I.J. Information Technology and Computer Science, **2024, 6, 74-88**
Published Online on December 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2024.06.06

Ranking of Machine Learning Algorithms Used in Disease Prediction: A Review-based Approach

Shunmuga Priya Subramanian*
Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil – 626126, Tamilnadu, India
E-mail: s.shunmugapriya@klu.ac.in
ORCID iD: https://orcid.org/0000-0002-4802-1367
*Corresponding author

Amuthaguka Duraipandian
Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil – 626126, Tamilnadu, India
E-mail: d.amuthaguka@klu.ac.in
ORCID iD: https://orcid.org/0000-0002-3906-148X

Received: 17 August 2024; Revised: 20 September 2024; Accepted: 17 October 2024; Published: 08 December 2024

Abstract: There are remarkable improvements in the healthcare sector particularly in patient care, maintaining and protecting the data, and saving administrative and operating costs, etc. Among the various functions in the healthcare sector, disease diagnosis is considered as the foremost function because it saves a life at the correct time. Early detection of diseases helps in disease prevention, letting the patients get vigorous and effective treatment and saving their lives. Several techniques were suggested by the researchers for disease prediction. Many literatures have been witnessed on disease prediction. This article reviews several articles systematically and compares various machine learning (ML) algorithms for disease prediction, including the Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR) algorithms. A thorough analysis is presented based on the number of publications year-wise, disease-wise, and also based on the performance metrics. This review thoroughly analyzes and compares various ML techniques applied in disease prediction, focusing on classification algorithms commonly employed in healthcare applications. From the systematic review, a multi objective optimization method named Grey Relational Analysis (GRA) is used to rank the ML algorithms using their performance metrics. The results of this paper help the researchers to have an insight into the disease prediction domain. Also, the performance of various ML algorithms aids the researchers to choose a better methodology to predict a disease.

Index Terms: Machine Learning, Disease Prediction, Review, Grey Relational Analysis.

1. Introduction

The present state of healthcare demands the collection and maintenance of databases. Patients' data contain a wide range of diagnostics and sickness features, and datasets permit the storage of data for more patients in a single file. Since healthcare databases contain redundant data and missing information, the healthcare data must be handled with caution. Recently, machine learning is gained popularity as a mainstream division of artificial intelligence (AI) due to its autonomous feature learning capabilities, increased accuracy rates, and improved performance. Because of the consistency and precision of noise removal of these techniques, disease prediction is easier and faster for data mining [1]. Haleem et al., [2] explored the revolution in the healthcare systems using digital twin technology. Digital twins utilize massive volumes of data from numerous IoT (Internet of Things) devices and AI powered models. A patient's personal digital twin uses information from historical data to predict the outcomes of certain surgeries, help choose the best medications and manage chronic disease. The data for training the ML technique were obtained from the medical centers at the University of California and Beth Israel Deaconess. The parameters considered were temperature, heart beat rate, diastolic blood pressure, blood oxygen level, and respiration rate. AI and ML can be utilized to recognize, examine, forecast, and categorize medical data [3]. In the previous 10 years, several AI algorithms have been successfully implemented in data processing in the healthcare sector [4]. This article's main goal is to increase understanding and awareness among researchers and medical practitioners about the recent and ongoing trend in disease prediction. Several

ML techniques, including the LR, SVM, DT, RF, and NB algorithms, are used to categorize the papers examined in this research. In addition to that the review on comparison of various algorithms is also presented.

2. Review on Machine Learning Approaches

Machine Learning is a methodology that covers a broad variety of methodologies for diagnosing and predicting the issues in a variety of healthcare fields [5]. For the prevention and treatment of any disease, accurate and timely examination and identification of the causes is important. In the case of a life-threatening disease, the traditional regular diagnosis procedure may not be adequate. The advancement of the healthcare diagnosis system using ML can assist doctors and medical professionals in a more accurate manner than the existing method [6]. ML is gradually becoming more popular in the medical diagnosis field [7]. ML has been successfully implemented to find the solution to biological problems [8]. ML can enrich the quality of medical data and lessen the medical costs. Hence when compared to other traditional methodologies, ML models are widely used in disease diagnosis [9]. ML is applied in the drug development and prediction of antibiotic interactions [10]. Commonly both unsupervised and supervised learning methodologies are used in the majority of medical fields. Due to the benefits in the data handling quality, data analysis time, calculating the learning features, and accuracy. ML based techniques were vastly applied in the disease prediction domain [11]. Medical data can be analyzed faster by using the ML models as compared with the conventional methods. So, there is a remarkable saving on medical costs [12]. A few machines learning (ML) algorithms used in healthcare prediction include LR, SVM, DT, RF, and NB algorithms [13]. Mahabub [14] has developed a robust voting methodology to predict diabetes by using eleven different ML algorithms. The various parameters such as precision, accuracy, f-measure, etc. were evaluated. Jeon et al., [15] predicted blood glucose levels using ML techniques and investigated the predictive accuracy of various ML methods to assign missing data.

2.1. Review on Logistic Regression (LR) Algorithm

This machine learning approach uses a set of accessible independent features to predict a certain dependent variable. LR is one type of supervised learning technique of ML. It was initially formulated by David Cox in 1958 [16]. Khandezamin [16] used the LR approach to classify the dataset. In that study, the dataset from the Wisconsin Diagnostic Breast Cancer Center was taken into account. The classification was performed by using two parameters such as mean texture and mean radius. 90.48 % accuracy was obtained when picking the maximum texture and when choosing the maximum radius, the accuracy was obtained as 96.5 %. Owen et al., [17] suggested the LR technique to quantify vascular variations of the bulbar conjunctiva. The study demonstrated adequate sensitivity in evaluating increased hyperemia in rigid contact lens wearers than the soft lens users. After the capturing of photographic images, the area of interest was manually extracted. The high contrast vasculature was obtained by applying spatial averaging and later contrast enhancement methods were applied to get better visualization of images. The result included the repeatability factor and the work compared the objective approach proposed with the subjective assessment. Brea et al., [18] used LR based classification and regression techniques of machine learning in grading the hyperemia. The work concluded that though the classification approaches resulted in low training error, it was the regression techniques that correctly classified the majority of the cases. As not all values of the standard grading scales were used by practitioners, so the study considered those few values as classes for classification purposes. The prime focus of the work was to establish how the classification algorithm works in converting the features extracted into clinical grades.

