignment problem with multiple objectives, Information Sciences 177 (2007) 787–803.

Wen-Yau Liang, Chun-Che Huang.: The generic genetic algorithm incorporates with rough set theory – An application of the web services composition. Expert Systems with Applications 36 (2009) 5549-5556.

Marek Sikora.: Fuzzy Rules Generation Method for Classification Problems Using Rough Sets and Genetic Algorithms. Computer Science, (2005), 44-100.

Mohammad Lutfi Othman, Ishak Aris, Mohammad Ridzal Othman, Harussaleh Osman. Rough-Set-and-Genetic-Algorithm based data mining and rule Quality Measure to hypothesize distance protective relay operation characteristics from relay even report.ElectricalpowerandEnergySystems33 (2011)1437-1

Big Data Mining using Soft Computing Techniques: An Overview

Mohamed A. Ghazala[1], Khaled T. Wassif [2] and Hesham A. Hefny [3]

[1] Dept. of Computer Science
Faculty of Graduate Studies & Statistical Research,
Cairo University, Cairo, P.O. 12613

[2] Dept. of Computer Science,
Faculty of Computers & Artificial Intelligence,
Cairo University, Cairo, P.O. 12613

[3] Vice Dean
Faculty of Graduate Studies & Statistical Research,
Cairo University, Cairo, P.O. 12613

Abstract

We are now in the era of Big Data; Big Data is now expanding in all the science and engineering fields in the sectors of physical, biological sciences. Big Data Mining is the applicability of extracting useful valuable information from these huge data sets that is due to volume, variety, variability, velocity and value. Now it is not possible and much harder to use traditional Data mining techniques to handle uncertainty, ambiguity and complexity in data, on other hand Soft Computing methodologies are efficient techniques for processing large data sets to exploit the tolerance of the imprecise and vagueness of Data. Soft Computing as Fuzzy learning models, ANNs, GAs algorithms are very effective to exploring and processing Big Data sets, in this paper, we will present different methods Mining Big Data with various soft computing interactions and retrieve better results in performance and accuracy with comparing with other systems.

Keywords: Big data, Data Mining, Soft Computing, FRBCS, Evolutionary Computation

1. Introduction

The world Big Data is nothing different from data, just a prefix big has used to define huge amount of size and complexity, but this type of data cannot be handled through traditional data mining techniques and tools. Big Data is nothing rather than data but available in non-identical formats, un structured, several sources and large amount, for example all data get stored on Facebook servers at most of us daily use Facebook and upload multiples of photos, videos and audios. This data is nothing but big data as it also characterized by its complexity. Rohit & Vijay (2014).

This paper describes how about this big data is stored and processed using different techniques of soft computing in classification of data. The digital revolution has made digitized information easy to retrieve and all the data from anywhere. With the development of computer hardware, software and the rapid computerization of business, huge amount of data have been collected and stored in databases. As a result, traditional ad hoc MIS and the statistical techniques and data management tools are no longer adequate for analyzing this vast collection of data. Sushmita et al. (2012).

The concept of big data refers to data that are unstructured data or data, which are huge sets and complex that are preventing traditional applications to process such data.
Big Data involves in data storage, analysis, searching, querying, and updating and information privacy. Christopher, E., & Mgbeafuluike (2018). The sources of big data are social networking data, sensors, web sites, etc.

The banking sector has been expanding into the banking technologies and increasing the focusing on data in data warehousing, and unstructured data to reach the data driven organizations. These financial services data have generated in consideration to threats or security issues, increase in customer evolution, the fright for loans, fraud, profitability and other banking facilitating issues. Dazhong et al. (2015). The healthcare sector has also experienced the Big Data through integrate huge various data for the diagnoses, treatments, different departments to analyze these data to get the optimum meditation. AI analytical tools have used to help managing data and providing resolutions. Sarwar et al (2017).

Soft computing which is sometimes called Computational Intelligence is abroad of Artificial Intelligence (AI).
The principal constituents of Soft Computing (SC) are Fuzzy Logic (FL), Evolutionary Computation (EC), Machine Learning (ML) and Probabilistic Reasoning (PR), with the latter subsuming belief networks and parts of learning theory. Soft computing (SC) solutions are unpredictable, uncertain and between 0 and 1.

The methodologies that are being used and focused in soft computing is related to mathematical solving problems to create the required models. These models are evaluated, analyzed in order to work on problems with nonlinear polynomial boundaries therefore, conventional computing cannot solve such problems.

2. Big Data Specifications & Features

Data is classified into four different categories:

The Fastest on is **Structured Data**, the data that does have a proper format associated to it. For example, the data that is represented in the databases, CSV files, and the excel spread sheets are being categorized as structured data.

The Second type is the **Semi Structured Data**, the data that does not have a proper format associated to it, for example, the data that are represented in the emails, log files, and word document.

The third type is the **Quasi Structured Data**, hat does have irregular format, for example, data about which webpages a user visited and in what order.

And the forth type is the **Un Structured Data**, the data that does not have any format associated to it. For example the images, audios, and videos.

These data can be generated from terabytes to petabytes and coming from diverse sources such as social networks, mobile devices, multimedia data, webpages or sensor networks among others. S. Madden. (2012).

2.1 Characteristics of Big Data

Usually the Big Data is characterized by the V's innovations technologies. The most important three characteristics are the 3V's (Volume, Velocity, Variety):

The fastest one is the Volume, which refers to the amount of data being generated. The next is the Velocity, which refers to the speed while the data is being generated. While the variety refers to the different types of data that is getting generated. Then the Big Data is expanding very fast introducing the 4 V'S (Volume, Velocity, Variety, and Veracity). Veracity refers to uncertainty of data can be found due to the inconsistency and incompleteness. Dazhong (2015). Recently, the 5 V's has appeared in the technology of handing Big Data (Volume, Velocity, Variety, Veracity and Value). Figure 1 shows how the big data can be characterized as by 5 V's.

Figure 1: 5Vs of Big Data

We must take into consideration the additional definitions up to 9V's & 11V's that are terms like Viability, Validity, Vulnerability, Vagueness and Visualization, among others. S Mitha el al (2013).

The management of storing and processing the large volume of data is issue related to DBMS and old relational models. The method of exploring these large volumes of data, which leads to discover useful information and knowledge for best future decisions. The Traditional Processing tools and techniques (Excel, RDMSs, etc.) are not capable to handle the scalability of fast-growing data sets. People using Excels know that a sheet with more than 1 million record will be hard to be processed and data bases that have reached tera bytes of storage are becoming more difficult to performing mining on the data. In addition to *daily huge volume of social media data and real time processing data that requires higher scale up machines and tools that can handle different characteristics of data that was clarified previously. Ishwarappa (2015); Magesh (2017).

The concept of the Massive Parallel Processing MMP is the main based of capability to store and process petabytes of data and more into shared partitioned multiple nodes in which each node have its own resource of processors and memory on its local. Jogalekar (2000). A single machine cannot efficiently manage high volumes of data. Therefore, big data technologies intended to solve these problems by adjusting the concept of Divide – And – Conquer.

Map-Reduce consists of two functions Map function which divides the input data and distributes it across multiple nodes to process the data parallelly then Reduce function that collects the results from the nodes and aggregate the output results. Jogalekar (2000). Figure 2 shows the Map Reduce workflow.

Hadoop is designed to be consist of HDFS and YARN (Yet Another Resource Negotiator) which is a part of Hadoop and is responsible for the resource management of all processes performed on the Hadoop.

There are several Hadoop-based tools that helps Bid Data processing are more efficient and convenient. Apache Sqoop is a tool that perform transferring data between RDBMS and Hadoop. Apache Flume is a tool to combine and aggregate data from different sources into Hadoop. HBase is tool for NoSQL data and the NoSQL is Not only SQL data which works to deal data that is not relational data and cannot be stored on RDBMS. Apache Hive is a data warehousing framework built on the top of Hadoop which allow data stewards to store data in tables and write queries SQL-like using HiveQL. Apache Spark is an in-memory distributed framework which is used for fast mining and analyzing data. Spark has proven that it is well in the speed optimization Jogalekar (2000).

3. Data Mining

The concept of Data mining deals with extracting the useful information from the available data. Foster (2013). The data mining process consists of several stages first to get and clean the data or what is named as ETLs (Extract Transfer Load) then building the model by choosing an appropriate learning model to be applied based on the business case of the data and according to what you want to extract from the data. Then finally developing and deploying the model. Jogalekar (2000).

The idea of parallel data mining has been emerged with big data to improve the usage of massive, complex amount of datasets by using Artificial Intelligence (AI) systems that make computers think like humans. Data Mining refers for searching, analyzing and extracting the valuable data from data warehouses to support making decision. It is well known as the Knowledge Data Discovery (KDD).

There are two main types of data mining models:

The first type is the Descriptive data mining which used for summarization analysis and the structure of the model.

The second type is the Predictive data mining which is used for developing a model based on existing data and extract classifications with more specifications. Commonly used in predictions and is awful with big data. Barbierato el al (2014).

Data mining models also are divided into 2 categories: the supervised models and the unsupervised models. The supervised models are considered to deal with historical

training data and make the predictions or classifications of new data. The unsupervised models deal to find patterns or clusters from the given data.

Data mining process can be involved into different techniques such as: Classification, Estimation, Prediction, Association Rules, Clustering, Description. The following table shows the methods and algorithms that are applied on these techniques

Technique	Description	Algorithms	Category
Classification	It is the analysis that summarize the important features about the data, and meta data	• Decision Trees • Decision Tables • Naïve-Based Algorithm • Nearest- neighbor Algorithm	Supervised / Unsupervised
Regression	Regression analysis is the data mining method of identifying and analyzing the relationship between variables and forecast new data	• Linear Models (Linear Regression, Logistic Regression) • Instance – based Models	Supervised
Prediction	This technique analyzes past events or instances in a right sequence to predict a future event	• Prediction Task • Bayesian Network • Statistical Model	Supervised
Association	This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.	• Aprioro Algorithm • Parallel Algorithm • Distributed Algorithm	Unsupervised
Clustering	The Clustering analysis is a data mining technique to identify group of data are like each other.	• K-Means Algorithm • Hierarchical clustering • Cure Algorithm	Unsupervised

Table 1: Data Mining Algorithms

3.1 Data Mining & Analytics

Big Data mining is referring to collective data mining or extraction techniques which are performed on huge volume of data sets. Big Data mining is primarily performed to extract useful information or patterns from homogenous big data.

This is usually performed on large quantity of unstructured data. Big Data mining works on data searching, extraction and comparison methods. It is also support computer devices to perform complex operations on large data sets.

Big data mining techniques and processes are also used within big data analytics and business intelligence to deliver summarized targeted and relevant information, patterns and/or relationships between data, systems, processes and more. Kirubakaran et al (2015).

3.2 Big Data Mining Tools

As the voluminous of the data is in a rapid growth traditional Data Mining tools like Weka and Rapid Miner are not scalable and do not support the processing of these data however there are many Big Data Mining open source tools, the following presents the most popular tools:

3.2.1 Apache Hadoop

Hadoop is an open source implementation of Map Reduce computational paradigm. Hadoop is a framework that supports data intensive distributed applications freely available set of tools and libraries, based on Java software framework for processing, development, and execution the large volume of distributed datasets and application. Smitha (2013).

3.2.2 Apache Hive

Apache Hive is a data warehouse software that facilitate querying and management on large data in distributed storage like HDFS. Hive enables a mechanism for project structure and query the data using a SQL like language which is HiveQL that is friendly to the SQL developers. In addition, this language allows map reduce to be plug in and to be implemented to express efficient processing to big data.

3.2.3 Apache Spark

Apache Spark is an in-memory distributed framework which is used for fast mining and analyzing data. Spark has proven that it is well in the speed optimization. Spark was built to address the shortcomings of Hadoop and it does this incredibly well. For example, it can process both batch data and real-time data, and operates 100 times faster than MapReduce.

4. Soft Computing

Soft computing is a methodology that is based on Artificial Intelligence (AI). Soft computing focuses on fusing processes that is used to create models to solve real complex problems that cannot be solved using traditional mathematics. These models are usually evaluated, analyzed and supported using advanced mathematical tools. It is serves to provide solutions for uncertainty and imprecise in data. Soft computing problems are modeled using imprecise boundaries: 0 to 1. Kirubakaran et al (2015).

Soft computing includes several techniques like fuzzy logic, neural networks, artificial intelligence, genetic algorithm and machine learning. Rohit (2014).

4.1. Fuzzy Logic

The models that solve vagueness in qualitative knowledge as well as the handling of uncertainty is different stages are probably using fuzzy sets

Fuzzy logic is able of supporting to a reasonable extent human type analysis in natural form. It is the earliest and most widely reported ingredient of soft computing.

Lotfi Zadeh first defined the Fuzzy Set Theory in 1965. In mathematics, fuzzy sets are sets whose elements have degrees of membership. A Membership is the Fuzzy sets have values between 0 and 1 that indicate the degree to which an element has membership in the set.

The degree 0 means that the element has no membership; at one, it has full membership. Fuzzy membership function can be presented as: $\mu X(x) \epsilon (0,1)$ where, X is a set and x is an element. The Concept of Linguistic Variables was introduced by Zadeh in 1975. For example, I am not tall, where tall is a linguistic variable and the height is the linguistic variable L.A. Zadeh. (1975).

Mamdani introduced a new fuzzy control at the end of 1970s; the basic idea was to utilize linguistic if-then rules of human experts. For example, if your height is short and your weight is heavy Then you should not eat too much. Mamdani & Assilian. (1975).

4.2 Neural Networks

Neural Networks were considered earlier for data mining because of the inherent it black-box nature. No information was on hand in symbolic form, suitable for verification or understanding by humans. Recently there has been widespread movement aimed at redressing this, by extracting the embedded knowledge in trained networks in the form of symbolic rules. Nour E et al (2015).

Artificial Neural Networks (ANNs), like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Artificial Neural Networks are collected of initial computational units called neurons in which these neurons are combined into different architectures. A connection layer networks consists of:

- Input layer, made of multiple n neurons.
- Hidden layer, consist of one or more intermediate hidden layers contains of m neurons.
- Output layer, composed of p neurons.

