Cairo University

The 56th Annual International

Conference of Data science

Information Systems & Technology

4 – 6, Dec, 2023

Index

Information Systems & Technology

A survey on diseases prediction using machine learning techniques

Mohammed Badawy [1]*, Nagy Ramadan [1], and Hesham Ahmed Hefny [2]
[1] Department of Information Systems &Technology, and [2] Department of Computer Sciences
Faculty of Graduate Studies for Statistical Research
Cairo University, Giza, Egypt
mbadawy@pg.cu.edu.eg, nagyrd@cu.edu.eg, hehefny@cu.edu.eg

***Abstract*— The ability to forecast healthcare outcomes has played a crucial role in saving lives in recent decades. In the field of healthcare, there is a rapid improvement of intelligent systems that analyze complex data relationships and convert them into useful information for use in the prediction process. Hence, artificial intelligence is rapidly transforming the healthcare sector. The role of machine learning systems is crucial in diagnosing and predicting diseases. These systems utilize clinical data and images to provide extensive clinical support by emulating human perception. It can even identify diseases that are challenging for human intelligence to detect. The accuracy of disease prediction can have a substantial impact on patient outcomes. Accurate and timely predictions can potentially save lives, while inaccurate predictions can pose a risk to patients' lives. Hence, it is imperative to make precise and reliable predictions and estimations of diseases. Therefore, it is crucial to have dependable and effective techniques for healthcare predictive analysis. Hence, the objective of this study is to provide an extensive analysis of the current machine learning and deep learning methods employed in healthcare prediction and to identify the inherent challenges in using these methods in the healthcare sector.**

Keywords*— *Healthcare, Healthcare Prediction, Machine Learning, Artificial Intelligence

I. INTRODUCTION

Every day, life changes, but the health of each age either gets better or worse. Never know what will happen in life. Sometimes come across a lot of people come across fatal health problems because diseases are found too late. It is estimated that more than 50 million adults around the world would have chronic liver disease. The sickness can be stopped, though, if it is found early. Based on machine learning, diseases can be predicted so that common illnesses can be found earlier. Health isn't a top priority right now, which has caused a lot of problems. A lot of people can't afford to see a doctor, and others are very busy and don't have time, but avoiding symptoms that keep coming back for a long time can have serious health effects [1].

Illness affects people all over the world, so doctors and specialists are doing everything they can to reduce deaths caused by diseases. Because there is a lot more healthcare data coming from multiple sources that don't necessarily function together, predictive analytics models have become increasingly significant in the medical industry recently. The task of managing, archiving, and examining the vast quantities of historical and continuous data generated by healthcare services has become a novel challenge when compared to conventional database storage [2].

It is now common practice in healthcare systems to digitize all medical information and look at clinical data. This is because computers have made these tasks easier. In 2003, the Institute of Medicine, which is part of the National Academies of Sciences, Engineering, and Medicine, picked the term "electronic health records" to describe the records that kept improving healthcare for both patients and doctors. People who work in healthcare say that electronic health records (EHR) are "computerized medical records for patients that include all information about a person's past, present, or future that occur in an electronic system used to capture, store, retrieve, and link data primarily to offer healthcare and health-related services" [3].

Every day, healthcare services generate a vast volume of data, making it progressively challenging to analyse and manage using traditional methods. By applying machine learning techniques, this data will be effectively analyzed to produce practical and actionable insights. Furthermore, genetics, medical data, social media data, environmental data, and various other data sources can be employed to complement healthcare data. Figure 1 visually illustrates the data sources. Machine learning can be applied to four important areas in healthcare: prognosis, diagnosis, therapy, and clinical workflow, as described in section [4].

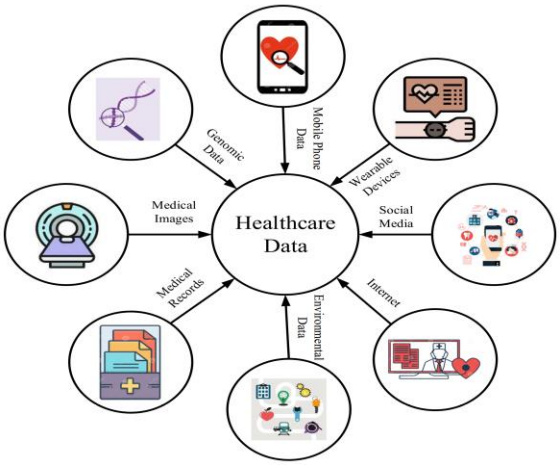

Fig. 1. Illustration of heterogeneous sources contributing to healthcare data [4].

The increased interest in predictive analytics techniques for healthcare is seen in the sustained investment in the development of innovative technologies utilizing machine learning techniques. Aims to improve individual health by accurately predicting future events. Previously known as clinical predictive models, these models aided in identifying individuals who had a higher likelihood of having an illness. These prediction algorithms are employed to determine clinical treatment choices and provide guidance to patients based on specific patient attributes [5].

The term "medical care" refers specifically to the management and coordination of curative treatments within the broader field of healthcare. White first developed the concept of medical care ecology in 1961. White suggested a framework for identifying patterns of health related to symptoms reported by specific demographics, as well as individuals' decisions regarding medical care. Within this framework, it can compute the percentage of the population that utilized medical services within a specified period. The hypothesis of the "ecology of medical care" has gained significant acceptance in academic circles in recent decades [6].

Because of the time constraints faced by medical workers, diseases move quickly, and patients' conditions change all the time, diagnosis is a very difficult process. But an exact way of diagnosing is needed to make sure that treatment goes quickly, and patients are safe [7]. Predictive analytics of diseases is a key industrial requirement. It can have a substantial impact on the accuracy of disease prediction, which can save patients' lives if the prognosis is accurate and timely but can potentially threaten patients' lives if the prediction is erroneous. As a result, diseases must be predicted and estimated exactly. As a result, dependable and effective healthcare predictive analysis methodologies are necessary.

The objective of this study is to provide a comprehensive review of common machine learning techniques applied in healthcare prediction, while also highlighting the inherent challenges involved in implementing these approaches in the healthcare domain.

The remainder of the paper is structured as follows: The "Background" section provides a theoretical foundation for machine learning techniques. The section " survey methodology " provides an overview of the literature on machine learning approaches used in healthcare prediction. The section "Results and Discussion" discusses the findings of prior studies on healthcare prediction. The section "Challenges" discusses the current issues relating to the survey's topic. Finally, Section "Conclusion" brings the paper to a close.

II. BACKGROUND

Predictive analytics methods are becoming more popular as a way to improve healthcare. This is shown by the large amount of research and development going into making cutting-edge tools based on machine learning for predicting individual health outcomes. Doctors could better find and treat patients who were more likely to get a dangerous illness with the help of clinical predictive models. These prediction algorithms are used to help patients and guide professional practice. They do this by looking at a number of factors that are unique to each patient.

Artificial intelligence (AI) is a system's ability to understand data. It uses computers and other tools to help people make better decisions, solve problems, and come up with new technologies [8]. Figure 2 depicts machine learning and deep learning as subsets of AI.

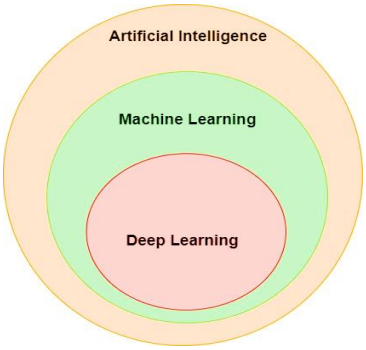

Fig. 2. AI, Machine Learning, and Deep Learning.

1. MACHINE LEARNING

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on creating predictive algorithms. These techniques are designed to enable machines to independently examine data and acquire knowledge [9]. Machine learning employs algorithms, methodologies, and procedures to identify fundamental relationships within data and generate descriptive and predictive tools that analyse those relationships. Machine learning is commonly linked with the fields of data mining, pattern recognition, and deep learning. While the boundaries between these domains are not well-defined and frequently intersect, it is widely acknowledged that deep learning is a relatively recent topic of machine learning. It employs sophisticated computational methods and vast quantities of data to establish intricate correlations among datasets. Figure 3 illustrates the categorization of ML algorithms into three distinct types: supervised learning, unsupervised learning, and reinforcement learning [10].

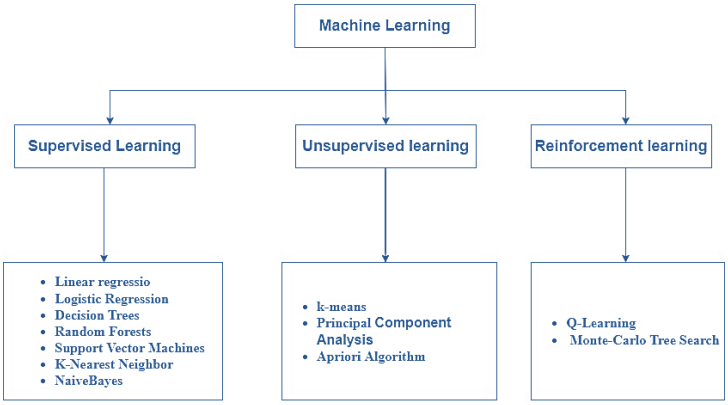

Fig. 3. Different types of Machine Learning Algorithms.