To forecast breast cancer, Nandagopal et al., [19] employed LR method to classify the cancer data by using the victim's gene. Among the dataset of 116 breast cancer patients, the average classification accuracy was closer to 94 %, and productivity with a lower error rate was attained by using the model. Latha and Jeeva [20] implemented the LR classification procedure to predict heart disease. To forecast heart disease, Desai et al. [21] used the LR technique and compared the results with those generated by artificial neural networks (ANNs). Tania and Raymond [22] predicted the foremost factors initiating cardiovascular disease. A dataset having 14 features with class labels was considered to identify the link between factors that cause cardiovascular disease. LR algorithm has been chosen to identify the relationship among the dependent variables and the independent variables considered. Hossain et al., [23] presented a model for risk prediction using LR based ML technique to identify the patients' risk conditions for heart diseases. Ambrish et al., [24] applied the LR technique to the UC Irvine dataset for the cardiac disease classification. The output of the suggested technique was improved by data pre-processing. The feature selection was done by the relationship using the target value for all the attributes and the higher associated attributes were obtained. After that, the classification was achieved by separating the dataset namely the training dataset and the testing dataset with five various ratios such as 50:50, 60:40, 70:30, 80:20, and 90:10. The training and testing ratio of 90:10 produced better accuracy of 87.10%. Several researchers [25-28] developed LR based prediction model for COVID-19 disease and its spread.

2.2. Support Vector Machine (SVM) Algorithm

Russian Professors Vladimir Vapnik and Alexei Chervonenkis developed the SVM algorithm in 1963 [29]. SVM is another effective supervised learning technique, which is typically utilized to solve regression and classification problems. SVM has been used for data classification by Son et al. [30]. The data included 11 variables that have been received from 76 patients in a hospital using a survey form. Mathematical simulations were performed to develop an SVM model to identify the features that would best forecast the disease. To assess the strength of the forecasts prepared using the SVM technique. The validation was done by Leave One Out Cross-Validation (LOOCV) on the dataset. Through this method, the prediction accuracy was obtained as 77.63%. Yu et al., [31] presented an SVM based approach to classify people having common diseases. The data have been received from the survey on National Health and Nutrition Examination in USA between 1999-2004. The classification model was proposed to classify the persons with pre-diabetes and diabetes among the sample of U.S. residents. For the classification, family history, age, weight, body mass index, waist circumference, height, race and ethnicity, and hypertension were considered as variables. Through SVM models 83.5% accuracy was attained. Huang et al., [32] have stressed the importance of the SVM classification algorithm for effective tumor tissue classification and are significantly involved in lymphoma.

Farooqui and Ahmad [33] have proposed a system to predict possible diseases based on symptoms. SVM based ML algorithm was proposed to predict the probable diseases. The prediction outcome of that SVM system has 87% accuracy. Owusu et al [34] proposed the SVM technique as a computer-aided investigative tool for predicting heart disease risk levels. The datasets considered in this study contained 13 factors namely age, gender, blood pressure level, and type of chest pain. This case study was conducted among 303 records taken from the Cleveland Clinic. SVM was applied to train and test the data. The results were compared with LR, NB, DT, and RF algorithms. Puri et al., [35] have recommended an SVM algorithm based model for heart stroke prediction by considering the attributes namely age, past history, blood pressure, average body glucose level, body mass index, and rate of smoking status. The prediction model is established using the SVM algorithm. The results proved that the SVM has achieved better in heart stroke prediction with superior accuracy values. Madhu and Ramesh [36] proposed SVM based supervised model to predict heart attacks. This study considered the following characteristics: heart rate, blood pressure, type of chest discomfort, and cholesterol level. The University of California, Irvine cardiac disease dataset was utilized to validate the model. The accuracy obtained was 83%. Guhathakurata et al., [37] have exercised the SVM based classification to classify the patient's COVID status into severe infection, minor infection, and null infection classifications. The COVID cases were predicted using SVM with 87% accuracy. The parameters considered were body temperature (> 103°C - high fever), cough, sputum, breathing rate (> 30 breath per minute - critical condition), hypertension (from 140 to 159 blood pressure level - stage 1; above 160 - stage 2), heartbeat rate (more than 100 beats per minute – high), chest pain (continuing chest pain in addition to respiratory problem is dangerous), and cardiac disease (patient having existing cardiovascular diseases is liable to COVID-19). Sarra et al., [38] recommended a SVM based heart disease classification. The SVM algorithm enhanced the accuracy of prediction from 85.29% to 89.7%.

2.3. Decision Tree (DT) Algorithm

DT is a supervised ML technique and is capable of solving classification and also regression problems [39]. Even though, most of the researchers proposed DT in the classification applications. This classifier resembles a tree structure. The internal nodes, branches, and leaf nodes denote the attributes of a dataset, the decision guidelines and the result respectively. Several DT algorithms such as Hunt's algorithm, ID3 algorithm, and C4.5 algorithm are used in the classification and prediction problems. Hunt's algorithm was developed in the 1960s by Hunt, the ID3 algorithm and C4.5 algorithms were introduced by J. Ross Quinlan in the late 70s and 80s respectively [40]. Bhatla and Jyoti [11] have presented a thorough survey of several data mining approaches that were used in computerized prediction systems for heart disease. The assessment demonstrated that the DT algorithm produced 99.62 % precision and the Neural Network has shown the greatest precision (100 %). Howlader et al., [41] have applied the DT algorithm in the prediction of heart diseases and proved that DT predicted the disease with 100% accuracy. Purushottam et al., [42] have designed a prediction system to envisage the threat level of patients by using their health parameters. The rules were arranged depending on the user's desires. DT was used for the prediction. The outcomes proved that the system abundantly predicted the heart disease threat level exactly. Bahrami et al., [43] implemented the DT algorithm and NB algorithm to classify datasets for diagnosing heart disease. The findings are compared by metrics like precision, specificity, area under ROC curve, accuracy, sensitivity, and F-measure. The results have been matched with the above algorithm and proved that J48 DT is the preeminent classifier for diagnosing heart disease by using the available dataset. Subburaj and Ramu [44] proposed an intelligent prediction system to forecast heart diseases using the DT algorithm. Several symptoms of heart diseases were fed as inputs and the ID3 algorithm was used for the prediction.

Ahmad et al., [45] explored the application of the DT algorithm to predict COVID cases. The several DT ensembles were compared using accuracy, recall, area below the precision-recall curve, F-measure, and area below the receiver operating characteristic curve. Based on those performance metrics, it appeared that DT ensembles designed for datasets with imbalances were operating more effectively. Kapoor and Singh [46] have predicted heart disease using the DT algorithm. Reddy et al., [47] predicted cardiovascular disease by applying DT based ML technique by considering the cardiac failure dataset with 13 attributes. Initially, the collected data should be pre-processed and the feature selection and reduction were performed later. DT predicted the disease with 82 % accuracy. Sateesh and Balamanigandan [48] used various medical parameters of several cardiac patients to determine heart diseases. DT was applied to enrich the accuracy of heart disease forecasts well in advance. The work was executed with five variety of datasets to record five samples at each time. The DT algorithm has shown a higher accuracy of 87.75%. Saraswathi et al., [49] have suggested the DT algorithm for diagnosing cardiac diseases. Age, blood pressure, blood sugar, and cholesterol were taken into account when predicting a person's risk of having a heart attack.