The simplest network is composed of a single neuron, with n inputs and a single output. The basic learning algorithm of perceptron. The Neural network with an input layer, one or more intermediate layers of neurons and an output layer is called multiple layer perceptron MLP. C. Gallo. (2015).

An illustrative drawing represents a multi-layer neural network is show in figure 2.

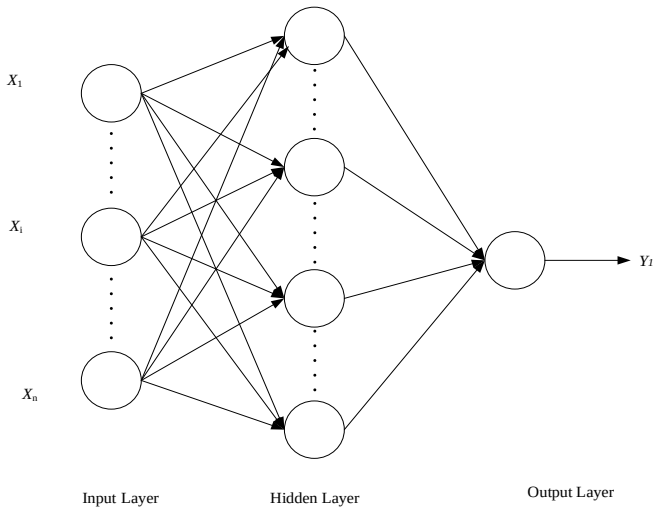


Figure 2: Multi-Layer Neural Network

Where X_i the i^{th} input to the neuron and Y_i is the output of the node. Eseye et al (2016) the efficiency is dependent on the larger volume of data for the training process in which increasing the number of hidden layers of the neural network, this process is called deep learning. The training phase that includes labeling large amount of data and determine their matching features. Deep Learning is such an approach to help system deal with complex perception activities with the maximum accuracy.

4.3 Genetic Algorithms

GAS are effective, robust, and adaptive global search mechanism, it is suitable for large space. Genetic Algorithm (GA) is based on search optimization techniques in which its methodology is principles of Genetics and Natural Selection. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. In Herrera F. (2016) a basic GA algorithm is illustrated as follows:

```
Basic Genetic Algorithm
Begin
 t = 0
        Initialization P(t)
        evalution P(t)
        While (the stop contition is not verified) do
         Begin
        t = t + 1
        selection P'(t) from P(t-1)
        P''(t) ← crossover P'(t)
        P'''(t) ← mutation P''(t)
        P(t) ← replacement (P(t-1),P'''(t)) evaluation P(t)
        End
End
```

4.4 Fuzzy Rule Base Classification Systems

Fuzzy Rule-based Classification Systems (FRBCSs) are effective and popular tools for pattern recognition and classification. López et al (2015). These techniques are able to obtain good accuracy results while providing a descriptive model for the end user through the usage of linguistic labels. Therefore, FRBCSs is an appropriate solution for dealing with uncertainty, ambiguity and vagueness in an effective way. It is also very effective when dealing with large data sets and Big Data.

The Fuzzy Rule-Based Classifier is a classifier with fuzzy classification rules in which a classification rule is an if-then rule with fuzzy conditions in the if-then part and a consequent class in the then part.

Usually any classification problem is determined by m training instances $X_P = (X_{P1}, \ldots, X_{Pn}, C_P)$, P = 1, 2, . . ., m with n attributes and M classes where X_{Pi} is the value of the attribute i(i = 1, 2, . . . ,n) and C_P is the value of class label $(C_1, \ldots, C_M)$ of the p-th training sample.

A FRBCS is composed by two elements: The Inference engine and the Knowledge Base (KB). In a linguistic FRBCS, the KB is formed of the Data Base (DB), which holds the membership functions of the fuzzy partitions connected to the input attributes and the Rule Base (RB), which comprises the fuzzy rules that describe the problem. A learning process is needed to build the KB from the available examples.

The two main components of FRBCSs are described as follows:

1) Knowledge base (KB): It is produced of both the rule base (RB) and the database (DB), as there is where to store the rules and membership functions used to model the linguistic variables.

2) Fuzzy Reasoning Method (FRM): This is the mechanism used to classify examples using the information stored in the KB.

In order to generate the KB, a fuzzy rule learning algorithm is applied using a training set R composed of N labeled examples $x_i = (x_{i1}, \ldots, x_{iN})$ with $i \in \{1, \ldots, N\}$, where x_{in} is the value of the n-th feature ($n \in \{1, 2, \ldots, N\}$) of the i-th training example. Each example belongs to a class $y_i \in C = \{C_1, C_2, \ldots, C_M\}$, where M is the number of classes in the problem. Río, Sara et al (2015).

The RB Rule base is generated from a set of rules with the following structure

$$\text{Rule } R_j: \text{If } x_1 \; is \; A_j^1 \text{ and } \ldots \text{and } x_n \text{ is } A_j^n \text{ then Class } C_j \text{ with } RW_j \qquad (1)$$

Where R_j is the label of the j-th rule, $x_i = (x_1 \ldots x_n)$ is an n-dimensional pattern vector that represents the example. A_{ji} is a linguistic label model by a triangular membership function, C_j is the class label and RW_j is the rule weight.

One of the most common ways to compute the rule weight is the Penalized Certainty Factor (PFC). H. Ishibuchi & Yamamoto (2005).

$$Rw_j = PCF = \frac{\sum_{x_P \in Class\ Cj} \mu\ A_j\ (x_P)\ - \sum_{x_P \notin Class\ Cj} \mu\ A_j\ (x_P)}{\sum_{P=1}^{P} \mu\ A_j\ (x_P)} \qquad (2)$$

Where $\mu\ A_j\ (x_P)$ is the membership degree of the example x_p with antecedent part of the fuzzy rule R_j and C_j is the consequent class of rule j.

In order to provide the final class associated with a new pattern $x_p = (x_{p1, \ldots, } x_{pm})$, the winner R_w is determined through the following equation:

$$\mu_w(x_P)\ .\ RW_w = \max\ \{\mu_j\ .\ RW_j\ ;\ j = 1\ \ldots L\ \} \qquad (3)$$

The Fuzzy reasoning method of the wining rule is used when predicting a class using the built KB for a given example. By this approach, the class assigned to the example X_p, $C_{p,}$ is the class indicated in the consequent result of the winner rule $R_{w.}$ Río, Sara et al (2015).

In the process that includes multiple rules, obtain the same maximum value for equation 3 but with various classes on the consequent.

5. Overview on the Related Researches

The researchers and practitioners was digging to many approaches that must deal Big Data and handling the scalability in order to Machine Learning and Data Mining algorithms to addresses such problems of complexity and uncertainty. This resulted frameworks are implied to design of divide-and-conquer map reduce processes. Fernnádez et al (2017).

Sara del Rio has proposed the Chi-FRBCS-Big Data algorithm, a linguistic fuzzy rule-based

classification system that uses the Map Reduce framework to learn and fuse rule bases.

The Chi- FRBCS is based on Chi et al. approach, this approach can generate several rules with the same antecedent. If the consequent of those rules belongs to the same class.

The Chi-FRBCS-BigData algorithm that have developed to deal with big data classifications problems. This method uses two stages of MapReduce processes:

- One MapReduce process to build the model of big data training set, this procedure is to build the KB with a MapReduce process in the schema of Chi-FRBCS-BigData with the following steps:
    I. Initial Phase: The first phase in which the DB is created using equally distributed triangular membership function as in Chi-FRBCS.

II. Map: In this second phase, the data is being partitioned over its data to build its associated RB_i and the rule weight is being computed using the set of examples of the map process.

III. Reduce: In the last phase the processing unit receives the output results from the map process RB_i and aggregate them to form final RB_R.

In figure 3, an illustrative diagram for how to build the model is represented.

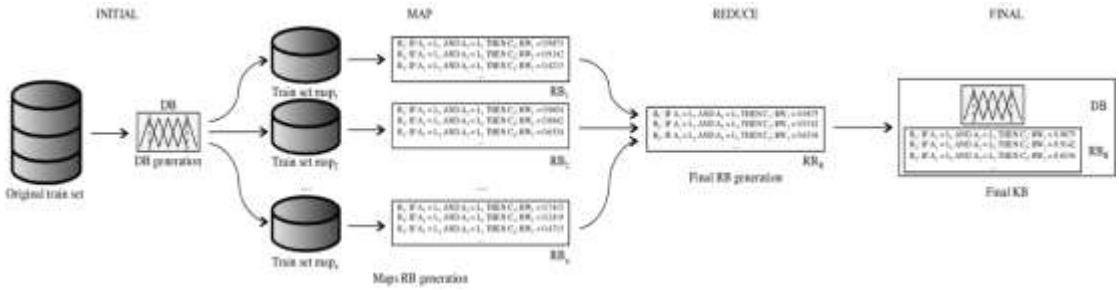


Figure 3: A flowchart of how the building of the KB in Chi-FRBCS-BigData

- The Second process of the Chi-FRBCS-BigData is classifying the BigData Samples, it also uses MapReduce process to predict the class of the examples that belong to Big Data sets using the KB built in the previous step. This process is similar to earlier one in respect to stages, the classification prediction process is depicted in Figure 4 and includes the following phases:

I. Initial Phase: In this first phase, the system detect the segments of the original data sets that needs to be classified.

II. Map: In this second phase the map task estimates the class for the examples that are included in the data partitions. It goes through dividing the examples of data into chunks and predict it's correspondent class according to the given KB.

III. Final: The last phase the results are being aggregated the results of the map task and computed the output results.

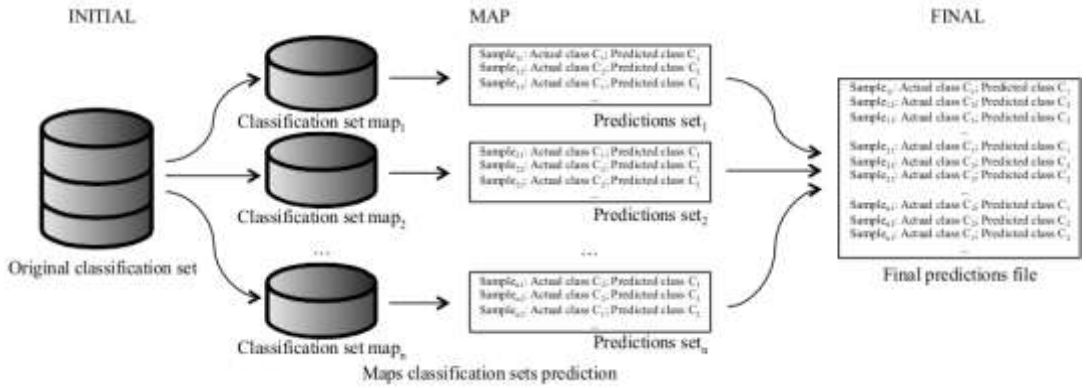


Figure 4: A flowchart of how the classification of a big dataset in Chi-FRBCS-BigData

Chi-FRBCS-BigData approach have been developed: Chi-FRBCS-BigData-Max and Chi-FRBCS-BigData-Ave. While both versions share most of their operations, they differ in the reduce step of the approach, where the generated rule bases are combined. Therefore, these variants of Chi-FRBCS-BigData algorithm obtain different final rule bases. Río, Sara et al (2015).

The chi-FRBCS approach testing was using forest cover type data set. Forest cover type dataset consists of 12 measure attributes (10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables) which is presented in below. Dataset consists of 5,81,012 records, which is divided into train data (consists of 4,64,810) and test data (consists of 1,16,202 records).

Table 2: Forest Cover Type Dataset

Data set	Attribute Characteristics	No of records	No of attributes	No of rows	No of columns
Cov_type	Categorical, integer	5,81,012	54	581012	55

Kumar proposed the feature selection method combing to the fuzzy rule-based classification system to increase accuracy in large data sets, he also identified the bagging method to divides the datasets into n equal datasets to let the feature selection select attributes of higher weight with class attributes. He examined the same Dataset that was previously clarified and resulted the following results

This proposed approach was based on the Chi-FRBCS-BigData method. This algorithm uses the MapReduce framework to handle huge amount of data through its parallel processing. So, the work was focusing to enhance the Chi-FRBCS-BigData to adapt higher accuracy and combining the feature selection and Bagging. Kumar et al (2015).

The accuracy for the FRBCS has increased by combining bagging and feature selection process with it than FRBCS without using these approaches. Table 3 shows the accuracy of two algorithms namely, chiBigData-Max and chi-BigData-Ave.

Table 3: Chi-BigData Accuracy Measures

Dataset	FRBCS without using Bagging and Feature selection		FRBCS using Bagging and Feature selection	
	Chi-Bigdata max	Chi-Bigdata ave	Chi-Bigdata max	Chi-Bigdata ave
Covtype_2_vs_1	74.63	74.61	79.04	77.55

A new CHI-BD approach was introduced by Mikel ElKano as a new classification system for Big Data that was also based on the original Chi-FRBCS, he took the

advantage experiment the Chi-FRBCS-BigDataCS using MapReduce regardless of the number of mappers used for its processing which provides the same classification performance Elkano et al (2017).