1.1. Supervised Learning

Based on a set of training cases that match the inputs and outputs, supervised learning is a machine learning model for looking into how inputs and outputs are related in a system [11]. A collection that has been labelled is used to train the model. It's like how a student learns basic math from a teacher. For this type of learning to happen, the data must be labelled, and the right answers must be forecast based on the algorithm's output [12]. Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbour, Naive Bayes, and Linear Regression are some of the most common supervised learning-based methods.

A. Linear Regression

Is a popular way to use statistics to make predictions. The independent input variable makes a good guess about the dependent output variable. In the equation, X and Y are linked. assuming that the inputs are real, number-based, and continuous.

$$Y = mX + c. \qquad (1)$$

The slope is shown by m, and the intersection is shown by c. Based on equation 1, we can figure out that the independent parameters (X) and the dependent parameters (Y) are linked [13].

A positive aspect of linear regression is that it is simple to learn, and regularization makes it easy to get rid of overfitting. An issue with linear regression is that it is hard to use with interactions that are not linear. It is not suggested for most practical uses, though, because it makes real-world problems much easier [14]. Some of the programmes that can be used to apply Linear Regression are Python, R, MATLAB, and Excel.

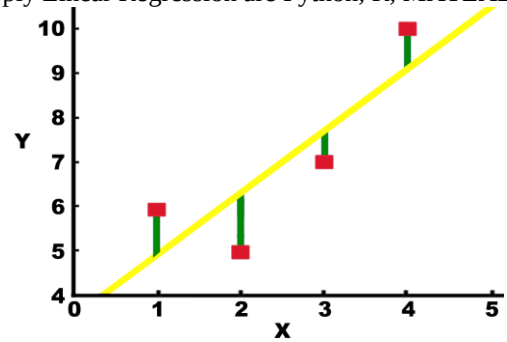

Fig. 4. Linear Regression model.

Figure 4 displays observations highlighted in red, together with random deviations indicated in green, which deviate from the main relationship depicted in yellow between the independent variable (x) and the dependent variable (y) [15].

B. Logistic Regression

The logistic model, or logistic regression, studies the association between several independent variables and a categorical dependent variable and determines event probability by fitting the data to a logistic curve [16]. Discrete mean values must be binary—true or false, 0 or 1, yes or no, superscript or subscript. Logistic regression predicts categorical variables and solves classification difficulties. Logistic regression can be done in R, Python, Java, and MATLAB [17]. Logistic regression illustrates the linear relationship between dependent and independent variables well. It's easy to grasp. However, it solely predicts numerical output, ignores non-linear data, and is susceptible to outliers [18].

C. Decision Tree

The supervised learning technique utilized for classification is the Decision Tree (DT). It mixes attribute values based on their ascending or descending order [19]. As a tree-based method, DT defines each path beginning at the root with a data-separating sequence until a Boolean conclusion is reached at the leaf node [20- 21]. DT is a node-and-link-based hierarchical representation of knowledge relationships. Nodes reflect purposes when relations are used to classify [22 - 23]. Figure 5 depicts an example of DT.

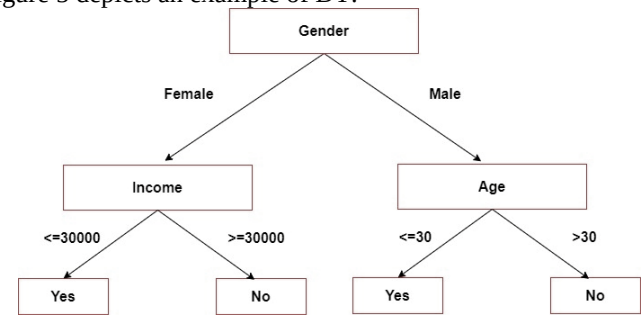

Fig. 5. Example of a DT.

DTs have several disadvantages, including greater complexity with increasing nomenclature, minor changes that may result in a different architecture, and longer processing time to train data [18]. Python (Scikit-Learn), KNIME, Orange, and R Studio are the implementation tools used in DT [17].

D. Random Foresta

Random Forest (RF) is a commonly used technique that consistently yields accurate results. It can be employed for categorization and prediction. The model generates a collection of decision trees then combines them [24].

The accuracy of the findings in the RF classifier increases proportionally with the number of trees in the forest. The RF has created a set of DTs known as the forest and merged them to enhance the accuracy of forecast outcomes. In the context of RF, each DT is constructed using only a subset of the provided dataset and trained using approximations. The RF algorithm combines many choice Trees DTs in order to achieve the most optimal choice [17].

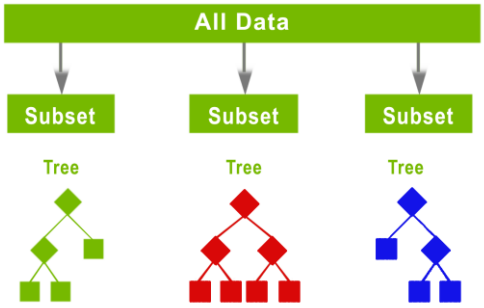

Fig. 6. Random Forest architecture.

As depicted in Figure 6, the RF algorithm takes a selection of features from the data and generates a specified number of random trees from each subset [15]. The random forest algorithm will aggregate the findings from all decision trees and present them in the final output.

E. Support Vector Machine

The Support Vector Machine (SVM) is a supervised machine learning approach used for classification problems and regression models. The SVM is a mathematical model that provides answers for both linear and nonlinear problems. As shown in Figure 7. The concept of margin calculation is the basis of its foundation. The dataset is partitioned into multiple groups to establish relationships between them [17].

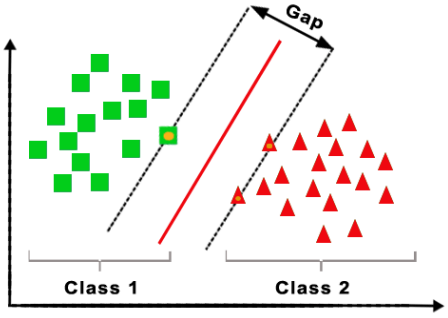

Fig. 7. Support Vector Machine.

SVM is a statistical learning method that aims to find decision boundaries in a usable N-dimensional space to accurately classify data points [26]. In SVM, the decision border between two classes is determined by the value of each data point. This boundary is defined by the support vector points that are located along the boundary between the classes [27].

F - Nearest Neighbor

K-nearest neighbor (KNN) is a "instance-based learning" or non-generalized learning algorithm [27]. It is also referred to as a "lazy learning" algorithm. KNN helps people figure out how to classify things. As shown in Figure

8, KNN figures out the distance between the nearest training data class labels and a new test data point in order to guess the target label of the new test data. After using the K value to find the number of closest data points, it ends with the name of the new test data class. The KNN usually sets the value of K to (1): k = n^(1/2) to find the number of nearest-distance training data points. The size of the dataset is given by n [18].

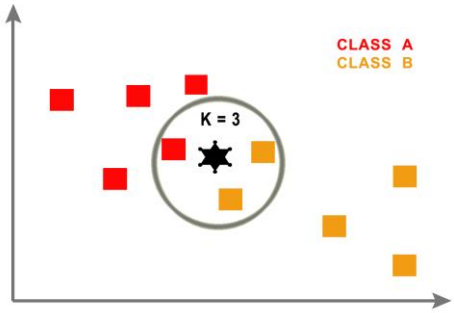

Fig. 8. K-Nearest Neighbor.

G. Naïve Bayes

Naive Bayes (NB) is a probabilistic model that relies on Bayes' theorem. It is straightforward to implement and does not require complex recursive parameter estimates, making it well-suited for large data sets [28]. The class membership degree is determined by NB based on a specified class identification [29]. By doing a single scan of the data, the categorization process becomes simplified [30]. Essentially, the NB classifier operates under the assumption that the presence of a specific feature in a class is independent of the presence of any other characteristic. Its primary focus is on the text categorization sector [31].

In summary, Table 1 presents the advantages and drawbacks of each model that were previously mentioned.

Table 1. Comparison of Various Supervised Learning Machine Learning Techniques.