2.4. Random Forest (RF) Algorithm

RF algorithm is also one of the supervised ML methods that is commonly utilized for classification problems. In addition, it can also be used in regression problems. RF was initially developed in 1995 by T. K. Ho and improved by Breiman in 2001 [50]. Szűcs [51] used the RF framework in a range of real-time medical prediction applications. Besides, the RF technique offers a variety of applications like decision making processes and prediction models. Using the RF approach, Ooka et al. [52] examined the predictive parameters and risk prediction of type 2 diabetes. The RF algorithm was employed to forecast the risk of cancer [53]. The Dataset collected from the breast cancer database available at the Wisconsin diagnosis laboratory was utilized to train the RF algorithm to assess the performance of disease prediction with reference to precision and accuracy. Bharati et al., [54] identified breast cancer by data mining data from individuals who have already been diagnosed with the condition. The UCI repository provided the dataset on breast cancer disease, which was then classified and predicted using LR, RF, NB, K-Nearest Neighbor (KNN), and Multilayer Perceptron.

Wu et al., [55] recommended RF based classification algorithm to identify fatty liver disease (FLD) patients using the bare minimum of clinical information. This technique might direct in greater insights into real-world clinical practice, allowing doctors to detect FLD more successfully for innovative diagnostic, preventive, and therapeutic objectives, lowering the disease's worldwide impact. Future research is needed to prove our model's capacity to predict FLD over a wide range of datasets. A study by Gates et al., [56] acquired the best textural analysis-based classification model for glioma grading using random forest algorithm. Ma et al., [57] created a prediction model using the RF technique for predicting clinical illness. Mohan et al., [58] have developed a framework to study the problem of constraining and summarizing various algorithms relating to data mining. Several attributes for the prediction of combinations of different attributes were focused. Finally, a prediction model was developed to analyze the heart attack. The result produces a high level of prediction accuracy and helps the physician to properly diagnose heart disease. The RF prediction model produced 88.7% accuracy.

Escamilla et al. [59] introduced a unique feature selection method for discovering critical variables in categorizing cardiac disease by merging chi-square and principal component analysis (CHI-PCA). The UCI database provided the dataset used in this investigation. The RF classifier selected the features with an accuracy of 98.7%. Shah et al., [60] developed an RF prediction model to forecast Cardiovascular Disease. Initially, the data pre-processing was performed for the removal of noisy data and the removal of missing data. Then the classification of attributes for prediction was done by using the RF algorithm. Pal and Parija [61] predicted the existence of heart disease among the patients using RF technique. The Kaggle dataset was used in the study. Totally 303 samples with 14 attributes were chosen from the dataset. The datasets were classified and predicted using RF based ML algorithm. Using the RF method, the accuracy, sensitivity, and specificity in percentage were determined. RF produced 86.9 % accuracy with 90.6% sensitivity and a specificity value of 82.7%. Jackins et al., [62] used RF based ML classification procedure to classify the various disease datasets such as heart disease, diabetes, and cancer to predict whether the patients were infected by those diseases or not. Shanthakumari et al., [63] have recommended RF based classification technique to identify several diseases using the patient's dataset.

2.5. Naive Bayes (NB) Algorithm

NB algorithm is working based on the principles of Bayes' Theorem. It is not an individual algorithm, but a group of algorithms where all share a common rule, i.e. all the pairs of attributes are being classified as independent of each other. Normally the Bayesian method works based on probability theory. Reddy et al., [47] predicted cardiovascular disease by using Naïve Bayes method. The heart failure dataset with 13 attributes was chosen for the prediction case study. The results were achieved by pre-processing the data and then by the feature selection and reduction. Naïve Bayes predicted the disease with 86 % accuracy. Wei et al., [64] have predicted brain disorders by implementing the NB algorithm. The results proved that the NB algorithm predicted the disease with 72% accuracy. Wang et al., [65] used NB-based ML to predict lung cancer. The data were collected from a Taiwanese database containing around 50,000 cancer patients between 1996 and 2010. Miranda et al., [66] improved the prognostic accuracy by using the NB classifier in the categorization of heart disease data. A dissertation procedure and genetic examination were employed in the elimination of fired features. A comparison had also been made over other techniques that endeavored to improve NB classifier achievement in classifying heart disease.

Singh et al., [67] framed a classification model using the NB algorithm and ANN. The NB classification method was suggested for the development of a decision support system. It was also established that the statistics based techniques needed modifications when used by data mining methods. Bzdok et al., [68] explained the use of NB based supervised learning algorithms for classification. It is one of the simplest and most effective ML algorithms out now, with uses in a variety of industries. In the NB technique, all predictors should be independent. As a result of this shortcoming, the algorithm's application in real-world situations is restricted. This zero-frequency issue can be solved using a smoothing method. Another downside of the Naive Bayes method is that it performs best with large datasets. Dulhare, [69] used the NB algorithm to create a model for predicting cardiac disease. The presence of heart disease was predicted using a range of input parameters, such as the patient's age, gender, type of chest pain, blood pressure at rest, blood cholesterol level, blood sugar level during fasting, results of the resting electrocardiogram, heart rate attained, exercise-induced angina, and depression-induced angina. The developed model predicted the disease with 87.9% accuracy. Anitha et al., [70] implemented an NB based healthcare prediction model to forecast the disease in the beginning stage. In that study, the experiment involving supervised machine learning techniques was conducted using the R programming language. While predicting cardiac illness, the NB method outperformed SVM and KNN, achieving an accuracy of 86.6%, compared to 77.7% and 76.67% accuracy, respectively. Hemanth and Pavani [71] described NB based techniques to diagnose coronavirus. Sore throat and cough, fever, runny nose, and difficulty in breathing were taken as input parameters.

2.6. Comparison of Algorithms

Kaur and Kumari [5] used supervised ML algorithms namely SVM, KNN, ANN, radial basis function (RBF), and multifactor dimensionality reduction (MDR) for the classification of patients into two categories, such as with and without diabetes. The sensitivity of these models was obtained as 87%, 90%, 88%, 83%, and 87% respectively. The precision of the models was found to be 88%, 87%, 85%, 85%, and 82% respectively. F1 scores of the models were 0.87, 0.88, 0.86, 0.83, and 0.84 respectively. The area under the curve (AUC) to measure the performance of those models was 0.90, 0.92 0.88, 0.85, and 0.89. Finally, it was resolved that KNN achieved well as compared with other models in the diabetic prediction model. Bahrami et al., [43] have used various classifier techniques namely DT, NB, and KNN to classify datasets to diagnose heart disease. The results were compared by metrics such as precision, specificity, area under ROC curve, accuracy, sensitivity, and F-measure. The outcomes were compared with the above algorithm and concluded that DT was performing well and became the best classifier for diagnosing heart disease.