This approach was presented in Río, Sara et al (2015). as an optimization of the original Chi algorithm to Big Data imbalanced binary classification problems. In order to move on this optimization, the authors proposed the usage of Hadoop framework. This algorithm is composed of two steps:

I. Generating the rule bases: It is an independent Chi Classifier that is learning from each mapper. The learning process of each classifier is performed considering only the input split of instances associated with the mapper. Once all classifiers have finished their learning phase, as many rule bases as mappers (classifiers) are obtained.
II. Fusion of rule bases: all the rules are being aggregated in one reducer to provide the final rule base.
The following figure illustrate the workflow on the CHI-BD

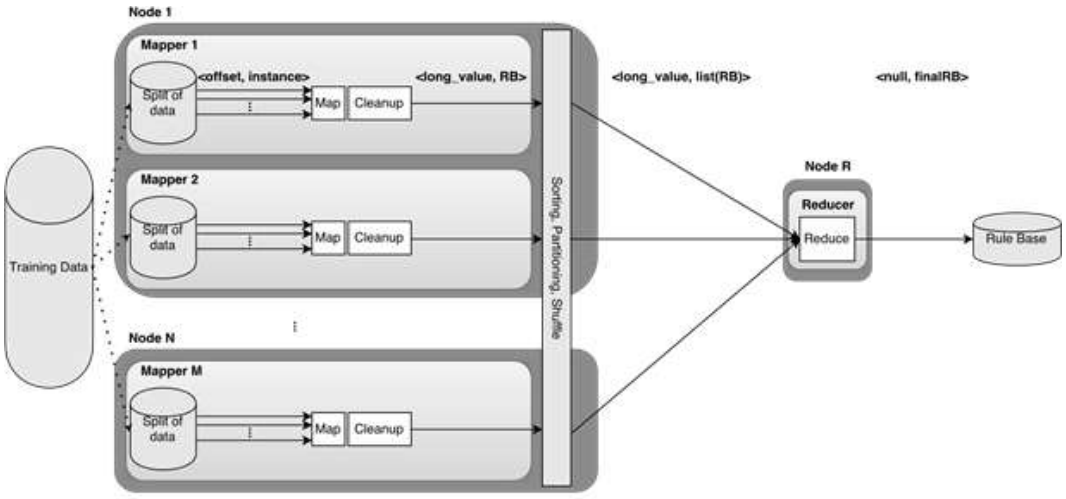


Figure 5: Data flow of the CHI-BD Algorithm

The algorithm was examined to outcomes a nearby results to Chi-FRBCS-BigData with a minor increase in performance.

Chi-Spark approach was designed by Alberto Fernandez to represent the benefits of Spark framework to work with big data classification problems and to work on the same model of the Chi-et al. foundation. To do such experiments on Spark framework it is must to understand the concept of Resilient Distributed Datasets (RDDs) as well as using different kind of optimized distributed operators over these RDDs. This process does

perform a given function over each element then aggregate the date for the final results. This allows spark applications to be executed more efficiently. Zaharia et al (2012). Below figure represents the Spark's dataflow.

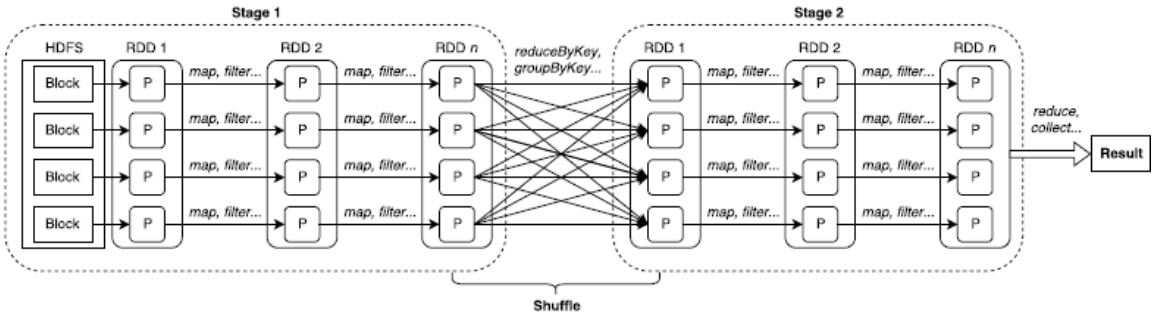


Figure 6: Spark's data flow (P = Partition).

This algorithm was implementing the original Chi-FRBCS-BigData algorithm on Spark environment and was donated by the Chi-Spark. In addition to an Evolutionary Fuzzy system for Rule selection in Big Data Classification EFS, was introduced to allow selecting the optimal parameters in building the FRBCS after the Chi-Spark finish for more improvement. Fernández et al (2017).

And recently, Mikel ElKano proposed a new distributed learning algorithm named CFM-BD to build the idea of enhancing accuracy with the fuzzy rule-based classification system as it is using Apriori algorithm. The proposed learning algorithm is composed of three phases, Data pre-processing and partitioning then rule induction and at the end rule selection. It compacted a competitive achievement in Big Data classification problems. Elkano et al (2019).

CFM-BD was newly proposed approach that is built on a new distributed fuzzy rule induction algorithm design especially for Big Data classification problems. The objective behind this proposal is by building a compact fuzzy models achieving high classification performance and interpretability. This learning algorithm is composed of three phases:

I. Pre-processing and partitioning: In this phase the fuzzy sets (linguistic labels) are built from the training data.
II. Rule induction process: In this phase after the pre-processing the fuzzy sets have been built, the rule bas is composed by applying a rule induction algorithm in which applying the concepts introduced in CHI-BD. Elkano et al (2017).
III. Evolutionary rule selection: When the rule base has been created, a rule selection process is carried out in order to obtain a compact and accurate model. The author proposed to apply the CHC evolutionary algorithm (EA) Jogalekar (2000).

because of its ability to deal with complex search spaces and the good results achieved by this EA in state of-the-art in FRBCSs

Below figures shows the Spark stages of the rule induction process

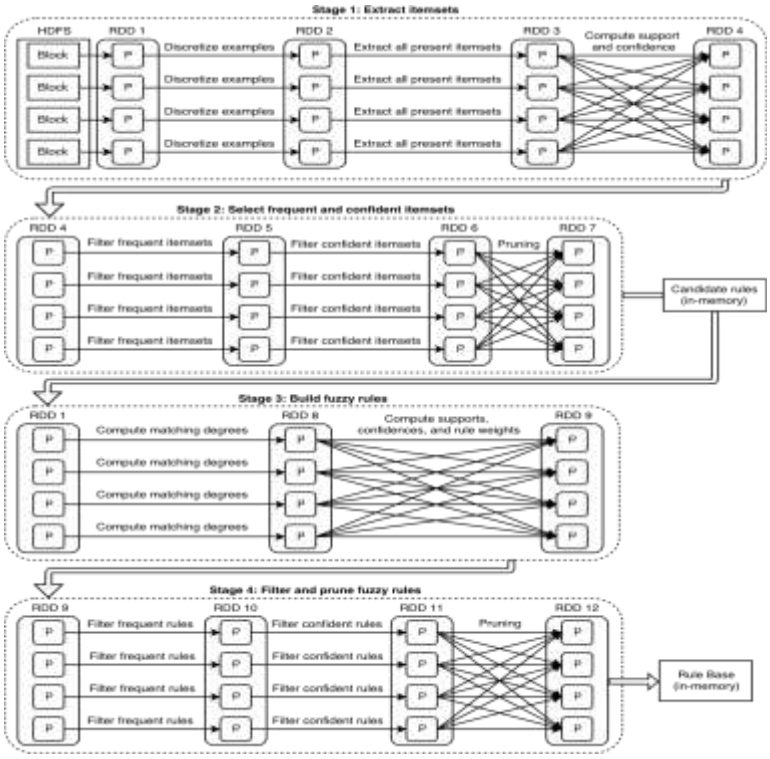


Figure 6: Spark stages launched during the rule induction process (P = Partition).

6. Discussion

In table 4, it shows a comparative study between the related researches that focus on Big Data mining techniques with combining of soft computing approaches specially the Fuzzy Rule-based Classification Systems.

Table 4: Review on the related work for Big Data and FRBCS techniques

No.	Author	Tittle	Goal	Strength	Weakness
1	Sara del Río (2015)	A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules	Addressing Big Data Problems using Linguistic Fuzzy Rules	Application was proposed the Chi-FRBCS-Big Data algorithm, a linguistic FRBC that uses the Map Reduce framework to learn and fuse rule bases	Chi-FRBCS-BigData-Max works faster with large number of mappers but with decreasing in accuracy

2	Akil Kumar (2015) Improved Fuzzy Rule-based Classification System Using Feature Selection and Bagging for Large Datasets	Improving the FRBCS using feature selection and Bagging	Obtains an interpretable model that can handle huge data with better performance	Low accuracy and for future work to combine both feature selection and bagging can increase accuracy
3	Mikel ElKano (2017) CHI-BD: A fuzzy rule-based classification system for Big Data classification problems	Addressing the Big Data classification problems using FRBCS	An application was proposed CHI-BD, to obtain good classification performance for Big Data	Increase the performance of Big Data Classification but still need to obtain more accuracy
4	Alberto Fernandez (2019) Chi-Spark-RS: a Spark-built Evolutionary Fuzzy Rule Selection Algorithm in Imbalanced Classification for Big Data Problems	Presenting an evolutionary fuzzy rule selection model in Big Data using Spark	Designed an Evolutionary Fuzzy Rule Selection algorithm within a Spark environment	Application in the early stage beside it took more time in the pre-sessions of the model for preprocessing the data
5	Mikel Elkano (2019) CFM-BD: a distributed rule induction algorithm for building Compact Fuzzy Models in Big Data classification problems	Proposed a new learning algorithm called CFM-BD to enhance accuracy	The proposed algorithm a new model with combining to the evolutionary optimization	Worked to enhance the accuracy, but with high complexity of the model and low scalability

The above represented related works all consider the measures of the quality by building a confusion matrix presented in table 5, which organizes the examples for each class in respect with their correct or incorrect labeling.

Table 5: Confusion matrix for a two-class problem

	Positive Prediction	Negative Prediction
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

The efficiency in classification for the proposed approaches had been evaluated using the most frequently used empirical measure, accuracy, which is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

Where, TP is True Positives, TN is True Negatives, FP is False Positives and FN is False Negatives from confusion matrix. TP represents the count of True positive prediction and WP represents the count of wrong prediction made by the classifier.

7. CONCLUSION

This paper has present different study on the Big Data with Data mining techniques and introduction to Soft Computing technologies that can be presented in Big Data. A descriptive illustration has been represented to shows various techniques used Fuzzy Rule Based Classifications systems with different approaches in Big Data tools and techniques.

A comparison between related works were introduced to make a reasonable study to the work, and for future work to design a model for combining Fuzzy rule-based classification systems with Apache spark and hive then attempt to customize a feature selection to enhance accuracy of big data sets.

Acknowledgements:

This work was supported by the vice dean of the faculty of graduate studies and statistical researches. All the thanks to family members and supported colleagues.

References

Rohit, P., & Vijay, K. (2014). A Survey Paper on Data Mining With Big Data. International Journal of Innovative Research in Advanced Engineering (IJIRAE) 1(1).

Sushmita, M., Sankar, K. P., Pabitra, M. (2012). Data Mining in Soft Computing Framework: A Survey. IEEE Transactions On Neural Networks, 13(1).

Christopher, E., & Mgbeafuluike (2018). Soft Computing: A Fundamental Approach in Analyzing Big Data. Journal of Information Technology & Software Engineering, 8(1).

S. Madden. (2012). From Databases to Big Data. IEEE Internet Computing, 16(3), 4–6.

Dazhong, W., David, R., Lihui, W., Dirk, S. (2015). Cloud-Based Design and Manufacturing: A New Paradigm in Digital Manufacturing and Design Innovation. Journal of Computer-Aided Design 59(1) 1-14.

M. Sarwar, Hanif K. H., Ramzan, T., Awais. M, & Muhammad, A. (2017). A Survey of Big Data Analytics in Healthcare. (IJACSA) International Journal of Advanced Computer Science and Applications 8(6).

Lee J, Lapira E, Bagheri B, Kao H (2013) Recent advances and trends in predictive manufacturing systems in big data environment. Manufacturing Letters 1: 38-41.

Ishwarappa, Anuradha J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. International Conference on Intelligence Computing, Communication & Convergence

Magesh M., P Swarnalatha. (2017). Big Data and Its Applications: A Survey. Research Journal of Pharmaceutical, Biological and Chemical Sciences, 8(2).

S Mitha T, MCA, M.Phil, & M.Tech, V. (2013). Application of Big Data in Data Mining. International Journal of Emerging Technology and Advanced Engineering, 3(7), 390-393.

P Anchalia, Prajesh, & Kaushik R. (2015). The K-Nearest Neighbor Algorithm Using MapReduce Paradigm. Fifth International Conference on Intelligent Systems, Modelling and Simulation.

H. Ishibuchi, T. Yamamoto. (2005). Rule weight specification in fuzzy rule-based classification systems. IEEE Trans. Fuzzy Syst. 13(4) 428–435.

P.C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, & G. Lapis. (2011) Understanding Big Data - Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media 1st edition Book.

Charurvedi. (2008). Soft computing: Techniques and its applications in electrical engineering. Springer, Germany.

L.A. Zadeh. (1975). The concept of a linguistic variable and its application to approximate Journal of Information Sciences 8(3) 199-249.

E. H. Mamdani, S. Assilian. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7(1) 1-13

X. Wu, X. Zhu, G.-Q. Wu, and W. Ding. (2014) Data mining with big data. IEEE Trans. Knowl. Data Eng., 26(1):97-107

SMITHA T, MCA, M.Phil, V. Suresh K., M.Tech. (2013). Application of Big Data in Data Mining International Journal of Emerging Technology and Advanced Engineering. 3(7)

Foster, P. and Tom F. (2013) Data Science for Business. What you need to know about data mining and data-analytic thinking. O 'Reilly Media, 1st edition Book.

Barbierato, E., Gribaudo, M., & Iacono, M. (2014). Performance evaluation of NoSQL big-data applications using multi-formalismmodels. Future Generation Computer Systems, 37, 345-353.

Kirubakaran R., Periya N., C.Mano P. (2015). A Survey on Data Mining in Big Data, International Journal of Engineering Development and Research, IJRSI, 3(1).

Nour E. Oweis1,4, Suhail S. Oweis, George, Waseem & Suliman. (2015). A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses). 10.1007/978-3-319-21206-7_10.

Big Data Mining https://www.techopedia.com/definition/30215/big-data-mining

Tom White. (2013). Hadoop The Definitive Guide O'Reilly Book 3th edition.