Method	Advantages	Disadvantages
Linear Regression	▪ Easy to understand for beginners. ▪ Training linear regression models is fast, even on large datasets. ▪ Linear regression models can forecast, classify, and predict. ▪ Robust to noise: Linear regression models can handle noisy data well.	▪ Linearity: Linear regression models require linearity between independent and dependent variables. This can limit non-linear relationships. ▪ Sensitive to outliers: Linear regression models can be inaccurate due to outliers. ▪ Sometimes wrong: Linear regression models aren't always the best. For tasks where the independent and dependent variables are not linear, they may be less effective than other models.
Logistic Regression	▪ Excellent performance with small datasets. ▪ Its output is interpretable as probability.	▪ Compliant data assumptions are required. ▪ It only offers linear solutions.
Decision trees	▪ Can manage categorical characteristics. ▪ There are a few parameters to tune.	▪ The interpretability of the ensemble is questionable
Random Forest	▪ Can achieve great accuracy even when dealing with noisy or imbalanced data. ▪ Interpretability: Random Forest models are easy to understand. ▪ Random forest scales to large datasets.	▪ Computational complexity: Random Forest training is computationally expensive, especially for large datasets. ▪ Sensitivity to hyperparameters: Random Forest performance can be sensitive to hyperparameters.
Support Vector Machine	▪ High-dimensional space for input. ▪ Few irrelevant features.	▪ A huge training set is required. ▪ Data collection is time-consuming.
K-Nearest Neighbors	▪ Simple algorithm	▪ The user must specify the number of neighbors. ▪ A high level of relative computational complexity
Naïve Bayes	▪ Simple and straightforward method. ▪ It combines effectiveness and reasonable precision.	▪ Used primarily when the size of the training set is smaller. ▪ It assumes the conditional independence of linguistic features.

1.2 Unsupervised learning

Unsupervised learning is different from supervised learning because there are no experts and no right answers [31]. It is based on the idea that a machine can figure out complicated patterns and processes without any help from a person. This method works especially well when experts don't know what to look for in the data and the data itself

doesn't have the goals written on it.

Unsupervised learning is commonly employed in multimedia content processing, as it is sometimes necessary to cluster and divide data without the presence of class labels [32]. Three widely used unsupervised learning methods include k-means, Principal Component Analysis (PCA), and the Apriori Algorithm.

A. k-means

The k-means algorithm is one of the most famous unsupervised learning methods for solving the well-known clustering problem. It is also the most common way to divide data into groups. The process sorts a certain set of data into a certain number of pre-selected (assuming k-sets) groups [33].

K means is better than hierarchical grouping in a number of ways, including being faster on computers when there are a lot of factors. When k is small, it makes groups that are closer together than hierarchical ones. Another benefit is that the results are easy to understand and put into action. But K-Means has some problems, like the fact that it's hard to guess what K will be. The performance is also changed because different beginning parts lead to different ending combinations. It works well for raw points and local optimisation, but there isn't just one answer for each K value. To find the best result, the average of the K value has to be run 20–100 times, and then the result with the lowest J is chosen [14].

B. Principal Component Analysis

Principal component analysis (PCA) is an important tool for modern data analysis because it helps you find the most important data in a dataset, reduce the size of the dataset by keeping only the most important features without losing much information, and make the description of the dataset easier to understand [34, 35].

PCA is often used to reduce the number of variables in data before classification models are applied. Also, unsupervised methods like dimensionality reduction or clustering algorithms are often used to reduce the amount of data, find similar patterns or behaviors, and make the data easier to understand [36].

C. Apriori

The Apriori algorithm operates on the principle of representing the strategy for filter generation. A filter element set (k + 1) is generated by utilizing the repeated k element groups. Apriori employs an iterative methodology known as planar search, which involves the exploration of (k+1) item sets using k item sets. After generating the set of repeating 1-items through a process of item collection that satisfies the minimum support requirement, the set of repeating 1-items is generated. The group produced is referred to as L1. Subsequently, L1 is employed to locate L2, L3 is located using a recursive set of two elements, and so forth, until no iteration of k elements is successful. To locate every Lk, a complete dataset scan is required. In order to enhance production efficiency with respect to repeated element groups, a critical property known as the Apriori property is implemented to decrease the size of the search space. All non-empty subsets of a recursive element group must be iterative, according to the Apriori property. To identify clusters of common elements, a two-step process is employed: join and prune activities [37]. In summary, Table 2 presents the advantages and drawbacks of each model that were previously mentioned.

TABLE 2. Comparison of Different Unsupervised Learning Machine Learning techniques.

Method	Advantages	Disadvantages
K-means	▪ Easy-to-understand algorithm with linear relative computational complexity	▪ The worker must specify the number of clusters or classes. ▪ Poor performance with irregular shape clusters
Principal Component Analysis	▪ The use of PCA reduced the complexity of image grouping. ▪ Noise reduction because the maximum variation basis is used, so small variations in the background are ignored automatically.	▪ It is difficult to evaluate the covariance matrix accurately. ▪ The PCA cannot capture even the most basic invariance unless the training data explicitly provides this information.
Apriori	▪ Makes use of the large itemset property. ▪ Simple to implement. ▪ It is simple to parallelize.	▪ There are a large number of candidates sets generated. ▪ A number of database scans are required. ▪ 3) The system I/O cost rises as a result of multiple scans of the transactional database.

1.3. Reinforcement learning

Reinforcement learning (RL) distinguishes itself from supervised learning and unsupervised learning. It is a learning technique that is focused on achieving certain goals. Reinforcement learning (RL) is intricately connected to an agent, also known as a controller, which assumes the responsibility of the learning process in order to

accomplish a specific objective. The agent selects actions, which subsequently cause the environment to alter its state and provide rewards. Rewards are given in the form of positive or negative numerical values. The primary objective of an agent is to optimize the total rewards obtained over its duration of operation. A job refers to a comprehensive specification of the environment that outlines the process of generating rewards [38]. The Q-Learning algorithm and the Monte-Carlo Tree Search (MCTS) are among the most widely used reinforcement learning-based algorithms.

A. Q-Learning

Q-Learning is a kind of RL that doesn't use models. It could be thought of as an asynchronous method to dynamic programming. It lets agents figure out how to work best in Markovian domains by looking at what happens when they do things, without having to make domain maps [39]. It was an incremental approach to dynamic programming that didn't require a lot of computer power. It works by making the evaluation of the quality of each person's action better over time in certain states [40].

The advantage about developmental Q-learning is that it can successfully find the reward value in a given multi-agent environment method, since agents in ant Q-learning talk to each other. It's possible for Q-learning to get stuck in the local minimum because its agents only look for the quickest path [41].

B. Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS) is a good way to solve problems with sequential selection. Its plan is based on a smart tree search that equally explores and exploits. MCTS gives out random samples in the form of simulations and keeps track of activity data so that each time around, people can make better decisions. MCTS is a decision-making method used to search through very large, complicated areas that look like trees. Each node in these kinds of trees represents a state, which is also called a problem setup. Edges show how one state changes into another [42].

The MCTS is directly linked to situations that can be shown by a Markov decision process (MDP), which is a kind of discrete-time random control process. The MCTS can be used with Partially Observable Markov Decision Processes (POMDP) [43] after some changes are made to it. Recently, MCTS and deep RL were used to build AlphaGo by Google DeepMind, which was described in [44]. In summary, Table 3 presents the advantages and drawbacks of each model that were previously mentioned.

TABLE 3. Comparison of Different Reinforcement Machine Learning Methods.

Method	Advantages	Disadvantages
Q-Learning	▪ It can be applied to a wide range of problems, including those that are difficult to solve using other methods. ▪ It can learn from experience, which means it can improve its performance over time.	▪ It can be difficult to learn, particularly for problems with large state spaces. ▪ It can be sensitive to initial conditions, which means that if it is not properly initialized, it may learn a suboptimal policy.
Monte Carlo Tree Search	▪ It is useful for solving problems with a large search space. ▪ It can learn from experience, which means it can improve its performance over time. ▪ It is relatively simple to implement and can be applied to a wide range of problems.	▪ It can be difficult to learn, particularly for problems with a large search space. ▪ It can be sensitive to initial conditions, which means that if it is not properly initialized, it may learn a suboptimal policy. ▪ Due to the requirement to store a tree of possible states, it can be difficult to scale to large problems.

III. SURVEY METHODOLOGY

The studies addressed in this paper have been presented and published in reputable journals and worldwide conferences published by prominent scientific publishers like IEEE, Springer, Elsevier, Hindawi, Frontiers, Taylor, and MDPI. The search engines employed include Google Scholar, Scopus, and ScienceDirect. The selected papers encompassed the time frame spanning from 2019 to 2023. Terms such as machine learning, healthcare, surgery, cardiology, radiology, hepatology, and nephrology are commonly used to search for papers related to these topics. The selected studies in this survey focus on the utilization of ML algorithms for healthcare prediction. This survey included empirical and review publications pertaining to the issues. This section examines current research endeavours in healthcare prediction utilizing several ML approaches. This survey provides an in-depth analysis of the methodologies and algorithms employed for making predictions, as well as the performance measures and tools utilized in their model.

The authors of [45], presented a system utilizing a machine learning-based technique to forecast Parkinson's Disease. The system employs a range of machine learning models, such as Gradient Boosted Tree, random forest, and logistic regression, to detect significant indicators and patterns linked to the disease. Logistic regression

achieved a higher accuracy rate of 85%. Although this research represents a notable advancement, there is still potential for further enhancement. Additional research and development efforts have the potential to enhance the accuracy of the prediction system by integrating supplementary data sources and optimizing the machine learning algorithms employed.