Das [72] compared four techniques of classification to predict Parkinson's disease. ANN, LR, and DT were used as classification models. ANN generated 92.9 percent of the highest prediction rate. Wu and Guo [73] provided three models for analyzing Parkinson's disease such as LR, DT, and ANN. Naraei et al., [74] have made a comparison between ANN and SVM techniques on the heart diseases dataset. The efficiency had been examined by 303 patient's datasets. The consequences revealed that SVM was capable of categorizing more precisely. Tapak et al., [75] utilized six various ML techniques and two conventional methods to compare and predict breast cancer metastasis and survival, including LR, SVM, RF, NB, and AdaBoost. Totally 550 breast cancer patient's databases were used in that investigation. These methodologies were used to assess performance metrics like accuracy, specificity, sensitivity, and likelihood ratio. As a result, RF produced the highest specificity (98%), SVM gave the best accuracy (93%) and the highest sensitivity was produced by NB (86%). Wu et al., [55] developed an ML model for liver disease prediction that can be used to accurately diagnose patients with greater risk factors. The data used in the study were gathered from January 1 to December 31, 2009, from New Taipei City Hospital. A variety of classification models were used, including NB, LR, RF, and ANN. The achievement of the selected ML models was assessed and found that the accuracy of NB, LR, RF, NB, and ANN was observed as 81.85, 76.96, 87.48, and 82.65% respectively.

Christodoulou et al., [76] comparatively evaluated how well machine learning techniques work in forecasting diseases. Data for the analysis were collected from Medline literature search between January 1, 2016 and August 8, 2017. The observed sample size spans from 1,250 to 3,994,872, with 19 predictor variables ranging from 5 to 563, and 8 predictor ranges ranging from 0.3 to 6,697. This study employed several ML algorithms namely LR, DT, SVM, RF, and ANN. The LR model performed better than other algorithms. Almustafa [77] carried out a study to classify heart disease even with a smaller amount of number of features. A comparison analysis using a dataset on heart illness was conducted with different classification techniques, including KNN, NB, DT, SVM, and AdaBoost. The undesirable attributes were removed by applying the Classifier Subset Evaluator. KNN achieved a better result of 99.70% accuracy as compared with other classifiers. Chiu et al., [78] have analyzed Chronic Kidney Disease (CKD) risk factors using the clinical data collected from 65,394 people between 2010 and 2015. Those data were obtained from a significant clinical testing and screening facility in Taiwan. For the analysis, 18 risk indicators were included in the dataset. Five ML based prediction models namely, LR model, DT model, stochastic gradient boosting (SGB) model, multivariate adaptive regression splines (MARS) model, and eXtreme gradient boosting (XGboost) models have been developed to determine the CKD. The accuracy of LR, SGB, and, XGboost models was obtained as 84.8%, 85.5% and 85.8% respectively. It was noted that those models produced very close classification performance levels and were better than the other two models.

Hassan et al., [79] have developed a prediction system to foresee heart disease with high accuracy. The performance of DT, SVM, and NB algorithms were compared. The outcomes proved that the DT algorithm is much superior to SVM and NB in terms of accuracy even with less training time also. Acheme and Vincent [80] compared DT, LR, and gradient boosting techniques used for predicting survivability in COVID-19 patients. The simulation's results demonstrated that the LR model's predictions were less accurate while the others predicted with 95% accuracy. Shanbehzadeh et al., [81] developed LR, DT, and RF models for prompt ICU admission prediction in COVID-19 patients when they were admitted to the hospital. The results proved that the LR model performed superior to other models. Cevik et al., [82] investigated ML based approaches for feature selection and highlighted the efficacy of active learning strategies. Two disease forecasting studies using RF and XGBoost were compared.

3. Observations and Discussions

3.1. Based on Articles Published

In this paper, several research articles which were published in the field of disease prediction using ML algorithms have been collected and reviewed systematically. The year-wise articles collected for the work is shown in Fig. 1. From Fig. 1, it is evident that around 75% papers considered are the latest articles published in 2018 and later years.

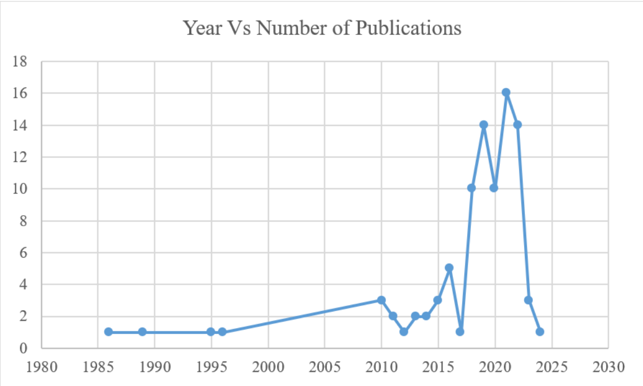

Fig.1. Year vs number of publications

The algorithms used in the prediction have been summarized and some observations based on these articles are made in Table 1. Most of the researchers used the ML algorithms to predict and categorize various diseases such as eye diseases, Parkinson diseases, liver diseases, cancer, heart diseases, diabetes, COVID-19, etc. The detailed analysis (percentage of literature on the prediction of various diseases using ML algorithms) is presented in Table 2 and Fig. 2.