Río, Sara & López, Victoria & Benítez, José & Herrera, Francisco. (2015). A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules. International Journal of Computational Intelligence Systems. 8. 422-437.

Akil Kumar A.1, Mithunkumar P. , Kiruthiga R., Anitha R., V. Priya,. (2015) Improved Fuzzy Rule-based Classification System Using Feature Selection and Bagging for Large Datasets. International Journal of Science and Research (IJSR), 6(3). 1206-1209.

López, Victoria & Río, Sara & Benítez, José & Herrera, Francisco. (2015). Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. Fuzzy Sets and Systems. 258. 5-38.

Elkano, Mikel, Galar, M., Sanz, J. & Sola, H. (2017). CHI-BD: A Fuzzy Rule-Based Classification System for Big Data classification problems. Fuzzy Sets and Systems. 348.

Fernández, A., Almansa, E., Herrera, F. (2017). Chi-Spark-RS: An Spark-built evolutionary fuzzy rule selection algorithm in imbalanced classification for big data problems. 1-6.
Elkano, Mikel, Sanz, José, Barrenechea, E., Sola, H. & Galar, Mikel. (2019). CFM-BD: a distributed rule induction algorithm for building Compact Fuzzy Models in Big Data classification problems. IEEE Transactions on Fuzzy Systems.
Eseye, Abinet & Zhang, J. & Zheng, D. & Shiferaw, Dereje. (2016). Short-Term Wind Power Forecasting Using Artificial Neural Networks for Resource Scheduling in Microgrids. International Journal of Science and Engineering Applications. 5. 144-151.
Herrera, F. (2016). Data Mining and Soft Computing, Academic Book at Dept. of Computer Science and A.I. University of Granada.
Fernnádez, A., Altalhi, A., Alshomrani, S. & Herrera, F. (2017). Why Linguistic Fuzzy Rule Based Classification Systems perform well in Big Data Applications?. International Journal of Computational Intelligence Systems. 10. 1211.
Zaharia, M., Chowdhury, Mridul, Das, Tathagata, Dave, A. & Ma, J. & McCauly, M. & Franklin, M.J. & Shenker, S. & Stoica, Ion. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In NSDI. 15-28.
Jogalekar, P.,& Woodside, M. (2000). Evaluating the scalability of distributed systems. Parallel and Distributed Systems, IEEE Transactions on. 11. 589 - 603.
Hashmi, A. & Ahmad, T. (2016). Big Data Mining: Tools & Algorithms. International Journal of Recent Contributions from Engineering, Science & IT (iJES). 4. 36.
C. Gallo. (2015). Artificial Neural Networks: tutorial, Encyclopedia of Information Science and Technology Edition: 3rd Ed.

An intelligent Diagnostic System for Renal Failure Detection

Ehab Shams El Deen Yaqut Aly, Shahira Shaaban Azab & Hesham Ahmed Hefney,

Faculty of Graduate Studies For Statistical Research Cairo University.

Abstract:

Kidney failure occurs suddenly without any symptoms, except in the final stage. Unfortunately, the disease in the final stage is difficult to treat. Thus, Kidney failure contributes to deaths, either directly or indirectly. Therefore, early detection of kidney disease contributes to reducing the probability of renal failure. Kidney failure has a high mortality specially in developing countries.

In this paper, we use a machine learning technique with a dataset from Aswan - Edfu General Hospital. The results are promising: Support vector machine with 98% and logistic regression 98% accuracy rate. The results of these experiments assure that a machine learning prediction tools may offer important prognostic capabilities for determining which patients are likely to have kidney failure, allowing to take steps for treatment before kidney damage manifests.

Keywords: AI, Machine Learning, Classification, Kidney Disease, KCD.

Introduction:

It was necessary to contribute to solving the problems facing our societies as possible, and one of the most important problems facing us is health problems. The devices such as X-ray, CT, sonar, and ECG have evolved. However, despite all this development, the machines were powerless, because they are do not diagnose or treat, and the person who controls the diagnosis and treatment is the doctor. Currently, with the development of machine learning, the machine has been able to help in diagnosing, treating and performing some surgeries.

In the latest years, machine learning techniques have been successfully applied in many fields [Ba, et al., 2014]. This success is due to the recent research in the field [Mnih, et al., 2015]. Large data sets are available to address problems in these fields [Russakovsky, et al., 2015], And the development of software platforms that allow easy use of large amounts of computer resources to train such models on these large data sets [Chilimbi, et al., 2014].
Chronic kidney disease (CKD) is an emerging global health problem. The increasing incidence of chronic kidney failure justifies the need for an approach to better understanding and prevention of the disease. While there is not enough information on how to prevent and investigate this disease better. There is evidence of global distribution indicating that chronic

kidney failure is a growing issue in developed and developing countries, while screening and intervention can prevent chronic kidney disease [Okorie, et al., 2018]

Machine Learning.

The ability of ML tools to detect key features from complex datasets reveals their importance [Kourou, et al., 2015].

Classification.

Classification is the process of dividing the dataset into different categories or groups has label. In classification learning these are the components we have a task T which has input and output, there is a performance metric P and there is experience E and the task T comprises of input which is a set of instances each instance has a set of features , and we have to output a set of predictions for the inputs. now we will look at some sample classification, tasks, let us say the task is medical diagnosis the instance our patient record comprising of the features of Age , Pedal Edema , Appetite, Coronary Artery Disease, Diabetes Mellitus , Hypertension , blood pressure , Red Blood Cell Count, White Blood Cell Count , Packed Cell Volume, Hemoglobin, Anemia, Potassium, Sodium, Serum Creatinine, Blood Urea, Blood Glucose Random, Bacteria, Pus Cell Clumps, Pus Cell, Red Blood Cells, Sugar, Albumin, Specific Gravity. and the labels are whether the person has is chronic kidney disease or not chronic kidney disease.

Features:

features play a very important role so observations that you see are analyzed into a set of quantifiable properties which are called features, so if you want to predict the price of a car or the price of a house you have to have the right features which will help us to come up with the price, if we want to predict what disease a patient as you have to come up with the right tests and parameters which will enable us to make that prediction now, features are of various types, features can be categorical, for example Appetite feature which has certain values like Yes-No. Anemia is a feature that has certain values like Yes-No etc. our features can be ordinal like large medium small, features can be integer-valued or a feature can be real valued.

Experimental Result: (Materials and methods)

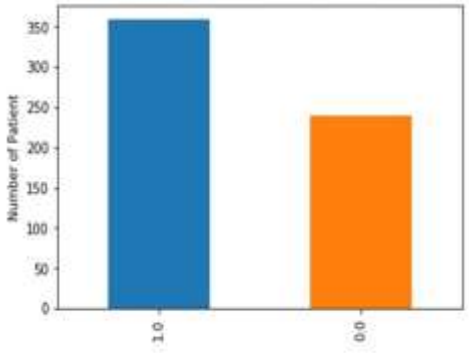
Dataset :

The dataset for diagnosis of chronic kidney disease is obtained from the medical records of 600 patients from Edfu General Hospital kidney department laboratories in Aswan Governorate. There are 600 instances with 25 different features as shown in table(1).

Table (1) The Attributes and the normal rates.

Attribute	Information
Age	Age in years
Blood Pressure	Blood Pressure in mm/Hg
Specific Gravity	Specific Gravity (1.005,1.010,1.015,1.020,1.025)
Albumin	Albumin - (0,1,2,3,4,5)
Sugar	Sugar - (0,1,2,3,4,5)
Red Blood Cells	Red Blood Cells - (normal,abnormal)
Pus Cell	Pus Cell - (normal,abnormal)
Pus Cell clumps	Pus Cell clumps - (present,notpresent)
Bacteria	Bacteria - - (present,notpresent)
Blood Glucose Random	.Blood Glucose Random - in mgs/dl
Blood Urea	Blood Urea - in mgs/dl
Serum Creatinine	Serum Creatinine - in mgs/dl
Sodium	Sodium - in mEq/L
Potassium	Potassium - in mEq/L
Hemoglobin	Hemoglobin - in gms
Packed Cell Volume	Packed Cell Volume
White Blood Cell Count	White Blood Cell Count- in cells/cumm
Red Blood Cell Count	Red Blood Cell Count - in millions/cmm
Hypertension	Hypertension - (yes,no)
Diabetes Mellitus	Diabetes Mellitus - (yes,no)
Coronary Artery Disease	Coronary Artery Disease - (yes,no)
Appetite	Appetite - (good,poor)
Pedal Edema	Pedal Edema - (yes,no)
Anemia	Anemia - - (yes,no)

Graphical display such as the histogram as shown in Fig (1) gives us a rough idea on the whole, very informative and clear to balance data .



Fig(1)

Chronic and Not Chronic Patient

The Correlations

The measures of association examined so far are useful for describing the nature of association between two variables which are measured at no more than the nominal scale of measurement [Ba, et al., 2014].

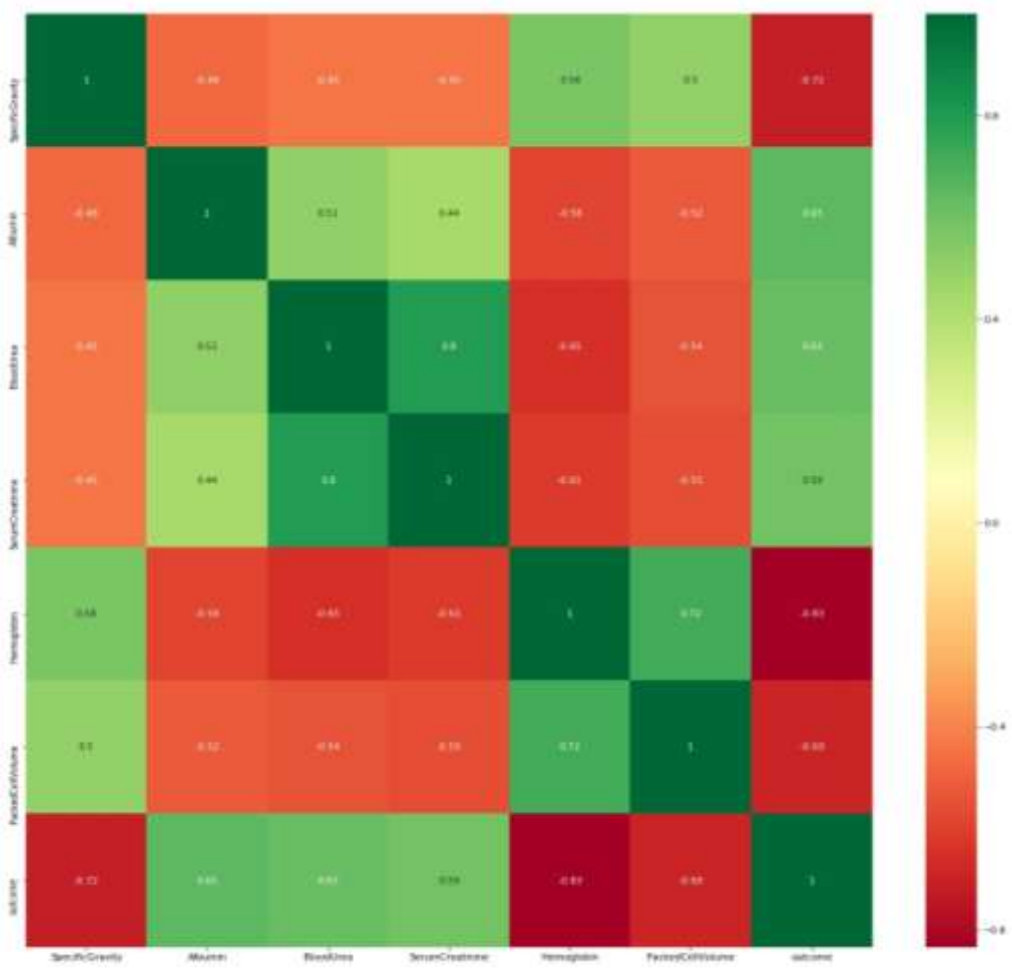


Fig (2)

Heat Map For Give Correlations Between Feature

Feature selection:

Feature selection is a method of selecting some features out of the data and discarding the irrelevant ones.

We have created a procedure that contains a function that gives the correlation between each feature and the results and we applied this procedure to data set and the results are as follows:

Feature selection from Edfu data set as the following :

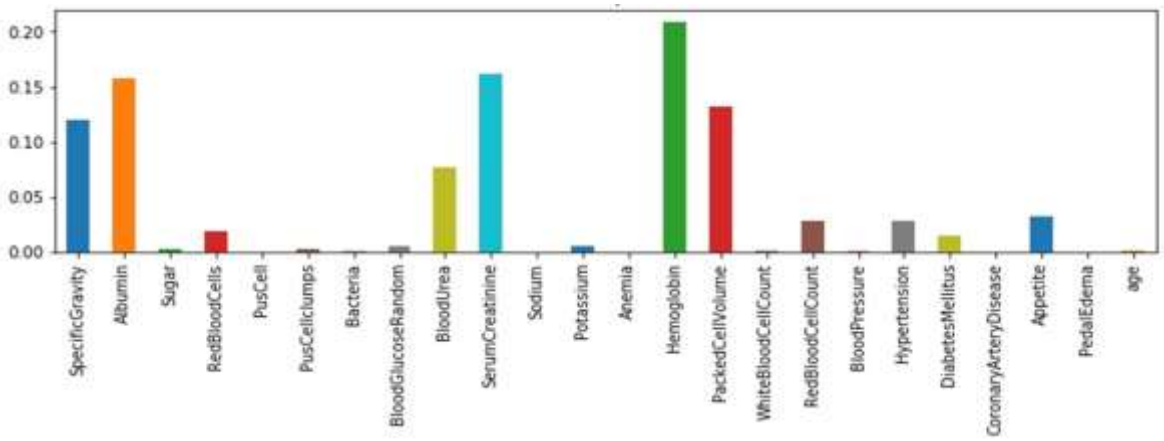


Fig (3)

Feature Important Edfu Dataset

We chose the features as the follows

specific gravity, Albumin, Blood Urea, Serum Creatinine, Hemoglobin, Packed Cell Volume

Training Test Split:

We used only 20 % for test and 80% for training from dataset, we applied all pre-processing procedures for all applications, applied all measures for all applications.