The authors of [46] attemped to employ machine learning algorithms for the prediction of probable heart problems in people. This research involves evaluating the performance of different classifiers, such as Decision Tree, Naive Bayes, Logistic Regression, SVM, and Random Forest. Applied to the dataset were Decision Tree, Random Forest, Naive Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting. The dataset has 76 variables that encompass the anticipated attributes associated with heart disease in individuals. From these, 14 relevant characteristics have been selected to aid in evaluating the system. Among the seven classifiers, the extreme gradient boosting classifier achieves the best accuracy of 81%.

The authors of [47] utilised a framework to create and test ML classification models such as Logistic Regression, KNN, SVM, and RF to predict diabetes patients. The ML approach was applied to the 768-row, 9-column Pima Indian Diabetes Database (PIDD). The forecasting accuracy is 83%. The implementation approach's results show that Logistic Regression outperformed other ML algorithms; additionally, only a structured dataset was chosen; unstructured data were not considered; and finally, other factors such as diabetes family history, smoking habits, and physical inactivity should be considered for diabetes prediction.

In [48], the authors developed a diagnosis system that predicts diabetes using two separate datasets (Frankfurt Hospital in Germany and PIDD from the UCI ML repository) and four prediction models (RF, SVM, NB, and DT). The SVM algorithm had an accuracy of 83.1 percent. Some components of this study should be improved, such as employing a DL technique to predict diabetes; also, the model should be tested in other healthcare domains, such as heart disease and COVID-19 prediction datasets.

The authors of [49] suggested three ML approaches (Logistic Regression, DT, and Boosted RF) to evaluate COVID-19 utilising OpenData Resources from Mexico and Brazil. The suggested model incorporates only the COVID-19 patient's geographical, social, and economic conditions, as well as clinical risk factors, medical reports, and demographic data, to forecast rescue and death. The model for Mexico has a 93 percent accuracy on the dataset used, and an F1 score of 0.79. On the other hand, the Brazil model has a 69 percent accuracy and an F1 score of 0.75 on the used dataset. The three ML algorithms were tested, and the findings revealed that Logistic Regression is the optimum method of data processing. The authors should be worried about the use of authentication and data privacy management.

The authors of [50] introduced a new model for predicting type 2 diabetes utilizing a network approach and ML techniques. Logistic Regression, SVM, NB, KNN, Decision Tree, RF, XGBoost, and ANN are some of the techniques used. The healthcare data of 1,028 type 2 diabetes patients and 1,028 non-type 2 diabetes patients were collected from de-identified data to predict the risk of type 2 diabetes. The experimental results show that the models are effective, with an Area Under Curve (AUC) ranging from 0.79 to 0.91. The RF model outperformed the others in terms of accuracy. This study is only based on a dataset containing hospital admission and discharge summaries from a single insurance company. For consumers who have several insurance carriers, external hospital visits and information from other insurance companies are unavailable.

The authors of [51] proposed a healthcare management system that patients could use to arrange doctor appointments and verify prescriptions. It aids ML in detecting diseases and determining medications. ML models such as DT, RF, logistic regression, and NB classifiers are applied to diabetes, heart disease, chronic kidney disease, and liver datasets. In the heart dataset, logistic regression had the highest accuracy of 98.5 percent out of all the models tested. The DT classifier has the lowest accuracy, coming in at 92 percent. Among all others in the liver dataset, logistic regression has the highest accuracy of 75.17%. The logistic regression, RF, and Gaussian NB all performed well with an accuracy of 1. The accuracy of 100% should be verified by employing k-folded cross-validation to assess the models' reliability. Random forest with a maximum accuracy of 83.67 percent in the diabetes dataset. The writers should include a hospital directory so that different hospitals and clinics may be accessible via a single gateway. Image datasets could also be supplied to allow image processing of reports and the use of DL to detect diseases.

The authors of [52] created an ML model to predict the occurrence of Type 2 Diabetes in the future year (Y + 1) using factors from the current year (Y). The dataset was obtained as an electronic health record from a private medical institute between 2013 and 2018. The authors used logistic regression, RF, SVM, XGBoost, and ensemble ML algorithms to predict non-diabetic, prediabetic, and diabetic outcomes. The three classes were quickly chosen using feature selection. Among the features chosen were FPG, HbA1c, triglycerides, BMI, gamma-GTP, gender, age, uric acid, smoking, drinking, physical activity, and family history. According to the trial data, the RF model had the highest accuracy (73%), while the Logistic Regression model had the lowest (71%). The authors provided a

model that relied solely on a single dataset. As a result, further data sources should be used to validate the models developed in this work.

To increase the model's accuracy, the authors of [53] classified the diabetes dataset using SVM and NB algorithms with feature selection. For analysis, PIDD is downloaded from the UCI Repository. The authors used the K-fold cross-validation model for training and testing. The SVM classifier performed better than the NB method, offering around 91% correct predictions; however, the authors acknowledge that they need to extend to the most recent dataset, which will include additional attributes and rows.

K-means clustering is an unsupervised machine learning approach developed by the authors of [54] to detect heart disease in its early stages using the UCI heart disease dataset. PCA is used to reduce dimensionality. The method's results show that it can predict early heart disease with 94.06% accuracy. The authors should employ more than one algorithm and dataset to implement the proposed technique.

The authors of [55] used the logistic regression classification technique to create a predictive model for the classification of diabetes data. The dataset contains 459 patients for training and 128 cases for testing. The prediction accuracy with logistic regression was 92%. The authors' biggest weakness is that they did not compare the model to other diabetes prediction algorithms, therefore it cannot be confirmed.

The authors of [56] created a prediction model that uses ML algorithms (DT classifier, RF classifier, and NB classifier) to assess the user's symptoms and predict the condition. The goal of this research was to solve health-related problems by helping doctors to predict diseases at an early stage. The dataset consists of 4920 patient records with 41 diagnoses. As a dependent variable, 41 disorders were considered. The accuracy score for all algorithms was 95.12%. Overfitting was observed when all 132 symptoms from the original dataset were analyzed rather just 95 symptoms. In other words, the tree appears to remember the dataset provided and, as a result, fails to classify fresh data. As a result, just 95 symptoms were evaluated during the data-cleansing process, and the best ones were selected.

The authors of [57] created a decision-making system that helps practitioners anticipate cardiac problems in exact classification using a simplified manner and will offer automated predictions about the patient's heart status. Four algorithms (KNN, RF, DT, and NB) were implemented and employed in the Cleveland Heart Disease dataset. The accuracy of various classification algorithms varies. When they used the KNN method with the Correlation factor, they achieved about 94 percent accuracy. The authors should expand the presented technique to forecast other diseases using more than one dataset.

The Cleveland dataset, which had 303 examples and 76 attributes, was used by the authors of [58] to test three alternative classification algorithms in addition to KNN: NB, SVM, and DT. Only 14 of these 76 features will be subjected to testing. The authors used data preprocessing to reduce noise from the data. The KNN achieved the highest accuracy of 90.79 percent. To enhance the accuracy of early heart disease prediction, the authors must employ more advanced models.

authors in [58] suggested a model to predict heart disease using a cardiovascular dataset that was then classified using supervised machine learning algorithms (DT, NB, Logistic Regression, RF, SVM, and KNN). The results show that the DT classification model predicted cardiovascular illnesses with a higher accuracy of 73% than previous algorithms. The authors emphasized that using ensemble ML techniques with the CVD dataset can result in a better sickness prediction model.

The authors of [60] aimed to improve the accuracy of heart disease prediction by utilizing Logistic Regression on a healthcare dataset to assess whether individuals had heart disease or not. The findings came from ongoing cardiovascular research on residents of Framingham, Massachusetts. The model predicted accuracy of 87%. The authors admit that further data and the deployment of more ML models could improve the model.

Because breast cancer affects one out of every 28 women in India, the author of [61] demonstrated an accurate classification strategy for examining a breast cancer dataset of 569 rows and 32 columns. Using a heart disease dataset and a lung cancer dataset, this study provided a novel approach to function selection. This selection method is based on evolutionary algorithms combined with SVM classification. Lung cancer 81.8182, Diabetes 78.9272 are the classifier results. It should be noted that the size, kind, and source of data used are not specified.

The authors of [62] predicted the risk factors that cause heart disease using the K-means clustering algorithm and analyzed the data with a visualization tool using a Cleveland heart disease dataset with 76 features of 303 patients, 209 records with 8 attributes such as age, chest pain type, blood pressure, blood glucose level, ECG in rest, heart rate, and four types of chest pain. The scientists anticipate cardiac disorders only based on the fundamental characteristics of four forms of chest pain, and K-means clustering is a common unsupervised ML technique.