Table 1. Literature vs machine learning vs disease prediction

ML Technique	COVID -19	Cancer	Heart Diseases	Diabetes	Others (Eye Diseases; Parkinson diseases; Liver diseases etc.,)
Logistic Regression (LR) algorithm	Majumder et al., [26]; Verghese et al., [27]; Kulenovic et al., [28]; Nopour et al., [29]; Acheme and Vincent [80]; Shanbehzadeh et al., [81]	Khandezamin et al. [19]; Nandagopal et al., [20]; Bharati et al., [54]; Tapak et al., [75]	Latha and Jeeva [21]; Desai et al., [22]; Tania and Raymond [23]; Hossain et al., [24]; Ambrish et al., [25]	Priyanka and Shahram [83]	Owen et al [17]; Brea et al [18]; Das [72]; Wu and Guo [73]; Christodoulou et al., [76]; Chiu et al., [78]
Support Vector Machine (SVM) algorithm	Guhathakurata et al., [38]	Huang et al., [33]; Tapak et al., [75]	Owusu et al., [35]; Puri et al., [36]; Madhu and Ramesh [37]; Sarra et al., [39]; Naraei et al., [74]; Almustafa [77]; Hassan et al., [79]	Kaur and Kumari [5]; Yu et al., [32]; Deepti and Dilip singh [84]	Son et al., [31]; Farooqui and Ahmad [34]; Christodoulou et al., [76]
Decision Tree (DT) algorithm	Ahmad et al., [45]; Acheme and Vincent [80]; Shanbehzadeh et al., [81]	Tapak et al., [75];	Bhatla and Jyoti [11]; Howlader et al., [41]; Purushottam et al., [42]; Bahrami et al., [43]; Subburaj and Ramu [44]; Kapoor and Singh [46]; Reddy et al., [47]; Sateesh and Balamanigandan [48]; Saraswathi et al., [49]; Almustafa [77]; Hassan et al., [79] Nai-arun and Moungmai [87]	Deepti and Dilip Singh [84]; Zou et al., [85]; Jayakumar et al., [86]	Das [72]; Wu and Guo [73]; Christodoulou et al., [76]; Chiu et al., [78]
Random Forest (RF) algorithm	Shanbehzadeh et al., [81]	Sharma et al., [53]; Bharati et al., [54]; Jackins et al., [62]; Tapak et al., [75]	Mohan et al., [58]; Escamilla et al., [59]; Pal and Parija [61]; Jackins et al., [62] Nai-arun and Moungmai [87]	Ooka et al., [52]; Zou et al., [85]; Alam et al., [88] Vijiya Kumar et al., [89]	Szűcs [51]; Christodoulou et al., [76]; Wu et al., [55]; Ma et al., [57]; Shanthakumari et al., [63]
Naive Bayes (NB) algorithm	Hemanth and Pavani [71]	Bharati et al., [54]; Wang et al., [65]	Bahrami et al., [43]; Reddy et al [47]; Miranda et al., [66]; Dulhare [69]; Anitha et al., [70]; Almustafa [77]; Hassan et al [79]	Deepti and Dilip Singh [84]	Wei et al., [64]; Singh et al., [67]; Bzdok et al., [68]

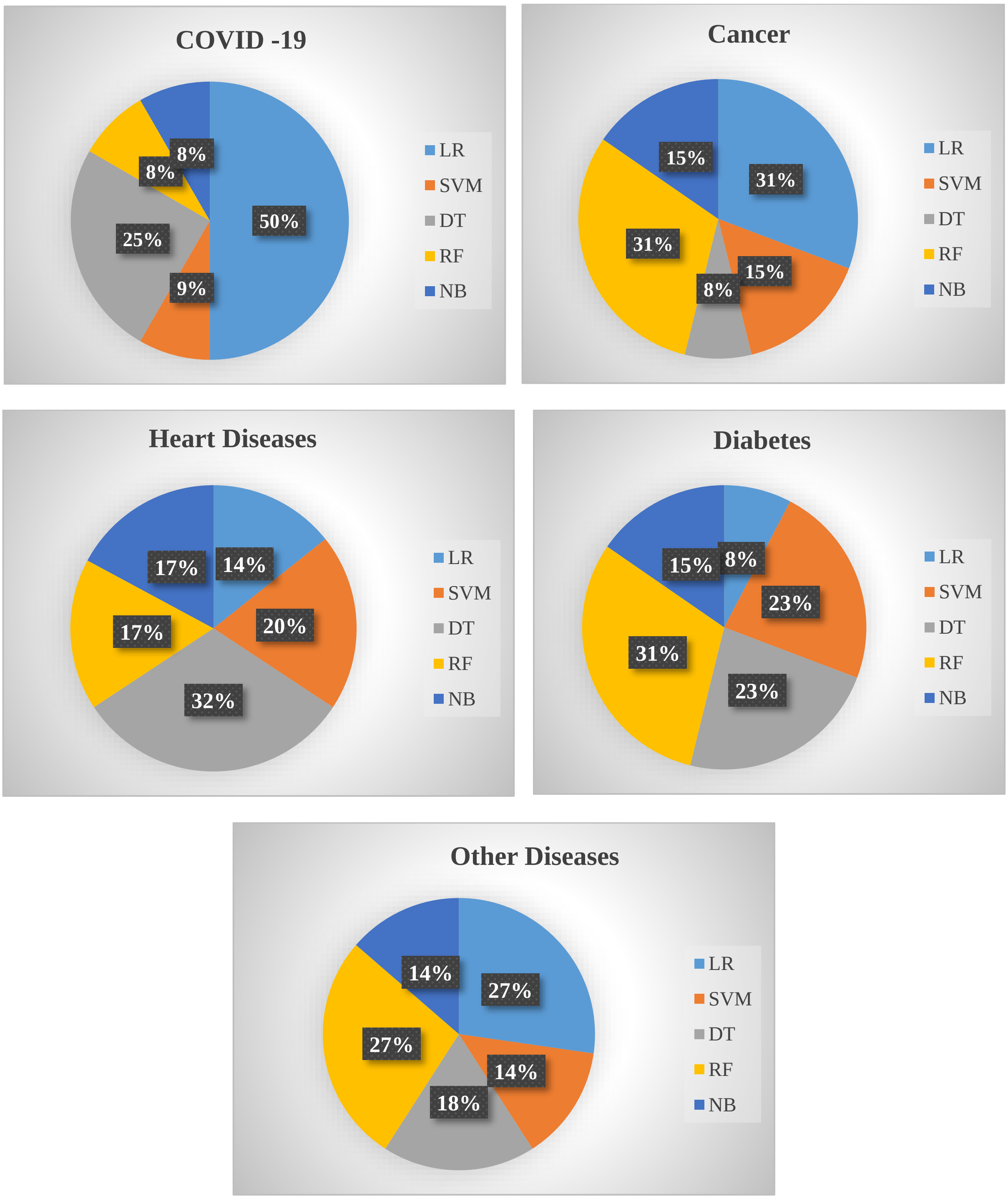

Fig.2. Percentage of literature vs machine learning vs disease prediction

Table 2. Percentage of literature vs machine learning vs disease prediction

ML	COVID -19	Cancer	Heart Diseases	Diabetes	Others (Eye Diseases; Parkinson diseases; Liver diseases etc.,)	Net
LR	50%	30.77	14.29%	7.69%	27.27%	23.16%
SVM	8.33%	15.38%	20%	23.08%	13.64%	16.84%
DT	25%	7.69%	31.43%	23.08%	18.18%	23.16%
RF	8.33%	30.77%	17.14%	30.77%	27.27%	22.11%
NB	8.33%	15.38%	17.14%	15.38%	13.64%	14.74%