The results:

Logistic Regression

Logistic regression is a generalized by linear regression. It is mainly used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modeled directly by linear regression.[Mnih, et al., 2015]. Logistic regression is a generalization of linear regression.[Russakovsky, et al., 2015]

Confusion Matrix:

Confusion Matrix is a table that describes the performance of a classification model. the confusion matrix function is also available in the metrics module and this case outputs a 2x2 matrix.
The size is 2 by 2 because this is a binary classification problem, so if there were 3 possible response classes this would be a 3 by 3 matrix, etc.
Another set of metrics that can be applied to the decision problem has been developed by diagnosticians (medicine, finance, etc.) and is based on the so-called confusion matrix [Candy & Breitfeller, 2013]

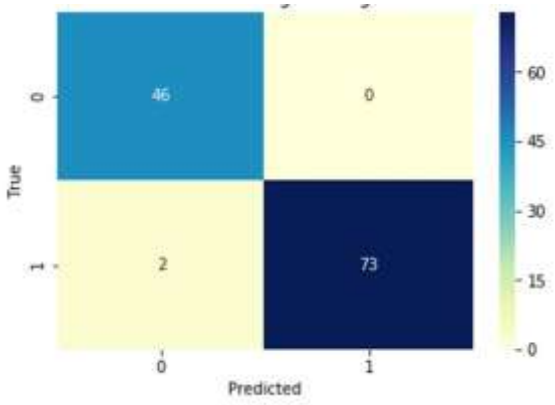


Fig (4)

Confusion Matrix (Logistic Regression)

Table 2 determine value of every measures we used as the following :

Classification Accuracy :

Overall how often is the classifier correct, classification accuracy can be calculated from the confusion matrix, you add the true positives and true negatives and divide that by the total number of observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad 1$$

Classification Error:

Also known as the misclassification rate, it is equal to the false positives plus the false negatives divided by the total or one minus the accuracy score.

.

$$\text{Classification Error} = \frac{FP + FN}{TP + TN + FP + FN} \quad 2$$

Or

Classification Error = 1 - Accuracy Score

Sensitivity:

Which answers the question when the actual value is positive. how often is the prediction correct?. sensitivity is calculated by dividing the true positives (TP) by the total (FN+TP).To me, the term sensitivity makes intuitive sense because it's a calculation of how sensitive the classifier is to detecting positive instances, however, it's also known by the terms true positive rate and recall, which of these terms is most commonly used is largely dependent on the field of study [Powers, 2011].

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad 3$$

Specificity :
Specificity is which answers the question when the actual value is negative how often is the prediction correct?
just like sensitivity, specificity is something we want to maximize examining this sample once again, specificity is calculated by dividing the true negatives by the total of the (TN+FP) [Tape, 2007].

Specificity is thus the ratio of Real Negative cases that are correctly Predicted Negative and is also known as the True Negative Rate (TNR). Conversely, Inverse Precision is the proportion of Predicted Negative cases that are indeed Real Negatives, and can also be called True Negative Accuracy (TNA) [Powers, 2011].

$$\text{Specificity} = \frac{TN}{TN + FP} \quad 4$$

False Positive Rate :

It answers the question when the actual value is negative how often is the prediction incorrect, it is calculated by this way. the false positives (FP) divided by the total of(TN+FP), also false positive rate is simply 1 minus specificity

$$\text{False Positive Rate} = \frac{FP}{TN + FP} \quad 5$$

OR

False Positive Rate = 1- Specificity

Precision:
it answers the question when a positive value is predicted how often is the prediction correct?
It is calculated by dividing true positives (TP) by the total of the predicted (TP+FP). That many other metrics can be calculated from the confusion matrix such as the f1 score and Matthews correlation coefficient [Powers, 2011].
Note : we always examine the confusion matrix for our classifier since it gives us a more complete picture of how our classifier is performing it, also allows us to compute various classification metrics which can guide us model selection process however us can't optimize our model for every one of these metrics so how do we choose between them the choice of metric ultimately depends on our business objective.

$$\text{Precision} = \frac{TP}{TP + FP} \quad 6$$

Measure

The F-Measure is the harmonic mean of precision and recall

$$F = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad 7$$

Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is used in machine learning as a measure of the quality of binary.

$$MCC = \frac{(TP* TN)-(FP * FN)}{\sqrt{(TP + FP) * (TP * FN)*(TN* FP)*(TN * FN)}} \quad 8$$

Table (2) Logistic Regression Results for Edfu Dataset (Feature Selection)

Accuracy	0.983
Classification Error or misclassification Rate	0.0165
Sensitivity True Positive Rate "or" Recall	0.973
Specificity	1.000
False Positive Rate	0.000
Precision	1.000
F Measure	0.986
Matthews Correlation Coefficient (MCC)	0.965
False Negative Rate	0.027

Logistic Regression Results for Edfu Dataset by chart as Fig(5)

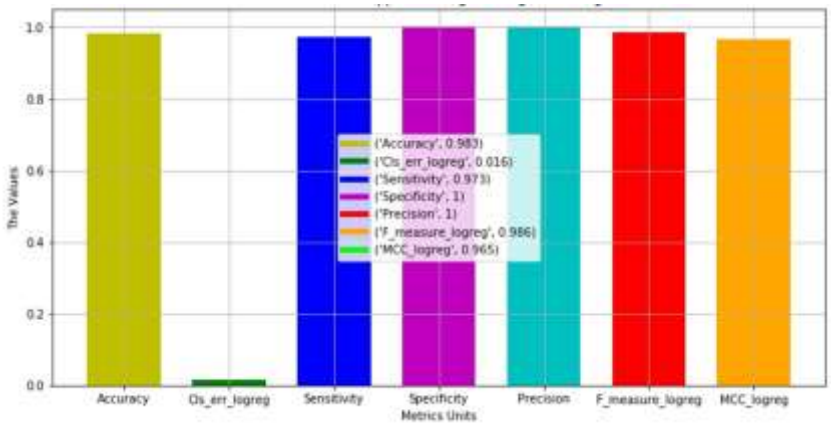


Fig (5)

Different Measures Applied to (Logistic Regression)

False Positive Rate and False Negative Rate For Confusion Matrix (Logistic Regression) as Fig(6)

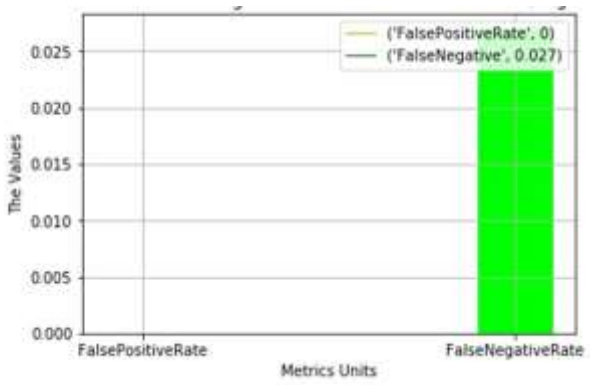


Fig (6)

False Positive Rate and False Negative Rate For Confusion Matrix (Logistic Regression)

Support Vector Machine

Support Vector Machines (SVM) is a powerful, state-of-the-art algorithm based on linear and nonlinear regression [Sinha & Sinha, 2015].

Confusion Matrix (Support Vector Machine) as shown in Fig (7)

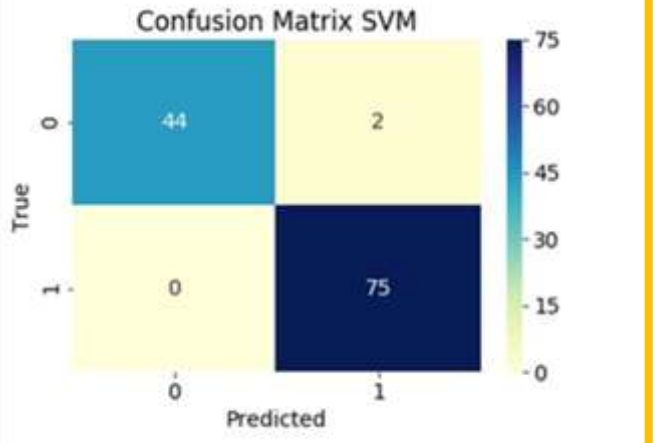


Fig (7)

Confusion Matrix (Support Vector Machine)

Table (3) Support Vector Machine Results for Edfu Dataset (Feature Selections)

Accuracy	0.983
Classification Error or misclassification Rate	0.016
Sensitivity True Positive Rate "or" Recall	1.000
Specificity	0.9583
False Positive Rate	0.042
Precision	0.973
F Measure	0.986
Matthews Correlation Coefficient (MCC)	0.966
False Negative	0

Results of SVM for Edfu Dataset by chart as Fig(8)

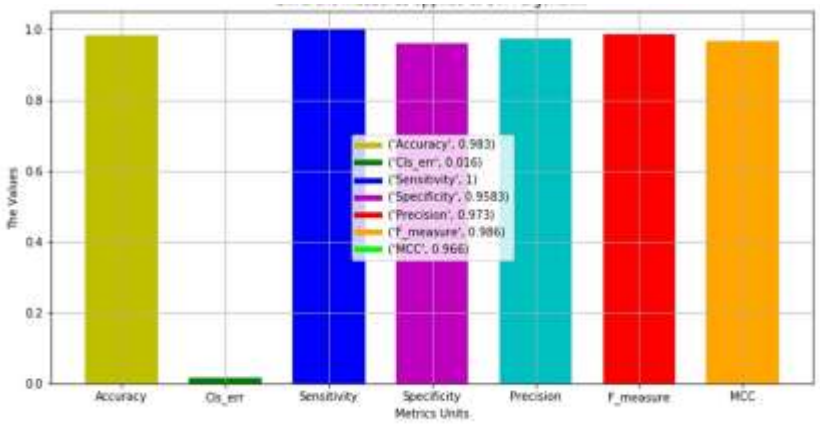


Fig (8)

Different Measures Applied to (SVM)

False Positive Rate and False Negative Rate For Confusion Matrix (SVM) as in Fig (9)

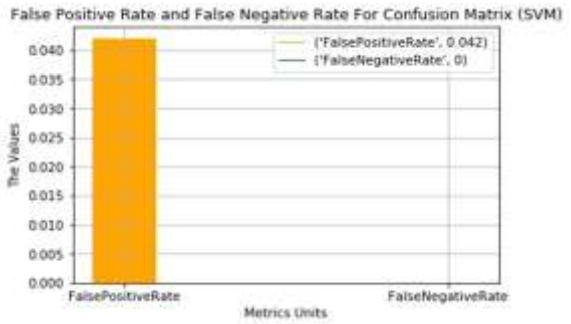


Fig (9)

False Positive Rate and False Negative Rate For Confusion Matrix (SVM)

Table (4) The Comparisons Between Logistic Regression and Support Vector Machine:

The Metrics	The Algorithms	
	Logistic Regression	Support Vector Machine
Accuracy	0.983	0.983
Classification Error	0.0165	0.016
Sensitivity	0.973	1.000
Specificity	1.000	0.9583
False Positive Rate	0.000	0.042
Precision	1.000	0.973
F Measure	0.986	0.986
Matthews Correlation Coefficient	0.965	0.966
False Negative Rate	0.027	0

fig (10) show the comparison between Logistic Regression and SVM For Accuracy

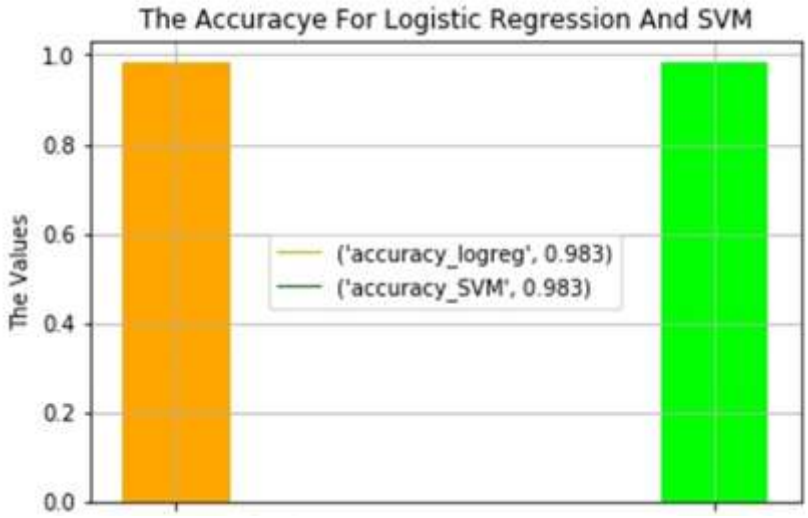


Fig (10)

Comparison between Logistic Regression and SVM For Accuracy

Comparison between Logistic Regression and SVM For False Negative Rate

Fig (11)

False Positive Rate and False Negative Rate For Confusion Matrix (SVM)

Discussion :

Metrics Computed From Confusion Matrix: as we know classification accuracy overall, How is often the classifier correct?

Classification accuracy is the easiest classification matric to understanding, but it does not tell us the underlying distribution of response values, and it does not tell us that types of error our classifier is making. This will be illustrated by the following:-

Both of Logistic Regression and SVM are has Accuracy equal 0.983

Classification Error :

Also Known as misclassification Rate. Also this scale does not show where the errors are, is in False Positive Rate or False Negative Rate.

Both of Logistic Regression and SVM are has Classification Error equal 0.0165

Author Contribution:

Basic Terminology of Confusion Matrix:

True Positive (TP) it Mean We Correctly Predicted That They do have a CKD
True Negative (TN) it Mean We Correctly Predicted That They do have a NotCKD
False Positive (FP) it Mean We Incorrectly Predicted That They don't have a CKD
False Negative (FN) it Mean We Incorrectly Predicted That They don't have a NotCKD

If We look at the previous terminology we note that False Negative it Mean We Incorrectly Predicted That They don't a Patients , Even though they are sick, This means that we do not perform the treatment, This leads to deterioration of cases or destruction of the kidneys, so It is a very important measure of which decisions are made.