The purpose of the study [63] was to describe the benefits of employing a number of Data Mining (DM) methodologies and proven heart disease survival prediction models. Based on their findings, the authors indicated that Logistic Regression and NB produced the maximum accuracy on the Cleveland hospital dataset when run on a

high dimensional dataset, whereas DT and RF produce superior results on small dimensional datasets. Because the RF classifier is an optimized learning algorithm, it outperforms the DT classifier in terms of accuracy. According to the author, this work can be expanded to other ML techniques, and the model can be produced in a distributed environment such as Map-Reduce, Apache Mahout, HBase, and so on.

The authors of [64] suggested a single algorithm called hybridization to forecast cardiac disease, which incorporates previously utilized techniques into a single algorithm. The provided Method is divided into three stages. Preprocessing, classification, and diagnosis are all phases. They used the Cleveland database as well as the algorithms NB, SVM, KNN, NN, J4.8, RF, and GA. NB and SVM always outperform others, whilst others are dependent on the features specified. The accuracy of the results was 89.2 percent. The primary focus should be authors. Because the dataset was small, the system was unable to train sufficiently, resulting in poor method accuracy.

The authors of [65] investigated several data representations using six algorithms (Logistic Regression, KNN, DT, SVM, NB, and RF) to better understand how to use clinical data for predicting liver disease. The original information was collected in the northeast of Andhra Pradesh, India, and contains 583 liver patient data, with 75.64 percent being male and 24.36 percent being female. According to the analysis results, the Logistics Regression classifier provides the most enhanced order exactness of 75 percent dependent on the f1 measure to forecast the liver sickness, while NB provides the lowest accuracy of 53 percent. The authors just looked at a few popular supervised ML algorithms; other algorithms can be chosen to develop a more precise model of liver disease prediction, and performance can be progressively improved.

The authors of [66] used ML technology to predict coronary heart disease (CHD) based on past medical data. The purpose of this work is to detect correlations in CHD data using three supervised learning approaches: NB, SVM, and DT. The dataset includes a retrospective sample of males from KEEL, a high-risk heart disease region in South Africa's Western Cape. The model made use of NB, SVM, and DT. NB was the most accurate of the three models. SVM and DT J48 surpassed NB with an 82 percent specificity rate but a sensitivity rate of less than 50 percent.

The authors of [67] created a chronic disease risk prediction framework that was created and evaluated in the Australian healthcare system to predict type 2 diabetes risk using DM and network analysis approaches. Using a six-year Australian private healthcare funds dataset and three different predictive methods (regression, parameter optimization, and DT). The prediction's accuracy ranges from 82 to 87 percent. The dataset's source is the hospital admission and discharge summary. As a result, it does not provide information on routine doctor visits or prospective diagnoses.

IV. RESULTS AND DISCUSSION

The study categorized the examined papers according to the diseases they predicted. Consequently, the discussion focused on five diseases: Diabetes, COVID-19, Heart disease, Liver disease, and chronic kidney disease. Table 4 displays the quantity of evaluated articles for each disease, together with the utilized prediction methods for each disease. Table 4 presents a comparison of related studies organized by diseases and the model that achieved the highest level of accuracy.

Table 4. Related works by disease and most realistic model.

Disease	ML Techniques	Highest Accuracy
DIABETES	Logistic Regression, KNN, SVM, RF, NB, DT, SVM, ensemble machine learning.	The Logistic Regression model achieved 92% accuracy rate.
COVID-19	DT, RF, logistic regression, NB, K-means.	Logistic regression achieved the highest accuracy with 98.5 %.
HEART	Logistic Regression, KNN, RF, DT, NB, SVM, DT, K-means, PCA.	PCA achieved the highest accuracy with 94.06% .
LIVER	Logistics Regression, KNN, DT, SVM, NB, and RF.	Logistic Regression achieved the highest accuracy with 75%.
MULTIPLE DISEASE DETECTION	DT, RF, logistic regression, and NB.	Logistic regression achieved the highest accuracy with 98.5% in the heart dataset.

V. CHALLENGES

While machine learning has made tremendous progress in recent years, it still has a long way to go before it can be effectively employed to solve the core challenges affecting healthcare systems. Some of the difficulties in adopting ML techniques in healthcare prediction are covered here.

The fundamental difficulty that must be addressed is the Biomedical Data Stream. Significant amounts of fresh medical data are being generated at an alarming rate, and the healthcare industry as a whole is evolving at an alarming rate. Measurements of blood pressure, oxygen saturation, and glucose levels are examples of real-time biological signals. While several DL architectural variants have attempted to solve this issue, there are still numerous obstacles to overcome before efficient analyses of fast evolving, large amounts of streaming data can be undertaken. Memory consumption, feature selection, missing data, and computational complexity are examples of these issues. Another difficulty for ML and DL is dealing with the complexity of the healthcare domain.

Healthcare and biological research are more complex than other professions. We still don't know much about the origins, transmission, and treatments for many of these highly diverse diseases. It's difficult to collect adequate data because there aren't always enough patients. However, a solution to this problem may be found. Due to the small number of patients, extensive patient profiling, novel data processing, and the incorporation of additional data sets are required.

VI. CONCLUSION

Applying machine learning techniques to predict healthcare outcomes has the potential to transform traditional healthcare delivery. Healthcare data is regarded as the most important aspect in ML applications for medical-care systems. The goal of this paper is to give a thorough examination of the most important machine learning algorithms utilized in predictive analytics for healthcare. It also addressed the challenges and barriers connected with using Machine Learning techniques in the healthcare domain. Following the survey, a complete evaluation of 23 papers from 2019 to 2023 was carried out. Furthermore, a thorough examination of the technique used in each study was carried out. The papers reviewed showed that artificial intelligence approaches, notably machine learning, have a significant impact on accurately diagnosing diseases. Furthermore, these strategies enable the prediction and analysis of healthcare data by connecting various clinical records and recreating a patient's medical history based on this data.

REFERENCES

[1] Latha, M. H., Ramakrishna, A., Reddy, B. S. C., Venkateswarlu, C., & Saraswathi, S. Y. (2022). Disease Prediction by Stacking Algorithms Over Big Data from Healthcare Communities. Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2021, 265, 355.

[2] Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works?. Journal of the American Medical Informatics Association. 2019 Dec;26(12):1651-4.

[3] Murphy, G. F., Hanken, M. A., & Waters, K. A. (1999). Electronic health records: changing the vision.

[4] Rahmani, A. M., Yousefpoor, E., Yousefpoor, M. S., Mehmood, Z., Haider, A., Hosseinzadeh, M., & Ali Naqvi, R. (2021). Machine learning (ML) in medicine: Review, applications, and challenges. Mathematics, 9(22), 2970.

[5] El Seddawy, A. B., Moawad, R., & Hana, M. A. (2018). Applying Data Mining Techniques in CRM.

[6] Xiong, X., Cao, X., & Luo, L. (2021). The ecology of medical care in Shanghai. BMC Health Services Research, 21, 1-9.

[7] Mirbabaie, M., Stieglitz, S., & Frick, N. R. (2021). Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health and Technology, 11(4), 693-731.

[8] Tang, R., De Donato, L., Besinović, N., Flammini, F., Goverde, R. M., Lin, Z., ... & Wang, Z. (2022). A literature review of Artificial Intelligence applications in railway systems. Transportation Research Part C: Emerging Technologies, 140, 103679.

[9] Singh, G., Al'Aref, S. J., Van Assen, M., Kim, T. S., van Rosendael, A., Kolli, K. K., ... & Min, J. K. (2018). Machine learning in cardiac CT: basic concepts and contemporary data. Journal of Cardiovascular Computed Tomography, 12(3), 192-201.

[10] Kim, K. J., & Tagkopoulos, I. (2019). Application of machine learning in rheumatic disease research. The