Among the literature collected on the development of COVID-19 prediction models (Table 2), the LR algorithm was used by 50% of the researchers while SVM, DT, RF, and NB algorithms were used by 8.33%, 25%, 8.33%, and 8.33% respectively. As far as Cancer diseases are concerned the percentage of literature considered by LR, SVM, DT, RF, and NB algorithms are 30.77%, 15.38%, 7.69%, 30.77%, and 15.38% respectively. For various heart diseases, 14.29%, 20%, 31.43%, 17.14%, and 17.14% literatures were taken for study on LR, SVM, DT, RF, and NB algorithms. For the prediction of diabetes, LR, SVM, DT, RF, and NB algorithms were utilized in 7.69%, 23.08%, 23.08%, 30.77%, and 15.38% literatures. Similarly, the percentage of literature reviewed on the prediction of other diseases like Eye Diseases; Parkinson diseases; Liver diseases, etc., using LR, SVM, DT, RF, and NB algorithms are 27.27%, 13.64%, 18.18%, 27.27%, and 13.64% respectively. It is also observed that 13%, 14%, 37%, 12%, and 24% articles were published in the prediction of COVID-19, Cancer, Heart Diseases, Diabetes and Others (Eye Diseases; Parkinson diseases; Liver diseases, etc.,). Similarly, from the collected articles, it is concluded that 23.16%, 16.84%, 23.16%, 22.11%, and 14.74% articles used LR, SVM, DT, RF, and NB algorithms respectively (Fig. 3).

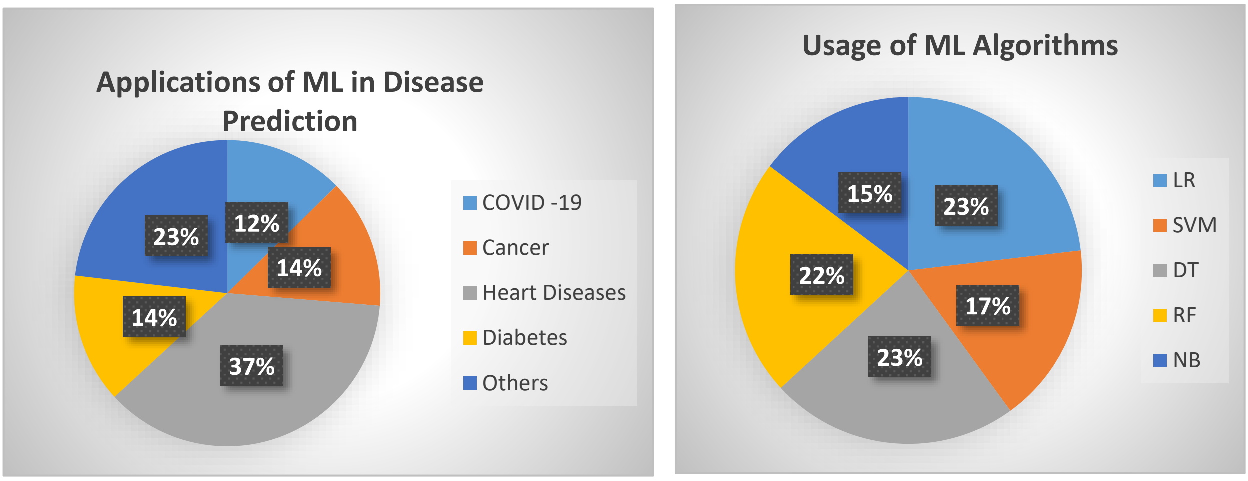

Fig.3. Summary of machine learning vs disease prediction

3.2. *Based on Performance Measures*

Majority of the literature published in the disease prediction domain used Precision (P), Accuracy (A), Specificity (Sp), F-measure (F), and Sensitivity (Sn) of the ML algorithms to evaluate, categorize, and predict the diseases. From the selected literature, the averages of the aforesaid performance metrics for various ML algorithms in the disease prediction domain is presented in Table 3 and Fig. 4.

From the table, the following inferences are made:

- SVM yields superior results in terms of F-measure and sensitivity
- RF predicts diseases with better accuracy and specificity
- NB is the better model for precision

These outcomes lead to confusion while selecting the ML technique for disease prediction. To overcome this problem, in this article, a multi objective optimization method is recommended to choose a better ML technique.

Table 3. Average performance measures vs ml algorithms

	Precision	Accuracy	Specificity	F-measure	Sensitivity
LR	0.8258	0.8426	0.8011	0.8403	0.7802
SVM	0.8560	0.8485	0.8391	0.8922	0.8378
DT	0.7557	0.8182	0.7416	0.6563	0.7529
RF	0.8347	0.8579	0.8487	0.8258	0.8130
NB	0.8629	0.7922	0.7961	0.8529	0.7579

3.3. *Limitations of the ML Algorithms*

From the detailed analysis, several limitations have been observed while using ML for disease prediction. Some key limitations are listed below [90-92]:

- Logistic Regression (LR) algorithm
  - LR assumes a linear correlation among both the dependent and independent variables.
  - In the case of too many predictors, LR may overfit the training data, and underfit the training data for a smaller number of predictors.