Based on The Above :

We are given utmost importance to False Negative Rate , and consider it the primary measure of kidney Failure.

Conclusion:

This paper presented a Medical aided diagnostic system for detection of kidney failure, using two machine learning Techniques namely logistic regression, Support vector machine. The results of the experiments using the medical records of 600 patients from Aswan - Edfu General Hospital, Egypt The prediction algorithm achieves the Support vector machine 99% and logistic regression 98%.

References

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014).“Multiple Object Recognition with Visual Attention,” pp. 1–10.

Mnih, V. *et al.*, (2015). “Human-level control through deep reinforcement learning.,”*Nature*, vol. 518, no. 7540, pp. 529–33.

Russakovsky, O. *et al.*, (2015).“ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252.

Chilimbi, T., Suzue Y., Apacible, J., & Kalyanaraman, K. (2014). “Project Adam: Building an Efficient and Scalable Deep Learning Training System,” *Proc. 11th USENIX Symp. Oper. Syst. Des. Implement.*, pp. 571–582.

Okorie. C *et al.*, (2018). “Epidemiology and management of chronic renal failure: a global public health problem,” *Biostat. Epidemiol. Int. J.*, vol. 1, no. 1, pp. 11–16,

Kourou, K., Exarchos, T.P.,Exarchos, K.P., Karamouzis, M. V. & Fotiadis, D. I. (2015).“Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17.

Gingrich, P. “ Association Between Variables,” *Introd. Stat. Soc. Sci. - http//uregina.ca/~gingrich/text.htm*, pp. 794–835, 2004.

Kumar, K., & Abhishek, A. (2012).“Artificial Neural Networks for Diagnosis of Kidney Stones Disease,” *Int. J. Inf. Technol. Comput.* .

Lakshmi, K. R., Nagesh, Y., & Veerakrishna, M. (2014), “Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability,” *Int. J. Adv. Eng. Technol.*, vol. 7, no. 1, pp. 242–254.

Candy, J. V. & Breitfeller, E. F. (2013) “Receiver Operating Characteristic (ROC) Curves : An Analysis Tool for Detection Performance,” pp. 1–27.

Powers, D. M. (2011) “Estimation of high affinity estradiol binding sites in human breast cancer evaluation: from percesion, recall and F-measure to ROC, informedness, markedness & correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63.

Tape, T. G. (2000). “Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model,” *Univ. Nebraska Med. Cent.*, pp. 1–3.

Sinha, P., & Sinha, P. (2015). “Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM,” *Int. J. Eng. Res. Technol.*, vol. 4, no. 12, pp. 608–612.

An Efficient Prediction Approach for Default Customers of Personal Loans Using Machine Learning

Mohamed H. Khedr[1] **Nesrine A. Azim**[2] **Ammar M. Ammar**[3]
[1]Department of Software Engineering, Faculty of Graduate Studies for Statistical Research (FGSSR), Cairo University, Egypt.
Mohamed.khedr@abe.com.eg
[2]Department of Information Systems and Technology, FGSSR, Cairo University, Egypt.
nesrinealiazim79@hotmail.com
[3] Department of Computer Science, FGSSR, Cairo University, Egypt.
ammar@cu.edu.eg

Abstract

In Egyptian banking sector credit approval decision for personal loans is made using a pure judgment by credit officers due to the fact that machine learning is not widely in practice yet. In this paper, we have taken upon the challenge of delivering prediction approach for default customers of personal loans using machine learning. Our main objective is to predict default customers and so we analyze the trustworthiness of customers for getting a loan through available personal data and historical credit data .we used ABE dataset for training and testing , also we used 10 features from the application form and i-score report class that could be a helpful tool to credit staffs to take the right decision to decrease credit risk in order to avoid random techniques for customer selection. The data obtained were analyzed with different machine learning classification algorithms based on certain features in order to achieve higher accuracy. We compared between several methods before and after feature selection. We have observed that the most important features are (activity – income – loan) that can lead to high accuracy and decision tree has the best performance than any other machine learning algorithms with significant prediction accuracy of almost 94.85%.

Keywords*:* Machine learning, Personal loans, Credit risk

1. Introduction

The history of banking began in Mesopotamia over 3000 years ago where the practice of lending money at interest started in Sumer; loans were made to farmers and traders who carried goods between cities. Also, there was evidence of money lending activity in ancient Greece during the Roman Empire, ancient China, ancient Egypt and India. (Graeber, 2011; Bank& Davies, 2002).

Banks in its current structure founded in the middle Ages, in the thirteenth century and fourteenth century. The first bank was named the bank was the "Bank of Barcelona" in (1401). (Idiab , Haron, & Ahmad, 2011)

During the 20th century, due to advances in communications and computing operations of banks have been changed and also an increase in size and geographical spread. Credit is defined as a period of time allowed to one party to provide a debt incurred for goods or service to another party. While credit risk is a risk that the party received credit may not be able to repay on time. (Collin, 2005).

It is sometimes called default risk, performance risk or counterparty risk. (Hasan,2016).

Personal loans mean that customers apply for household or other personal use, not for business use loans (investment purposes - enterprises), but for credit cards, buying electrical or electronic devices loans and car loans. (Collin, 2005)
Avoiding default customers of investment and enterprises loans may be easier than personal loans, because feasibility studies and financial statements are analyzed in addition to customer and company reputations in the market. For personal loans, Banks have to observe and gather accurate information about their potential borrowers and monitor the performance of accepted borrowers over time. (Hasan, 2016).
However, the development of a scoring system (in Egypt: Egyptian credit bureau I. Score company) will give information about historical credit of the customers, and also will help to avoid default customers, but it is not a trivial task due to availability, accessibility and interpretation techniques of customers' data are often scarce.
In this paper, we propose an efficient approach of applying machine learning techniques to make a prediction for default customers of personal loans, depending on analyzing data in the application form and i-score report class that could be a helpful tool to credit staffs to take the right decision and decrease credit risk. We use dataset of personal loans customers from 27 branches of ABE which are the most branches deal with personal loans. The paper compares between several methods before and after applying feature selection. The rest of this paper is organized as follows: Section 2 presents previous studies which compared performance of using machine learning techniques to predict default customers of personal loans. Section 3 introduces a background about machine learning techniques. Section 4 presents data collection and the processes that had been done in this dataset. Section 5 shows the results before and after features selection. Finally, the conclusion is in Section 6.

2. Related work

Xiao et al (2006)
They made a comparison of the performance of Logistic regression, KNN, decision tree, SVM classifiers, MARS and ANN, for classifying the loan applications.
A personal loans dataset was provided by German banks (it consists of 1000 applicants),Australian banks (it consists of 700 applicants) and The third credit data is from major financial institutions in US, where there are (it consists of 1225 applications).
The features used are including (credit history, account balances, loan purpose, loan amount, employment status, age, housing, and job.
Research findings indicate that in terms of the classification ratio the best performance classifier was Logistic regression as it yields (74.2%, 88.2%, and 63.5%).
However, it has to be noted that LDA and decision tree significantly more accurate than any other model in identifying bad credit risks for German and American credit scoring data sets.
Abdou & Pointon (2009)
They made a comparison of the performance of Discriminant analysis model (DA), Logistic regression model, and neural network models for classifying the loan applications.
A personal loans dataset was provided by one of the commercial banks in Egypt. It consists of 1262 personal loans with 851 good loans and 411 bad loans. Each bank customer in this dataset is linked to19 features, as : loan amount, loan duration, age,

gender, dependents, profession, education, house status, telephone, monthly income, CBE report, guarantees, field visit, feasibility study, and credit card status, loans from other banks, car ownership and marital status.

Research findings indicate that Neural Networks models gave a better average correct classification rate (87.64%) which is higher than the other techniques.

Ince & Aktan (2009)

They made a comparison of the performance of discriminant analysis, logistic regression, decision trees (C5, CART), and neural networks, for classifying the loan applications.

Credit card dataset provided by a Turkish bank is used. Each bank customer in the data set contains nine features, namely: gender, age, marital status, educational level, occupation, job position, income, customer type and credit cards from the other banks. The response variable is the credit status of the customer (good or bad credit). The dataset is composed of 1260 customers' records. Research findings indicate that analytic results demonstrate that CART has (65.58%) which is the best average correct classification rate in comparison with discriminant analysis, logistic regression, and neural networks. On the other hand, neural network credit scoring model has lower Type II errors (a customer with bad credit is misclassified as a customer with good credit) associated with high misclassification costs and therefore has better overall credit scoring capabilities.

Yeh & Lien (2009)

They made a comparison of the performance of discriminant analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, and classification trees for classifying the loan applications. Used dataset of one of the important banks in Taiwan and the targets were credit card holders of the bank (25,000)observations .the customer features including (Gender - Amount of the given credit- Education- Marital status- Age- History of past payment).

Research findings indicate that artificial neural networks perform classification more accurately than the other five methods with rate(54%).

Sudhakar & Reddy (2014)

They made a comparison of the performance of logistic regression, radial basis network, and multilayer perceptron model, SVM and decision tree for classifying the loan applications.

Data used is of 500 customer's data. The data was collected from different banks in India such as SBI, Andhra, ICICI and Syndicate banks. It consists of different 10 independent features (Age of customer- Sex- Marital status- Service period- Current account- Saving account- Payment history- Occupation- Home ownership- Address time) and one dependent variable(Credit (Approved or Not).

Research findings indicate that accuracy for SVM(85.97%), decision tree(84.77%) and radial basis network (86.53%) which are the best methodologies for classifying the loan applications. By analyzing the performance of these models on standard dataset it is found that in case of missing data multilayer perceptron model and logistic regression is also good. The recommendation is to develop new integrated model which takes advantages of all the five models.

Sousa & Figueiredo (2014)

They made a comparison of the performance of Decision Tree and Neural Network for classifying the loan applications. The data were collected of 211 individual members from Credit Union of the SICOOB in Brazil.The study used 26 features, such as (gender-

age – level of education- birth place-place of residence-work or activity-Marital status-Years of experience in activity/work).
Research findings indicate that the decision tree's performance for the problem is (94.08%) which is better than Neural Network(91.59%).

Jafar & Ahmed (2016)

They made a comparison of the performance of decision tree (j48), bayes Net and naïve Bayes for classifying the loan applications. A collection of data from banking sector in Sudan (1000 instances) has been selected using 7 features (Credit history, Purpose, Gender, Credit amount, Age, Housing, Job and one dependent variable (the loan Class (good –bad).
Research findings indicate that applying classification's data mining techniques algorithms which are decision tree (j48), bayes Net and naïve Bayes ,found that the best algorithm for loan classification is decision tree (j48)algorithm (78.38%) , because it has high accuracy and low mean absolute error.

Based on the above mentioned studies, it is useful to highlight the following points:-

1- They did not use large datasets (not exceeding 25,000 customers).
2- Most of them did not compare performance of wide variety of classification algorithms.
3- They did not compare performance of algorithms before and after feature selection.

3. Background

Term of machine learning stands for learning from past experience (which in this case is previous data) to improve future performance. (Das & Behera, 2017)
Machine Learning algorithms can be grouped by learning style into the following categories:

Supervised learning: Input data has a pre-determined label e.g. True/False, Positive/Negative, and Spam/Not Spam, etc. a classifier is built and trained to predict the labels of the test data.

Unsupervised learning: Input data is not labeled. A classifier is built by deducing existing patterns or clusters in the training datasets.

Semi-supervised learning: The training dataset contains both labeled and unlabeled data. The classifiers train to learn the patterns to classify and label the data as well as to predict.

Reinforcement learning: The algorithm is trained to map action to a situation so that the feedback signal is maximized. The classifier is not programmed directly to choose the action but instead trained to find the most rewarding actions by trial and error.

Transduction: it tries to predict the output based on training data, training labels, and test data. However, it shares similar traits with supervised learning, but it does not develop an explicit classifier.

Learning to learn: The classifier is trained to learn from the bias it induced during previous stages. (Das & Behera, 2017)

3.1. Classification Algorithms:

Different machine learning techniques for classification exist, for example, K-nearest neighbor classifiers, Logistic regression, Discriminant analysis, Naive Bayesian classifier, SVM, Artificial neural networks and Classification trees.

3.1.1 K-nearest neighbor (KNN) classifiers: the training data (well-labeled) is fed into the classifier. When the test data is introduced to the classifier, it compares both of the data. k most correlated data is taken from the training set.(Dey , 2016)

3.1.2 Logistic regression: it can predict a discrete outcome from a set of variables which may be continuous, discrete, dichotomous, or a mix of any of these. (Keramati & Yousefi ,2011)

3.1.3 Discriminate analysis: it is an alternative to logistic regression and is based on the assumptions that, for each given class of response variable, the explanatory variables are distributed as a multivariate normal distribution with a common variance-covariance matrix. (Keramati & Yousefi ,2011)

3.1.4 The naive Bayes classifier: it is a probabilistic classifier that assumes the existence or nonexistence of a specific feature of a class is unrelated to the existence or nonexistence of any other feature. (Keramati & Yousefi ,2011)

3.1.5 Support vector machine: it determines an optimal separating hyper plane that separates the two classes of data points justly and is equidistant from both of them.(Das & Behera ,2017)

3.1.6 Decision tree: type of trees which groups attribute by sorting them based on their values. Each tree consists of nodes and branches.

Each node represents attributes in a group that is to be classified and each branch represents a value that the node can take. (Dey, 2016)

3.1.7 Random Forest: is a generalization of recursive partitioning that combines a collection of trees called an ensemble. (Clarke, Fokoue , & Zhang ,2009, p.254).

3.1.8 Artificial neural networks: based on the function of the human brain. it is a powerful tool for unknown data relationship modeling. it is able to recognize the complex pattern between input and output variables then predict the outcome of new independent input data. (Keramati & Yousefi ,2011)

4. Proposed Work

4.1. Dataset Description:

we collected a dataset provides the ABE bank customers' information including 122,572 records and 11 fields of customers who have applied for a personal loan to the ABE bank for five years from 2013 to 2018, with loan duration ranging from 12 to 60 months.