Korean journal of internal medicine, 34(4), 708.
[11] Liu, B. (2011). Web data mining: exploring hyperlinks, contents, and usage data (Vol. 1). Berlin: springer.
[12] Haykin, S., & Lippmann, R. (1994). Neural networks, a comprehensive foundation. International journal of neural systems, 5(4), 363-364.
[13] Gupta, M., & Pandya, S. D. (2022). A Comparative Study on Supervised Machine Learning Algorithm. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 10(1), 1023-1028.
[14] Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.
[15] Srivastava, A., Saini, S., & Gupta, D. (2019, June). Comparison of various machine learning techniques and its uses in different fields. In 2019 3rd International conference on electronics, communication and aerospace technology (ICECA) (pp. 81-86). IEEE.
[16] Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing, 43(2), 154-164.
[17] Gupta, M., & Pandya, S. D. (2022). A Comparative Study on Supervised Machine Learning Algorithm. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 10(1), 1023-1028.
[18] Obulesu, O., Mahendra, M., & ThrilokReddy, M. (2018, July). Machine learning techniques and tools: A survey. In 2018 international conference on inventive research in computing applications (ICIRCA) (pp. 605-611). IEEE.
[19] Dhall, D., Kaur, R., & Juneja, M. (2020). Machine learning: a review of the algorithms and its applications. Proceedings of ICRIC 2019: Recent Innovations in Computing, 47-63.
[20] Yang, F. J. (2019, December). An extended idea about decision trees. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 349-354). IEEE.
[21] Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert systems with applications, 42(5), 2670-2679.
[22] Shamim, A., Hussain, H., & Shaikh, M. U. (2010, June). A framework for generation of rules from decision tree and decision table. In 2010 International Conference on Information and Emerging Technologies (pp. 1-6). IEEE.
[23] Eesa, A. S., Abdulazeez, A. M., & Orman, Z. (2017). A DIDS based on the combination of Cuttlefish Algorithm and decision tree. Science Journal of University of Zakho, 5(4), 313-318.
[24] Bakyarani, E. S., Srimathi, H., & Bagavandas, M. (2019). A survey of machine learning algorithms in health care. International Journal of Scientific & Technology Research, 8(11).
[25] Han J, Kamber M, Mining D. Data Mining Concepts and Techniques, Elevier, 2011.
[26] Aldahiri, A., Alrashed, B., & Hussain, W. (2021). Trends in using IoT with machine learning in health prediction system. Forecasting, 3(1), 181-206.
[27] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.
[28] Ting, K. M., & Zheng, Z. (1999). Improving the performance of boosting for naive Bayesian classification. In Methodologies for Knowledge Discovery and Data Mining: Third Pacific-Asia Conference, PAKDD-99 Beijing, China, April 26–28, 1999 Proceedings 3 (pp. 296-305). Springer Berlin Heidelberg.
[29] Oladipo ID, AbdulRaheem M, Awotunde JB, Bhoi AK, Adeniyi EA, Abiodun MK (2022) Machine learning and deep learning algorithms for smart cities: a start-of-the-art review. In: IoT and IoE driven smart cities, pp 143–162.
[30] Shailaja, K., Seetharamulu, B., & Jabbar, M. A. Machine learning in healthcare: A review. In2018 Second international conference on electronics, communication and aerospace technology (ICECA) 2018 Mar 29 (pp. 910–914).
[31] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.
[32] Greene, D., Cunningham, P., & Mayer, R. (2008). Unsupervised learning and clustering. Machine learning techniques for multimedia: Case studies on organization and retrieval, 51-90.
[33] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90-95.
[34] Smith, L. I. (2002). A tutorial on principal components analysis.
[35] Mishra, S. P., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., ... & Laishram, M. (2017).

Multivariate statistical data analysis-principal component analysis (PCA). International Journal of Livestock Research, 7(5), 60-78.
[36] Mahdi Kamani, Mohammad, Farzin Haddadpour, Rana Forsati, and Mehrdad Mahdavi. "Efficient Fair Principal Component Analysis." arXiv e-prints (2019): arXiv-1911.
[37] Singh, J., Ram, H., & Sodhi, D. J. (2013). Improving efficiency of apriori algorithm using transaction reduction. International Journal of Scientific and Research Publications, 3(1), 1-4.
[38] Coronato, A., Naeem, M., De Pietro, G., & Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. Artificial Intelligence in Medicine, 109, 101964.
[39] Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine learning, 8, 279-292.
[40] Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-learning algorithms: A comprehensive classification and applications. IEEE access, 7, 133653-133667.
[41] Juang, C. F., & Lu, C. M. (2009). Ant colony optimization incorporated with fuzzy Q-learning for reinforcement fuzzy control. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 39(3), 597-608.
[42] Świechowski, M., Godlewski, K., Sawicki, B., & Mańdziuk, J. (2022). Monte Carlo tree search: A review of recent modifications and applications. Artificial Intelligence Review, 1-66.
[43] Lizotte, D. J., & Laber, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. The journal of machine learning research, 17(1), 7378-7405.
[44] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.
[45] Patil, S., Jaybhaye, S., Bokariya, S., Jain, P., Phapale, S., & Hande, T. (2023). Parkinson's Disease Prediction System in Machine Learning. In ITM Web of Conferences (Vol. 56, p. 05002). EDP Sciences.
[46] Valle, H., Uppala, R., Vanumu, A., Sabbi, L., & Varalakshmi. (2023). Heart disease prediction using machine learning. Journal of Engineering Sciences, Vol 14 Issue 04.
[47] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, 2022.
[48] Edeh, M. O., Khalaf, O. I., Tavera, C. A., Tayeb, S., Ghouali, S., Abdulsahib, G. M., ... & Louni, A. (2022). A classification algorithm-based hybrid diabetes prediction model. Frontiers in Public Health, 10.
[49] Iwendi, C., Huescas, C. G. Y., Chakraborty, C., & Mohan, S. (2022). COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients. Journal of Experimental & Theoretical Artificial Intelligence, 1-21.
[50] Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Applied Intelligence, 52(3), 2411-2422.
[51] Chugh, M., Johari, R., & Goel, A. (2022). MATHS: Machine Learning Techniques in Healthcare System. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 3 (pp. 693-702). Springer Singapore.
[52] Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. International journal of environmental research and public health, 18(6), 3317.
[53] Gupta, S., Verma, H. K., & Bhardwaj, D. (2021). Classification of diabetes using Naive Bayes and support vector machine as a technique. In Operations Management and Systems Engineering: Select Proceedings of CPIE 2019 (pp. 365-376). Springer Singapore.
[54] Islam, M. T., Rafa, S. R., & Kibria, M. G. (2020, December). Early prediction of heart disease using PCA and hybrid genetic algorithm with k-means. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.
[55] Qawqzeh, Y. K., Bajahzar, A. S., Jemmali, M., Otoom, M. M., & Thaljaoui, A. (2020). Classification of diabetes using photoplethysmogram (PPG) waveform analysis: Logistic regression modeling. BioMed Research International, 2020.
[56] Grampurohit, S., & Sagarnal, C. (2020, June). Disease prediction using machine learning algorithms. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.
[57] Moturi, S., & Srikanth Vemuru, D. S. (2020). Classification model for prediction of heart disease using correlation coefficient technique. International Journal, 9(2).
[58] Barik, S., Mohanty, S., Rout, D., Mohanty, S., Patra, A. K., & Mishra, A. K. (2020). Heart disease prediction using machine learning techniques. In Advances in Electrical Control and Signal Systems: Select Proceedings of AECSS 2019 (pp. 879-888). Springer Singapore.
[59] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020, May).

Prediction of cardiac disease using supervised machine learning algorithms. In 2020 4th international conference on intelligent computing and control systems (ICICCS) (pp. 570-575). IEEE.

[60] Saw, M., Saxena, T., Kaithwas, S., Yadav, R., & Lal, N. (2020, January). Estimation of prediction for getting heart disease using logistic regression model of machine learning. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[61] Soni, V. D. (2020). Chronic disease detection model using machine learning techniques. International Journal of Scientific & Technology Research, 9(9), 262-266.

[62] Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart disease prediction using exploratory data analysis. Procedia Computer Science, 173, 130-139.

[63] Wu, C. S. M., Badshah, M., & Bhagwat, V. (2019, July). Heart disease prediction using data mining techniques. In Proceedings of the 2019 2nd international conference on data science and information technology (pp. 7-11).

[64] Tarawneh, M., & Embarak, O. (2019). Hybrid approach for heart disease prediction using data mining techniques. In Advances in Internet, Data and Web Technologies: The 7th International Conference on Emerging Internet, Data and Web Technologies (EIDWT-2019) (pp. 447-454). Springer International Publishing.

[65] Rahman, A. S., Shamrat, F. J. M., Tasnim, Z., Roy, J., & Hossain, S. A. (2019). A comparative study on liver disease prediction using supervised machine learning algorithms. International Journal of Scientific & Technology Research, 8(11), 419-422.

[66] Gonsalves, A. H., Thabtah, F., Mohammad, R. M. A., & Singh, G. (2019, July). Prediction of coronary heart disease using machine learning: an experimental analysis. In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies (pp. 51-56).

[67] Khan, A., Uddin, S., & Srinivasan, U. (2019). Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes. Expert Systems with Applications, 136, 230-241.

A Method for Traffic Flow Prediction in cloud computing

Sello Prince Sekwatlakwatla[1] and Vusumuzi Malele[2]

Department of Computer Science and Information Systems, North-West University, South Africa

Sek.prince@gmail.com[1] Vusi.Malele@nwu.ac.za[2]

***Abstract*—Cloud computing is an advanced technology that is easy to manage and run. This makes the business more flexible and reliable, increases performance and efficiency, and reduces IT costs. By eliminating storage and maintenance concerns, cloud computing has increased adoption and resulted in an increase in traffic in the cloud computing sector. As organisations cannot predict when and how cloud computing traffic will increase, they are unable to manage it effectively, creating a bottleneck in their system and driving users away. The paper proposes linear regression, time series decomposition, and autoregressive integrated moving average forecasting. Simulations indicate that ARIMA is more accurate at predicting cloud computing traffic than linear regression and time series decomposition. The ARIMA and linear regression results were very similar, at 96% and 95%, respectively. A time-series decomposition model yielded 69 percent accuracy. Future research should consider ensemble techniques combining multiple models to improve the results.**

Keywords—linear regression, time series decomposition, and autoregressive integrated moving average forecasting.