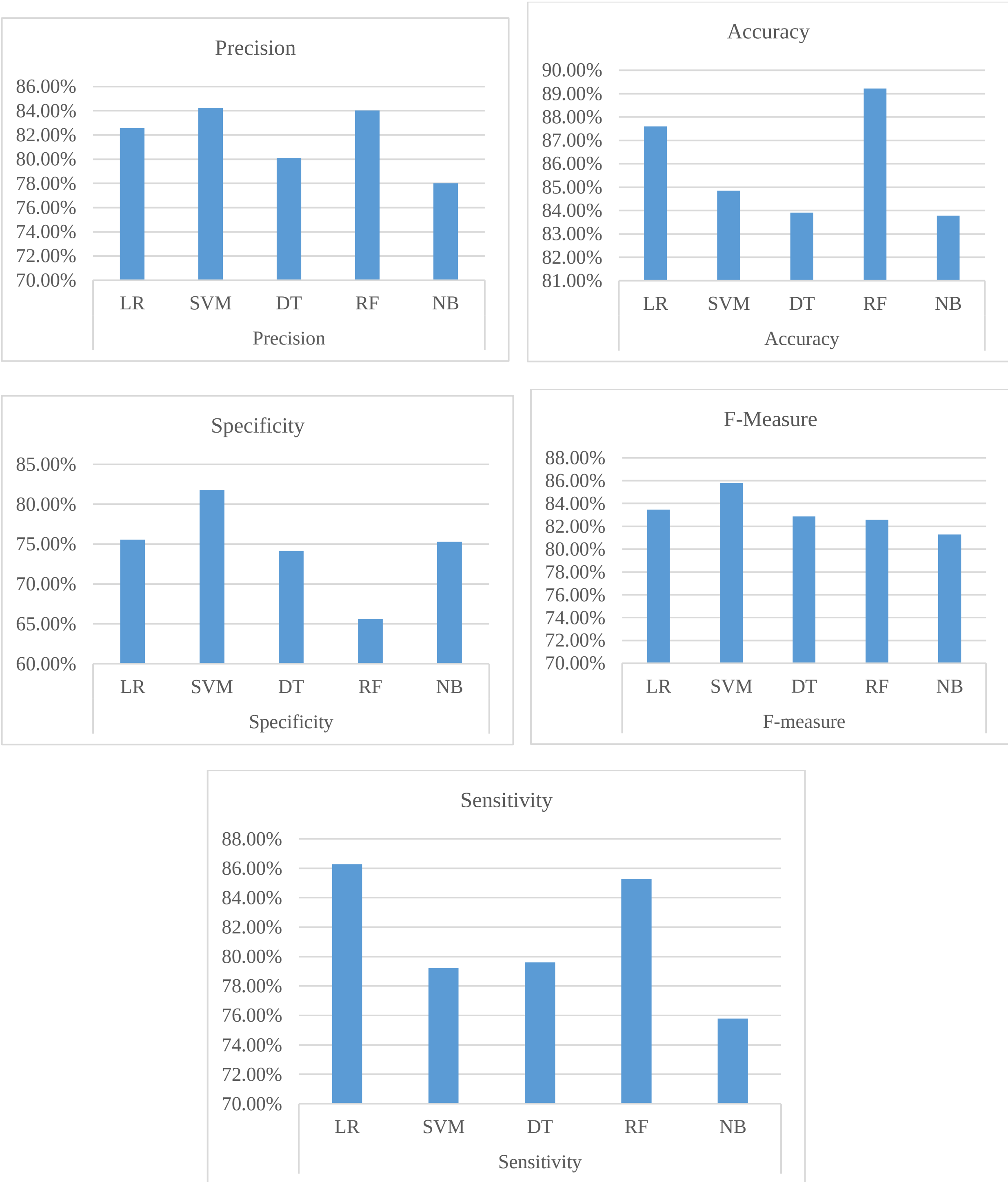

Fig.4. Performance measures vs ml algorithms

- LR is not appropriate for feature selection

- Support Vector Machine (SVM) algorithm.
  - SVM becomes more complex with bulk datasets.
  - More memory resources are needed.
  - Normalization / pre-processing is must, otherwise SVM performs less effectively.

- Decision Tree (DT) algorithm
  - DT easily overfits the training data if there are no constraints.
  - Minor changes in the data can make major changes in the DT structure.

- DT is biased if the features have more levels / categories.
- Computational cost is high for very large datasets.

- Random Forest (RF) algorithm
  - RF is more complex as compared with DT.
  - More computational time since the training is much slower.
  - RF decisions are difficult to interpret due to its complexity.
  - Overfit to the training data if the number of trees is excessively high.
  - Requires more memory due to the large number of trees.
  - RF may not perform well with highly imbalanced datasets.

- Naive Bayes (NB) algorithm
  - If the features are highly correlated, NB performs poorly.
  - NB is highly sensitive which might affect the accuracy
  - Requires more training data

3.4. Recommendations Based on Multi Objective Optimization Technique

In this article, Grey Relational Analysis (GRA) [93] is used to find a better ML technique with a combined objective of maximizing the performance measures. GRA was introduced by a Chinese Professor Deng [94]. It consists of the following steps:

- Generation of data set
- Grey normalization
- Determination of grey relational coefficient
- Computation of grey relational grade (GRG)

The data set (Table 3) is depicted using equation (1) as X_{Data}

$$X_{Data} = \begin{bmatrix} x_{11} & x_{12} & . & . & . & x_{1n} \\ x_{22} & x_{22} & . & . & . & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & . & . & . & x_{mn} \end{bmatrix} \tag{1}$$

where ‘m’ is Alternatives (ML algorithms) & ‘n’ is the performance measures.

The data set is normalized as $X_i\ (j)$ where $(0 \leq x_i\ (j) \leq 1)$ by using the following equation (2) to reduce the variability.

$$X_i^*(j) = \frac{x_i(j) - min\ x_i(j)}{max\ x_i(j) - min\ x_i(j)} \tag{2}$$

where, i= 1, 2, …. m; j=1, 2 … n.

Table 4. Data set & normalized data set

ML Technique	Data Matrix					Normalized matrix				
	P	A	Sp	F	Sn	P	A	Sp	F	Sn
LR	0.8258	0.8426	0.8011	0.8403	0.7802	0.654	0.767	0.556	0.780	0.322
SVM	0.856	0.8485	0.8391	0.8922	0.8378	0.936	0.857	0.910	1.000	1.000
DT	0.7557	0.8182	0.7416	0.6563	0.7529	0.000	0.396	0.000	0.000	0.000
RF	0.8347	0.8579	0.8487	0.8258	0.813	0.737	1.000	1.000	0.719	0.708
NB	0.8629	0.7922	0.7961	0.8529	0.7579	1.000	0.000	0.509	0.833	0.059

The absolute difference ($[\Delta_i(j)]$ is determined by using equation (3).

$$\Delta_i(j) = abs\ (X_o{}^*(j) - X_i{}^*(j)) \tag{3}$$

where $X_o{}^*(j)$ = optimum value of jth performance measure; $X_i{}^*(j)$= normalized cell value of jth performance measure.

The following formula (4) is used to determine the grey relational coefficient from the normalized data.

$$\gamma_i(j) = \frac{\Delta \min + \xi \Delta \max}{\Delta_i(j) + \xi \Delta \max} \tag{4}$$

where $\Delta \min = \min_i \min_j \Delta_i(j)$; $\Delta \max = \max_i \max_j \Delta_i(j)$; and ξ =distinguished coefficient $\xi \varepsilon [0,1]$ (assumed as 0.5)

Finally, using the equation (5), the grey relational grade (GRG) is obtained.

$$\tau_i = \sum_{j=1}^{m} [w_j * \gamma_i(j)] \tag{5}$$

where W_j = weightage of performance measures (Assumes as equal).

The best alternative is the one (ML method) that has the highest GRG. Tables 4 & 5 present the GRG calculations.