The dataset was collected from 27 branches of ABE bank which are the most branches deal with personal loans including (Cairo –Qaliobia – Alexandria – Matrouh - Kafr Elsheikh- Damietta -Ismailia- port said - Behira – Giza – Dakahlia –Suez –Sharkia – Gharbia – Menofia - North of Sinai –South of Sinai – Luxor –Red sea - Fayoum – Minia –Beni swif- Sohag – Qena– Assiut – New valley – Aswan).

This dataset contained both categorical attributes and numeric ones as described below:

1- Address: the town or city in which customer lives (categorical).
2- Activity: type of job or activity of the customer (categorical).

3- Gender: sex type of the customer (categorical)
4- Marital: Marital Status of the customer (categorical)
5- Age: age of the customer (numeric)
6- Income: amount of customer annual income (categorical)
7- Loan: amount of loan requested for (numeric).
8- Duration: time of loan requested for (categorical)
9- Type: purpose of loan requested for (categorical)
10- I-score: customer credit score report status (categorical)
11- Status: Current credit status of a consumer (categorical)

4.2. Dataset preprocessing:

We made sure that there are no missing values as it could affect the training of the model. For the dataset preprocessing we used Python library scikit-learn (Sklearn).
several steps have been done to preprocess The dataset :-

4.2.1. Feature Encoding:

dataset contained values in form of strings and numbers. Since the machine learning models cannot run datasets with string attributes, we converted every categorical feature into numerical ones.
We used LabelEncoder for every categorical feature. We used the python library 'numpy' to label the dataset after converting outputs into numerical values.

4.2.2. Feature Scaling:

In order to train and test the data by the machine learning algorithms, we needed to scale the data down to such a range which will be easily evaluated by the algorithms and will be clearly understandable.
The continuous attributes don't mean any effect on calculations, so we used MinMaxScaler to scale the data.

4.3. Training and testing dataset

We split our data into two subsets:
(1) Training Data.
(2) Testing Data.
This is done only to fit our model into training data and make predictions on the test data.
After going through all previous steps, we would want to avoid both over-fitting and under-fitting since they both will affect the predictability. Due to over fitting the model will be very accurate on training data, but will perform poorly on new entries or the data which were not trained. This happens because the model is not generalized anywhere which means anyone can now generalize the outcomes and from here can't make any inferences on other data.
On the contrary, when the model being under-fitted means the model doesn't fit into the training data and thus misses out on the pattern of the data. So it can't be generalized to new entries as well. Therefore, to avoid both the scenarios we can do Train/test-split.
Using scikit-learn library and specifically the train_test_split method we can split our dataset into desired ratios. test_size=0.2 inside the train_test_split function describes that the split ratio is 80-20 , which means 80% of data is assigned for training the model and the rest 20% is assigned for testing the model.

We will train and test dataset using seven classification algorithms including:

1- Logistic regression
2- Random forest
3- Support vector machine
4- Naïve bayes
5- Discriminant analysis
6- Decision tree
7- K nearest neighbor

4.4. Feature Selection

Selection of attributes is a critical and important point because ,we need to be careful enough in case of including redundant attributes, as redundant features can skew the predictions.

Due to the presence of many categorical attributes, feature selection process turns into a difficult one.

We used model selection sklearn to select the attributes, but each time the set of selected attributes was slightly different from the previous set of selected attributes, that's why, we selected the features with the highest consensus among all the rankers.

We could determine the selected set of features after comparing importance indices from (logistic regression – random forest – naïve bayes- decision tree) :

Table1: importance indices of features using model selection from sklearn

Feature /Algorithm	Random forest	Logistic regression	Naïve bayes	Decision tree
Address	False	False	False	False
Activity	True	False	True	True
Gender	False	False	False	False
Marital	False	False	True	False
Age	True	False	False	False
Income	True	True	False	True
Loan	True	True	False	True
Duration	False	True	False	False
Type	False	True	True	False
I-score	False	False	False	False

According to the previous table, the selected set of features after comparing importance indices from the previous algorithms will be:

1- Activity
2- Income
3- Loan

5. Experimental Results

To measure the performance of every algorithm we can use a wide range of performance measurements in sklearn library:-

5.1. Precision: is the number of correct positive divided by the number of all positive results.

We can obtain the precision score from sklearn, which takes as inputs the actual labels and the predicted labels.

Table 2: precision results before feature selection

Algorithm	Precision
Logistic Regression	86.49%
Random Forest	95.12%
Support Vector Machine	86.05%
Naïve Bayes	86.72%
Discriminant Analysis	88.89%
Decision Tree	94.99%
K-Nearest Neighbor	94.02%

Table 3: precision results after feature selection

Algorithm	Precision
Logistic Regression	86.74%
Random Forest	96,68%
Support Vector Machine	86.87%
Naïve Bayes	86,76%
Discriminant Analysis	86,76%
Decision Tree	96,70%
K-Nearest Neighbor	95,60%

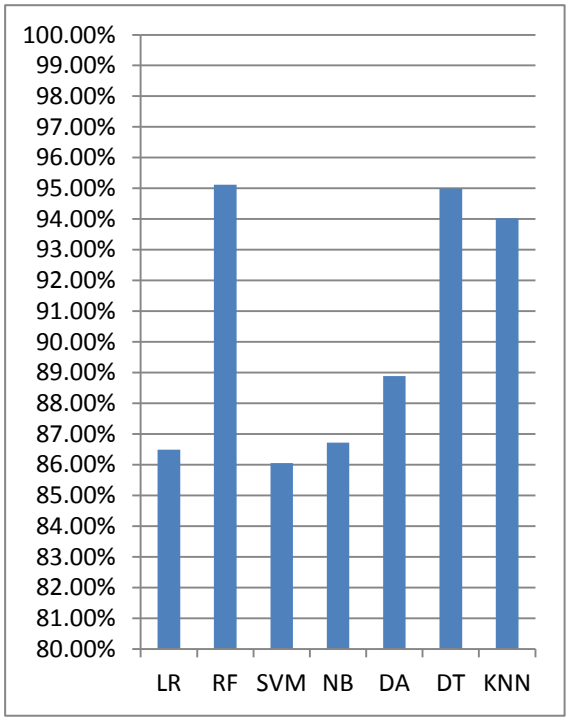


Figure 1: Precision before feature selection

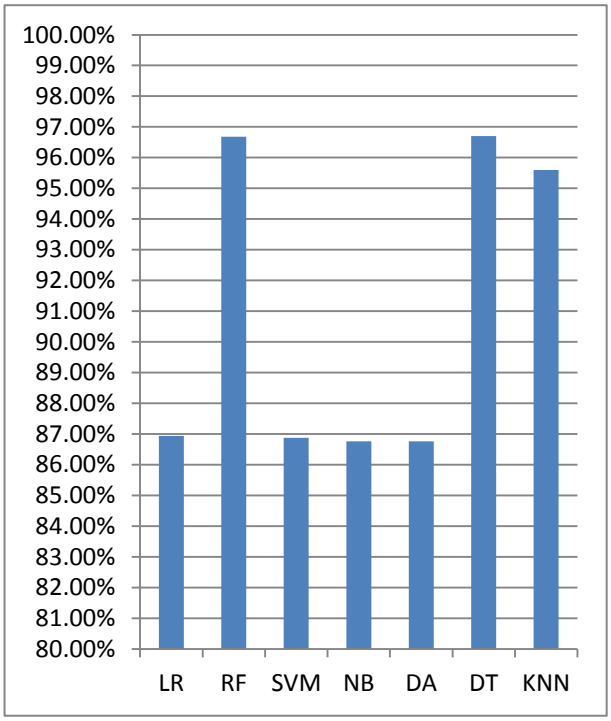


Figure 2: Precision after feature selection

5.2. Recall: is the number of correct positive results divided by all samples should have been identified as positive.

We can obtain the accuracy score from sklearn, which takes as inputs the actual labels and the predicted labels.

Table4: recall before feature selection

Algorithm	Recall
Logistic Regression	**94.69%**
Random Forest	**95.49%**
Support Vector Machine	**94.52%**
Naïve Bayes	**94.51%**
Discriminant Analysis	**92.06%**
Decision Tree	**93.74%**
K-Nearest Neighbor	**95.34%**

Table5: recall after feature selection

Algorithm	Recall
Logistic Regression	**94,45%**
Random Forest	**96,80%**
Support Vector Machine	**99,99%**
Naïve Bayes	**94,45%**
Discriminant Analysis	**94,45%**
Decision Tree	**96,81%**
K-Nearest Neighbor	**97,05%**

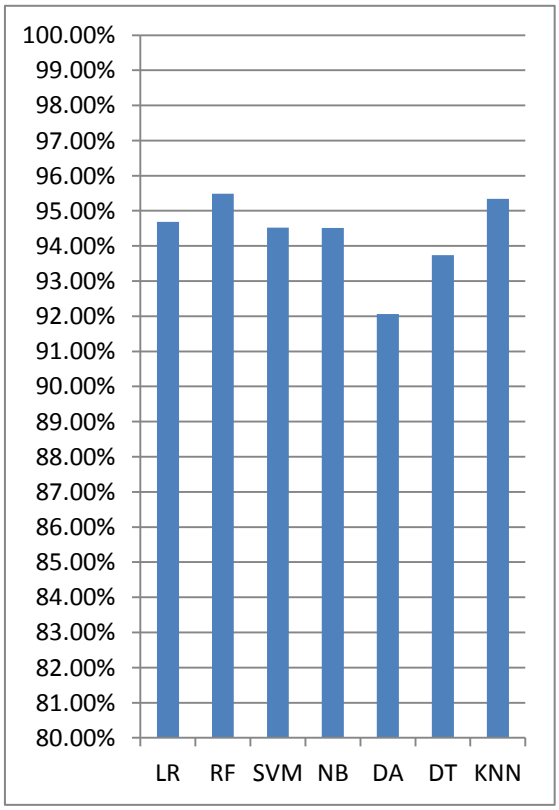


Figure 3: recall before feature selection

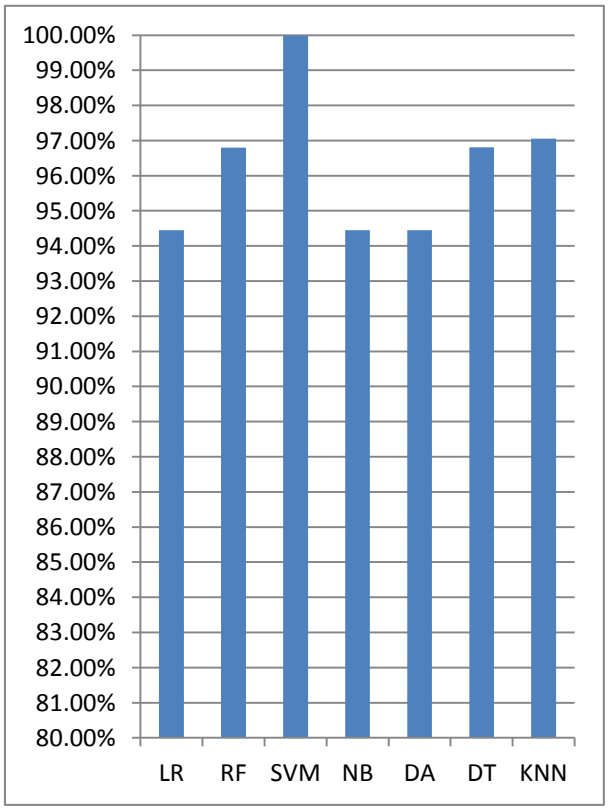


Figure 4: recall after feature selection

5.3. Classification Accuracy: is the ratio of number of correct results to the total number of samples.

We can obtain the accuracy score from sklearn, which takes as inputs the actual labels and the predicted labels.

Table6: accuracy before feature selection

Algorithm	Accuracy
Logistic Regression	**84.08%**
Random Forest	**92.54%**
Support Vector Machine	**83.52%**
Naïve Bayes	**84.19%**
Discriminant Analysis	**84.59%**
Decision Tree	**91.12%**
K-Nearest Neighbor	**91.51%**

Table7: accuracy after feature selection

Algorithm	accuracy
Logistic Regression	**84,16%**
Random Forest	**94,83%**
Support Vector Machine	**88,02%**
Naïve Bayes	**84,18%**
Discriminant Analysis	**84,18%**
Decision Tree	**94,85%**
K-Nearest Neighbor	**94,13%**

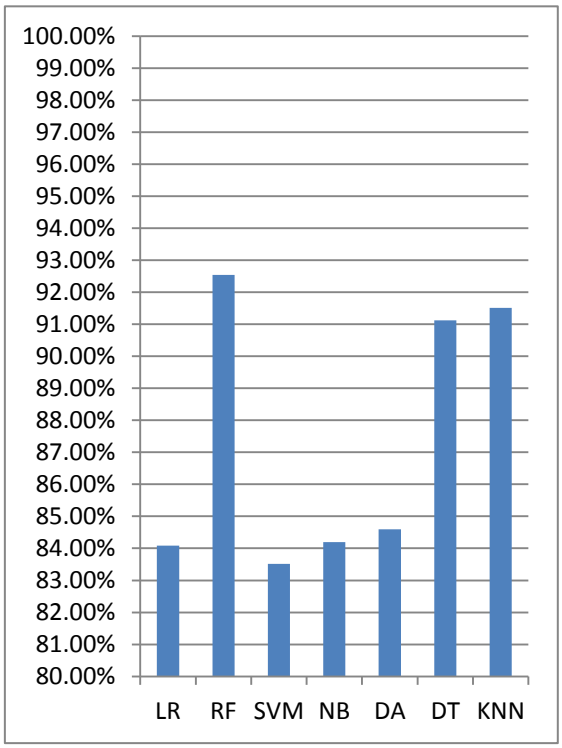


Figure 5: Accuracy before feature selection

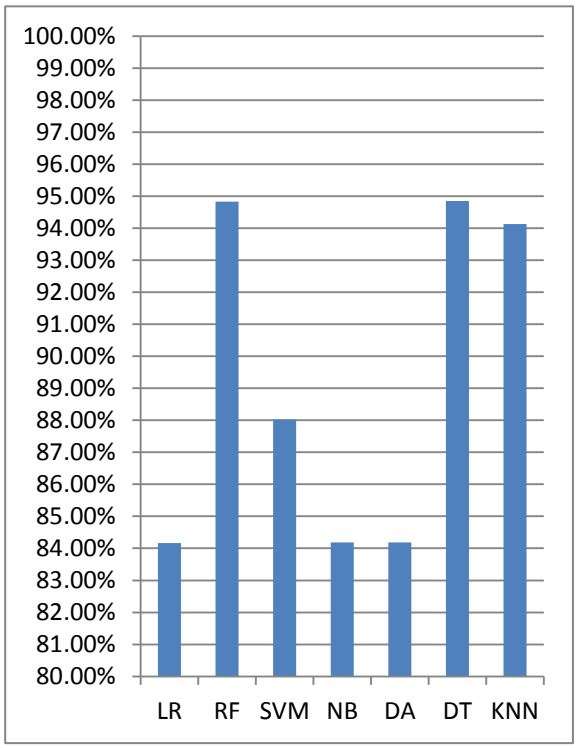


Figure 6: Accuracy after feature selection

5.4. F1-score: conveys the balance between recall and precision.it is computed as:-
2*((precision*recall)/(precision+recall))

We can obtain the F1- score from sklearn, which takes as inputs the actual labels and the predicted labels.