I. INTRODUCTION

The concept of cloud computing can simply be defined as the sharing of computing resources remotely on demand, so people can access data quickly and easily from anywhere in the world using an internet connection [1-2]. Organisations are responsible for keeping their data centers productive and efficient. The adoption of cloud computing as an innovative way of implementing information technology has grown significantly over the last decade [2]. Managed cloud services minimize the risk of data loss, making cloud computing extremely attractive [3]. As organisations cannot predict when and how cloud computing traffic will increase, they are unable to manage it effectively, creating a bottleneck in their system and driving users away. Hybrid IT solutions will be created by integrating cloud-based software products with on-premises computers [4].

The purpose of this paper is to propose a proactive traffic prediction model for network managers to keep abreast of congestion as well as accurately estimate future resource requirements. In this paper, a propose conceptual solution to the problem of predicting traffic flow in cloud computing by analyzing the three techniques, linear regression, time series decomposition, and auto-regressive integrated moving average.to predict the traffic. This paper is organized into four sections: introduction, related work, methods, results, and conclusions.

II. RELATED WORK

A cloud computing application, software, or service is the one that operates over the internet rather than on a remote server [6]. Businesses are increasingly utilizing cloud technology to simplify their existing operations, store and host data, and deploy apps, among other things. As a result, cloud traffic flow increases, making it crucial to predict the future of cloud traffic to manage resources properly. The combination of artificial neural networks (ANNs) and dynamic parameter methods (DPMs) increased accuracy, but when data dimensions were low, results were poor [2]. according to Fatima [13] Artificial Neural Networks and Long Short-Term Memory (ANN-LSTM) require a lot of time for training in traffic prediction, which delays data processing, which delays traffic prediction in cloud computing. As a result, this model is not suitable for real-time traffic prediction. To improve network traffic prediction accuracy, simulation-annealing-optimized autoregressive integrated moving average model back propagation neural networks (SA-optimized ARIMA-BPNN) [12] have been developed. It is unclear which field will lead to the future of optimizing network infrastructure architecture.

According to Fana [15] a simple ARIMA model can be used to forecast the monetary fund. Overall, the model improves the accuracy of traffic flow predictions. However, there are still some differences between the predicted and test data. There is difficulty in exploring and utilizing the correlation mechanism between traffic parameters using the Traffic Factor State Network (TFSN) framework [14]. A convolutional neural network-based bidirectional spatial-temporal network (CNN-BDSTN) is proposed. The potential accuracy of the network has been delayed until the validation results for accuracy are available. A time series decomposition forecast was proposed, and the results demonstrate that it improves accuracy by reducing errors by 36.60%, but it requires a lot of training [18].

ARIMA-RTS (Rauch Tung Striebel) technique can eliminates asynchrony and hysteresis, predicting errors caused by stimulation can be effectively reduced. According to Vidya [17] A Gaussian Process Regression Model (GPR) can predict accurately with very little data, and the training and prediction process is much simpler and faster. However, this method is not suitable for interactive real-time prediction due to the inconsistence of the data. Spatio-temporal Shared GRU (STSGRU) [19] improves the ability of the model to recognize traffic flow patterns. Users experience a delay as the model only focuses on weekly data. According to Ding [5], one limitation of linear regression is that it can't be conducted through such many cloud computing data centers, since the method experiments cannot be conducted through such a large number.

III METHODS

As shown in Figure 1, we collected data from organizations using cloud computing for the research study. Data preprocessing involves cleaning the data, removing bank columns, and transforming the data. The simulation process requires all methods (autoregressive integrated moving averages, linear regression, and time series decomposition) to process the same data and evaluate each model for accuracy.

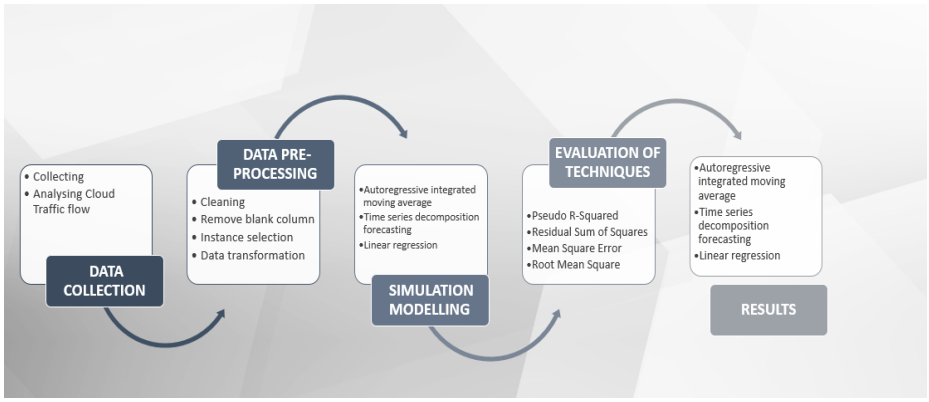

Fig 1: Proposed process

The article selected the Autoregressive Integrated Moving Average, linear regression, and time series decomposition methods as proposed techniques from the related works in Section 2.

A. Autoregressive integrated moving averages (ARIMA)

The ARIMA model was first developed by Box and Jenkins in 1970 [9]. ARIMA's flexibility with a wide variety of time-series data types, along with its accuracy at predicting outcomes, are generally favored by users [8], ARIMA combines AR and MA models as well as differencing. Regression models use past values to predict future values, while moving average models use previous residuals to predict future events. A basic process is to:

$$T_t = \theta_0 + \emptyset_1 D_{t-1} + \emptyset_2 D_{t-2} + \cdots + \emptyset_a D_{t-a} + V_t - \emptyset_1 V_{t-1} - \theta_2 V_{t-2} - \cdots - \theta_c E_{t-c} \quad (1)$$

Where, T_t is the actual viewed value at time t and V_t is random error. $\emptyset_1$ (I = 1, 2, ⋯, a) and θ_a(a = 0, 1, 2, ⋯, c) It is important to note that parameters a and c indicate the order in which the model runs. Random errors are independent and non-correlated, with zero mean and constant variance.

B. Linear regression method

By using linear regression analysis, it is possible to determine the value of a variable depending on another variable's value [7], below is a standard liner regression formula.

$$F_i = g(X_i, \beta) + e_i \quad (2)$$

Where, F_i = Variable dependent, g = Function, X_i = An independent variable, β= Parameters not known e_i= errors terms

C. Time series decomposition method

The decomposition of a time series is frequently performed before predictions, but if you know how to structure the series, you can use it to predict on its own [8]. Time series can be decomposed in two ways:

1. Multiplicative: $X(t) = Z(t) * P(t) * V(t)$ (3)
2. Additive: $X(t) = Z(t) + P(t) + V(t)$ (4)

Where $X(t)$ It consists of raw data, $Z(t)$ is the trend-cycle component at time t, $P(t)$ an indicator of seasonality at a particular point time t and $V(t)$ is the residual component at time t.

IV RESULTS

A. *Autoregressive integrated moving averages (ARIMA)*

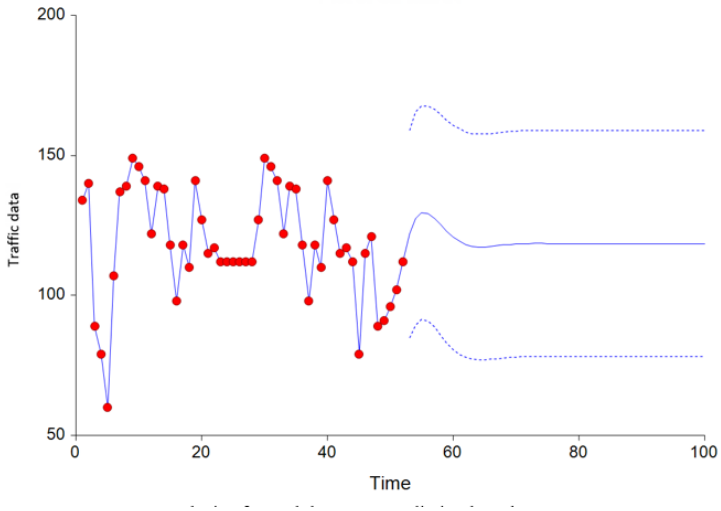

Fig 3: An analysis of actual data vs. a prediction based on ARIMA.

ARIMA results indicated by Figure 3 are 96% accurate. As cloud traffic increases, error evaluations also increase, while traffic decreases by almost half, which results in a lower error evaluation.