Table8: f1-score before feature selection

Algorithm	F1-score
Logistic Regression	90.40%
Random Forest	95.30%
Support Vector Machine	90.09%
Naïve Bayes	90.45%
Discriminant Analysis	90.45%
Decision Tree	94.36%
K-Nearest Neighbor	94.68%

Table9: f1-score after feature selection

Algorithm	F1-score
Logistic Regression	90.43%
Random Forest	96.74%
Support Vector Machine	92.97%
Naïve Bayes	90.44%
Discriminant Analysis	90.44%
Decision Tree	96.75%
K-Nearest Neighbor	96.32%

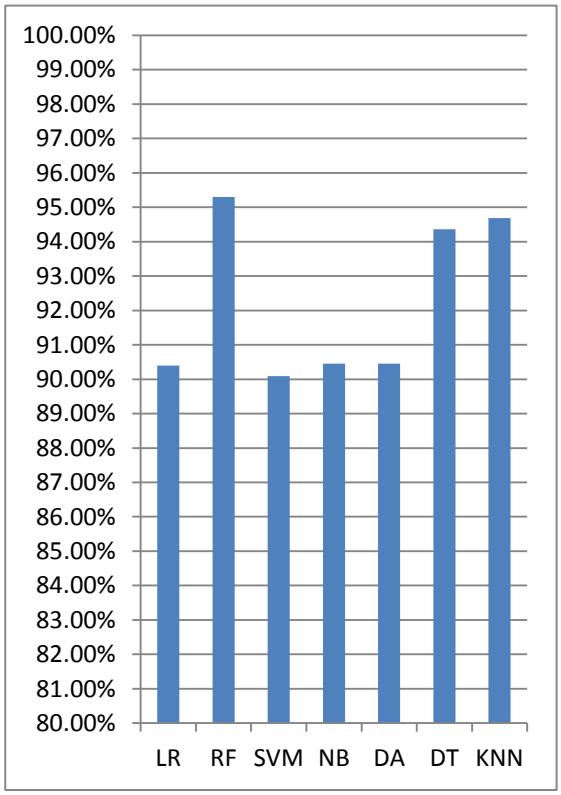


Figure 7: f1-score before feature selection

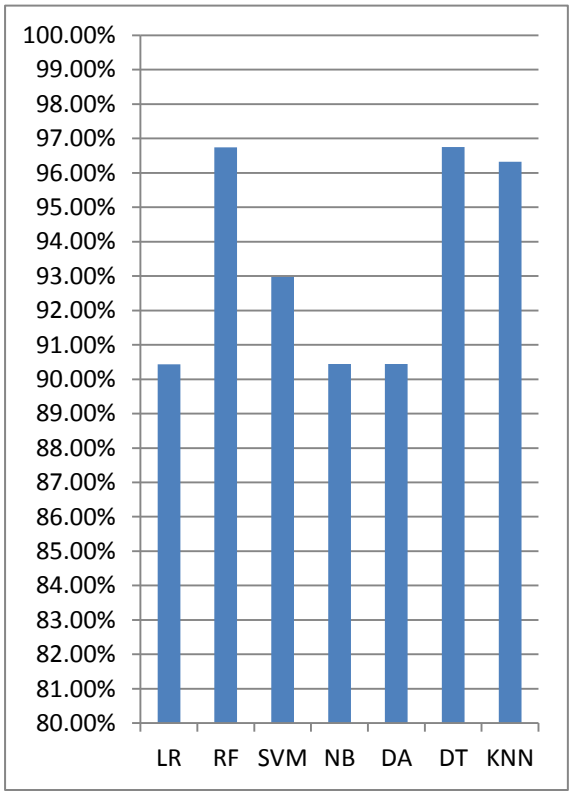


Figure 8: f1-score after feature selection

We can summarize results of (Precision – Recall –Accuracy – F1-score) before and after feature selection in the following two tables:-

Table10: results before feature selection

Algorithm	Precision	Recall	F1-score	Accuracy
LR	86.49%	94.69	90.40%	84.08%
RF	95.12%	95.49	95.30%	92.54%
SVM	86.05%	94.52	90.09%	83.52%
NB	86.72%	94.51	90.45%	84.19%
DA	88.89%	92.06	90.45%	84.59%
DT	94.99%	93.74	94.36%	91.12%
KNN	94.02%	95.34	94.68%	91.51%

Table11: results after feature selection

Algorithm	Precision	Recall	F1-score	Accuracy
LR	86.74%	94,45%	90.43%	84,16%
RF	96,68%	96,80%	96.74%	94,83%
SVM	86.87%	99,99%	92.97%	88,02%
NB	86,76%	94,45%	90.44%	84,18%
DA	86,76%	94,45%	90.44%	84,18%
DT	96,70%	96,81%	96.75%	94,85%
KNN	95,60%	97,05%	96.32%	94,13%

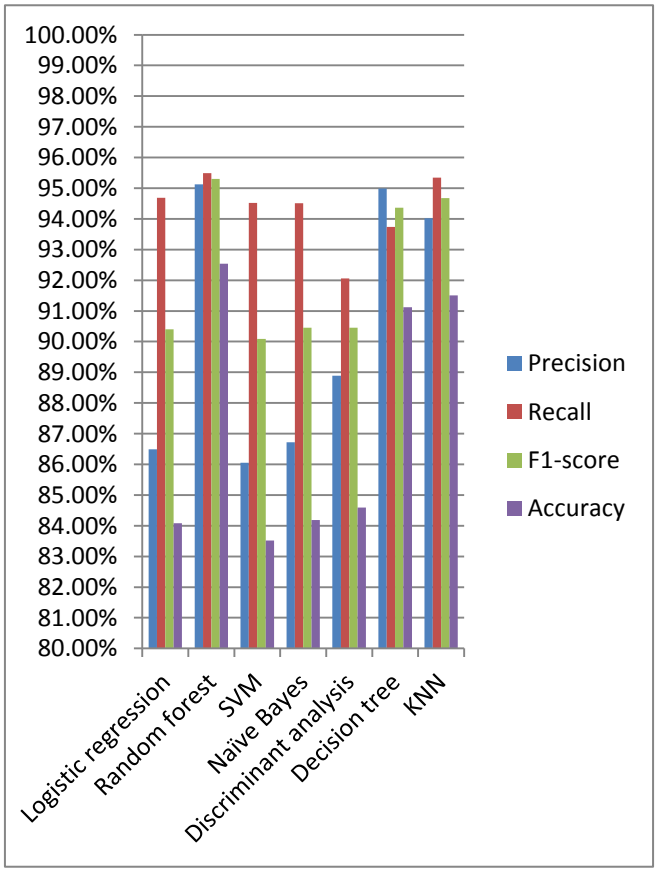


Figure 9: results before feature selection

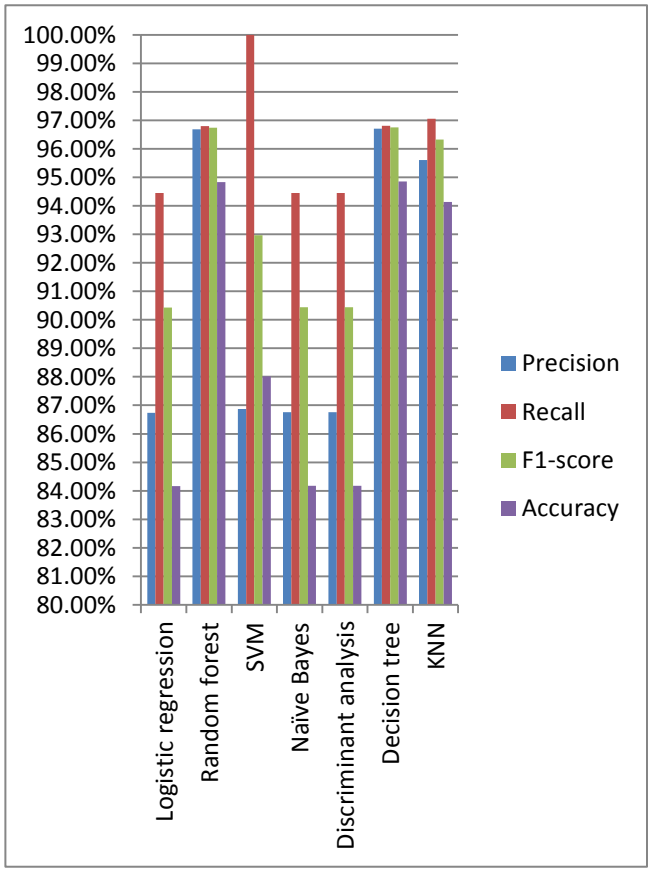


Figure 10: results after feature selection

6. Conclusion

The work presented in this paper is a step towards an efficient prediction approach for default customers of personal loans using machine learning.
Through our research we have tried to get the best approach to predict Credit Defaults. A description of the ABE personal loans data has been given earlier as a basis for evaluation. We found that we can use the most important features (activity – income – loan) that can lead to high accuracy for prediction of default customers out of ten features. We have observed that Decision tree has the best performance than any other machine learning models for our dataset with significant prediction accuracy of almost 94.85% although, Random forest and KNN also performed well but with lower accuracy than the Decision tree. However, some significantly popular Machine learning models could not provide satisfactory outcomes such as SVM, discriminant analysis, logistic regression and naïve bayes.

References

Abdou, H. A., & Pointon, J. (2009). Credit scoring and decision making in Egyptian public sector banks. International Journal of Managerial Finance, 5(4), 391–406.

Bank, J. H., & Davies, G. (2002). History of Money: From Ancient Times to the Present Day. University of Wales,pp.48-53.

Clarke, B., Fokoue, E., & Zhang, H. H. (2009). Principles and theory for data mining and machine learning. Springer Science & Business Media.

Collin, P. H. (2005). Dictionary of banking and finance,pp.85-261.

Das, K., & Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. International Journal of Innovative Research in Computer and Communication Engineering, 5(2), 1301-1309.

Dey, A. (2016). Machine learning algorithms: a review. International Journal of Computer Science and Information Technologies, 7(3), 1174-1179.

Graeber, d. (2011). Debt: The First Five Thousand Years. New york: Melvillehouse Brooklyn, pp.52-87, 211-237.

Hasan, K. (2016). Development of a credit scoring model for retail loan granting financial institutions from frontier markets. International Journal of Business and Economics Research, 5(5), 135-142.

Idiab, A. I. M., Haron, M. S., & Ahmad, S. (2011). Commercial banks & historical development. Journal of Applied Sciences Research, 7(7), 1024-1029.

Imtiaz, S., & J., A. (2017). A Better Comparison Summary of Credit Scoring Classification. International Journal of Advanced Computer Science and Applications, 8(7).

Ince, H., & Aktan, B. (2009). A comparison of data mining techniques for credit scoring in banking: A managerial perspective. Journal of Business Economics and Management, 10(3), 233–240.

Jafar Hamid, A., & Ahmed, T. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. Machine Learning and Applications: An International Journal, 3(1), 1–9.

Keramati, A., & Yousefi, N. (2011, January). A proposed classification of data mining techniques in credit scoring. In International Conference on Industrial Engineering and Operations Management (pp. 416-424).

Kumar, R., & Verma, R. (2012). Classification algorithms for data mining: A survey. International Journal of Innovations in Engineering and Technology (IJIET), 1(2), 7-14.

Siami, M., & Hajimohammadi, Z. (2013). Credit scoring in banks and financial institutions via data mining techniques: A literature review. Journal of AI and Data Mining, 1(2), 119-129.

Sousa, M. de M., & Figueiredo, R. S. (2014). credit analysis using data mining: application in the case of a credit union. Journal of Information Systems and Technology Management, 11(2), 379–396.

Srivastava, S., & Garg, A. (2013). Data mining for credit card risk analysis: a Review. International Journal of Computer Science, 3(2), 193-200.

Sudhakar, M., & Reddy, D. C. K. (2014). credit evaluation model of loan proposals for banks using data mining. International Journal of Latest Research in Science and Technology, 126-131.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International journal of forecasting, 16(2), 149-172.

Xiao, W., Zhao, Q., & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. Journal of Systems Science and Systems Engineering, 15(4), 419–435.

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480.

جامعة القاهرة
كلية الدراسات العليا للبحوث الإحصائية

المؤتمر السنوى الرابع والخمسون للإحصاء
وعلوم الحاسب وبحوث العمليات

علوم الحاسب والمعلومات

9-11 ديسمبر 2019