B. *Linear regression method*

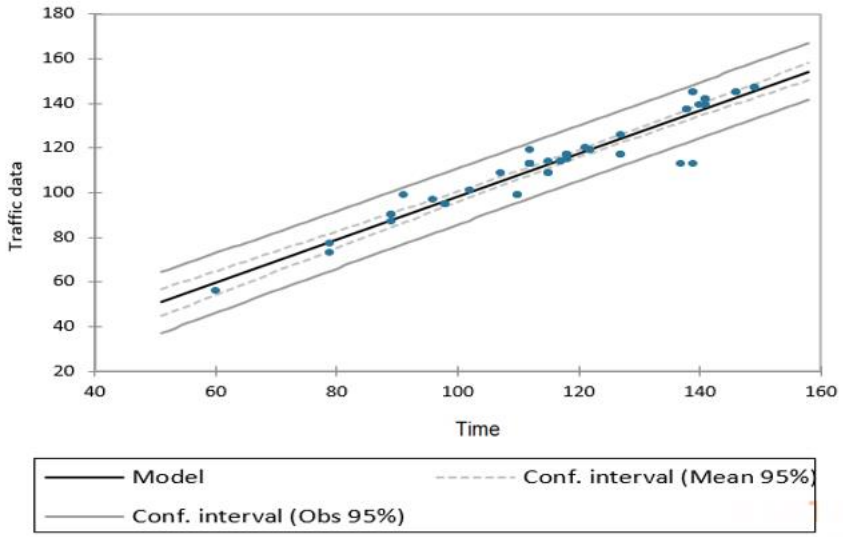

Fig 5: An analysis of actual data vs. a prediction based on linear regression.

A linear regression model showed a 95% accuracy in Figure 5. This technique produced better results when traffic increased.

C. Time series decomposition method

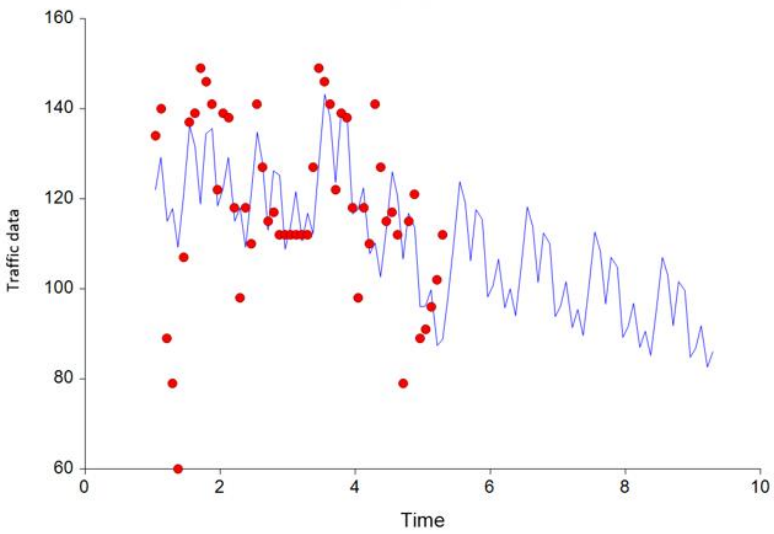

Fig 4: An analysis of actual data vs. a prediction based on Time series decomposition forecasting.

As shown in Figure 4, the time series decomposition forecasting model produced an accuracy of 69 %. When traffic decreased, the model produced better results, but when traffic increased, incorrect results were generated.

IV EVALUATION MEASURES

TABLE I. EVALUATION MEASURES

Prediction Tool	Evaluation Measures for Techniques			
	Root Mean Square	***Mean Square Error***	***Residual Sum of Squares***	***Pseudo R-Squared***
Autoregressive integrated moving average	14,20	262,37	10856	33,81
Time series decomposition forecasting	45,22	118	13300	39,40
Linear regression	15,09	186	11498	34,91

Table 1 indicates a lower root mean square for autoregressive integrated moving averages at 14.2% compared to linear regression at 15.09%, but the time series decomposition forecast root mean square was very high at 45.22%. Furthermore, the mean square error for autoregressive integrated moving averages was 262, followed by linear regression at 186 and time series decomposition at 118.

A smaller residual sum of squares indicates a better fit for the data. For the cloud data, autoregressive integrated moving averages were lower than other methods with 10856, followed by linear regression with 11498 and time series decomposition with 13300. However, pseudo-R-squared for autoregressive integrated moving averages was 33.81, followed by linear regression at 34.91% and time series decomposition forecasting at about 39.40%. The

autoregressive integrated moving average in this study outperformed linear regression and time series decomposition in terms of overall results.

V CONCLUSIONS

A cloud computing traffic prediction tool is proposed in this paper using autoregressive integrated moving averages, linear regression, and time series decomposition. As a result of this study, ARIMA is better than linear regression and time series decomposition at predicting cloud computing traffic.96% and 95% of ARIMA results and linear regression results were very similar. Time series decomposition forecasting model produced an accuracy of 69 %. Further work should consider using ensemble methods combining multiple models to produce better results.

ACKNOWLEDGMENT

The paper is part of the North-West University PhD program in Computer and Information Sciences with Information Technology.

CONFLICT OF INTEREST DECLARATION

There are no known financial interests or personal relationships that could have affected the authors' findings.

DATA AVAILABILITY

Data will be made available upon request.

REFERENCES

[1] I.A.Saroit, D.Tarek, "LBCC-Hung: A load balancing protocol for cloud computing based on Hungarian method,Egyptian Informatics Journal 24 (2023)", https://doi.org/10.1016/j.eij.2023.100387.

[2] M. Zheng, et al., "Do firms adopting cloud computing technology exhibit higher future performance? A textual analysis approach ".Journal of International Review of Financial Analysis, 90 (2023).https://doi.org/10.1016/j.irfa.2023.102866

[3] B.Godavarthi, et al.," Cloud computing enabled business model innovation", Journal of High Technology Management Research,34 (2023) https://doi.org/10.1016/j.hitech.2023.100469

[4] H.Materwala, L.Ismail,Hassanein,H.,S. " QoS-SLA-aware adaptive genetic algorithm for multi-request offloading in integrated edge-cloud computing in Internet of vehicles"Journal of Vehicular Communications,43 (2023) https://doi.org/10.1016/j.vehcom.2023.100654

[5] A.Ding ,M.Haizhou, "Method for evaluation on energy consumption of cloud computing data center based on deep reinforcement learning ". Journal of Electric Power Systems Research,208 (2022), https://doi.org/10.1016/j.epsr.2022.107899

[6] Y.Li,J.Liu,Y.Teng, "A decomposition-based memetic neural architecture search algorithm for univariate time series forecasting ",journal or Applied Soft Computing,130 (2022) https://doi.org/10.1016/j.asoc.2022.109714.

[7] R.Chaudhari, V.Venkatraman,.C.R.Kant, "Analysis of photocurrent trends in hybrid PS-BiI3 composites for direct X-ray detector via linear regression model ", journal of Composites Communications https://doi.org/10.1016/j.coco.2023.101681

[8] F.,Zhang,T.,Guo,H.Wang, "DFNet: Decomposition fusion model for long sequence time-series forecasting",Journal of Knowledge-Based Systems,277 (2023) https://doi.org/10.1016/j.knosys.2023.110794

[9] G.E.P. Box, G. Jenkins, " Time Series Analysis, Forecasting and Control, Holden-Day", San Francisco, CA, 1970.

[10] G.P.Zhang, "Time series forecasting using a hybrid ARIMA and neural network Model", Neurocomputing 50 (2003) 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

[11] K.W. Wang, et al., "Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network", Epidemiol. Infect. 145 (6) (2017) 1118–1129.

[12] K.Benmouiza, A. Cheknane, Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models, Theor. Appl. Climatol. 124 (3-4) (2016) pp.945–958.

[13] A.A.Fatima, M. Ghurab, "ANN-LSTM: A deep learning model for early student performance prediction" in MOOC Journal of Heliyon,Volume 9, Issue 4, April 2023 https://doi.org/10.1016/j.heliyon.2023.e15382

[14] W. Fana," Prediction of Monetary Fund Based on ARIMA Model". Journal of Procedia Computer Science vol.208 277–285(2022), https://doi.org/10.1016/j.procs.2022.10.040.

[15] L.Junyi, G.Fangce G. Sivakumar,D.Yanjie,R. Krishnan, "Transferability improvement in short-term traffic prediction using stacked LSTM network". Journal of Transportation Research Part C: Emerging Technologies Vol 124 (2021), https://doi.org/10.1016/j.trc.2021.102977.

[16] H.Lin, et al., "Traffic Flow Prediction Using SPGAPSO-CKRVM Model". Journal of Revue d'Intelligence Artificielle Vol. 34,257-265(2020), https://doi.org/10.18280/ria.340303.

[17] A' Parslov et al., "Short-term bus travel time prediction for transfer synchronization with intelligent uncertainty handling" journal of Expert Systems with Applications, 322 (2023), https://doi.org/10.1016/j.eswa.2023.120751

[18] X. Sun et al.,"Short-term traffic flow prediction model based on a shared weight gate recurrent unit neural network" journal of Physica A: Statistical Mechanics and its Applications,Volume 618, 15 May 2023 https://doi.org/10.1016/j.physa.2023.128650

[19] I.A.Saroit D.Tarek, "LBCC-Hung: A load balancing protocol for cloud computing based on Hungarian method ", Egyptian Informatics Journal 24 (2023),https://doi.org/10.1016/j.eij.2023.100387