Table 5. GRG calculations

	Δ Calculations (Abs)					Coefficient calculations					GRG
ML technique	P	A	Sp	F	Sn	P	A	Sp	F	Sn	P
LR	0.346	0.233	0.444	0.220	0.678	0.591	0.682	0.529	0.694	0.424	0.584
SVM	0.064	0.143	0.090	0.000	0.000	0.886	0.778	0.848	1.000	1.000	0.902
DT	1.000	0.604	1.000	1.000	1.000	0.333	0.453	0.333	0.333	0.333	0.357
RF	0.263	0.000	0.000	0.281	0.292	0.655	1.000	1.000	0.640	0.631	0.785
NB	0.000	1.000	0.491	0.167	0.941	1.000	0.333	0.504	0.750	0.347	0.587

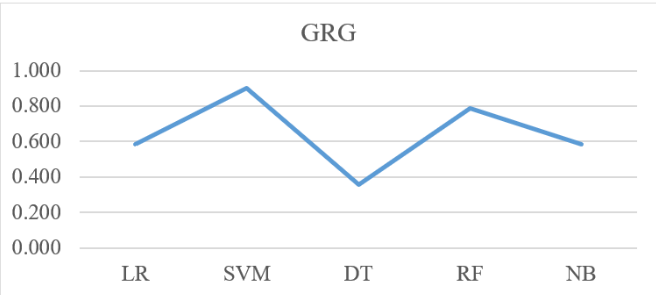

Fig.5. GRG

From Fig. 5, SVM has been recommended for disease prediction followed by RF.

4. Conclusions

The papers that were published in the field of disease prediction using machine learning algorithms were gathered and methodically reviewed for this study. Specifically, five distinct ML algorithms such as logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF), and Naive Bayes (NB) algorithms were considered in this study. The prediction models' performance metrics, including accuracy, sensitivity, F-measure, specificity, and precision, were ascertained by the researchers. Numerous health disorders / diseases, including COVID-19, cancer, heart disease, diabetes, eye disease, Parkinson disease, liver disease, and others, were predicted by the prediction models that were constructed. LR algorithm was used in half of the articles available to predict COVID-19. For cancer prediction, LR and RF were used by 30.77% of the researchers. For various heart diseases, 31.43% literatures used DT for the prediction. RF was utilized by 30.77% of the researchers for diabetes prediction. The detailed analysis was also presented after the complete review. Finally, GRA based multi objective optimization was proposed to choose a better ML technique. Through GRA SVM has been ranked as '1' followed by RF.

References

[1] D. Theng, and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowledge* and *Information Systems,* vol. 66, pp. 1575–1637, 2024, DOI: 10.1007/s10115-023-02010-5.

[2] Abid Haleem, Mohd Javaid, Ravi Pratap Singh, and Rajiv Suman, "Exploring the revolution in healthcare systems through the applications of digital twin technology," *Biomedical Technology,* vol. 4, pp. 28-38, 2023, DOI: 10.1016/j.bmt.2023.02.001.

[3] S. Phani Praveen, T. Bala Murali Krishna, C. H. Anuradha, Srinivasa Rao Mandalapu, Pappula Sarala, and S. Sindhura, "A robust framework for handling health care information based on machine learning and big data engineering techniques," *International Journal of Healthcare Management,* pp. 1-18, 2022, DOI: 10.1080/20479700.2022.2157071.

[4] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny, "A review on deep learning approaches in healthcare systems: taxonomies, challenges, and open issues," *Journal of Biomedical Informatics*, vol. 113, Article no. 103627, 2021, DOI: 10.1016/j.jbi.2020.103627

[5] H. Kaur, and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning Approach," *Applied Computing and Informatics,* vol. 18, no. 1/2, pp. 90-100, 2022, DOI: 10.1016/j.aci.2018.12.004

[6] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications,* vol. 23, no. 7–8, pp. 2031–2038, 2013, DOI: 10.1007/s00521-013-1362-6

[7] M. Nilashi, O. B. Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Computers* & *Chemical Engineering, vol.* 106, pp. 212–223, 2017, DOI: 10.1016/j.compchemeng.2017.06.011

[8] F. Z. Esfahlani, K. Visser, G. P. Strauss, and H. Sayama, "A network-based classification framework for predicting treatment response of schizophrenia patients," *Expert System with Applications*, vol. 109, pp. 152–161, 2018, DOI: 10.1016/j.eswa.2018.05.005

[9] R. H. Ekpo, V. C. Osamor, A. A. Azeta, E. Ikeakanam, and B. O. Amos, "Machine learning classification approach for asthma prediction models in children," *Health and Technology,* vol. 13, pp. 1–10, 2023, DOI: 10.1007/s12553-023-00732-8

[10] T. Turki, and Y. H. Taguchi, "Machine learning algorithms for predicting drugs–tissues relationships," *Expert System with Applications*, vol. 127, pp. 167-186, 2019, DOI: 10.1016/j.eswa.2019.02.013

[11] N. Bhatla, and K. Jyoti, "A novel approach for heart disease diagnosis using data mining and fuzzy logic," *International Journal* of *Computer Applications, vol.* 54, no. 17, pp. 16-21, 2012, DOI: 10.5120/8658-2498

[12] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "*Foundations of Machine Learning*," The MIT Press, London, UK, 2018.

[13] M. M. Eissa, M. Elmogy, and M. Hashem, "Rough – granular computing knowledge discovery models for medical classification," *Egyptian Informatics Journal*, vol. 17, no. 3, pp. 265–272, 2016, DOI: 10.1016/j.eij.2016.01.001

[14] A. Mahabub, "A robust voting approach for diabetes prediction using traditional machine learning techniques," *SN Applied Sciences,* vol. 1667, 2019, DOI: 10.1007/s42452-019-1759-7

[15] J. Jeon, P. J. Leimbigler, G. Baruah, M. H. Li, Y. Fossat, and A. J. Whitehead, "Predicting glycaemia in type 1 diabetes patients: Experiments in feature engineering and data imputation," *Journal* of *Healthcare Informatics Research,* vol. 4, pp. 71–90, 2020, DOI: 10.1007/s41666-019-00063-2

[16] Z. Khandezamin, M. Naderan, and M. Javad Rashti, "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier," *Journal of Biomedical Informatics,* vol. 111, Article no. 103591, 2020, DOI: 10.1016/j.jbi.2020.103591.

[17] C. G. Owen, F. W. Fitzke, and E. G. Woodward, "A new computer assisted objective method for quantifying vascular changes of the bulbar conjunctivae," *Ophthalmic and Physiological Optics,* vol. 16, pp. 430–437, 1996, DOI: 10.1016/0275-5408(96)00037-3

[18] L. Brea, N. Barreira, A. Mosquera, H. Pena-Verdeal, and E. Yebra-Pimentel, "Comparing machine learning techniques in a hyperemia grading framework," *in International Conference on Agents and Artificial Intelligence 2016,* vol. 2, pp. 423-429, 2016, DOI: 10.5220/0005756004230429

[19] V. Nandagopal, S. Geeitha, K. V. Kumar, and J. Anbarasi, "Feasible analysis of gene expression –a computational based classification for breast cancer," *Measurement,* vol. 140, pp. 120–125, 2019, DOI: 10.1016/j.measurement.2019.03.015

[20] C. B. C. Latha, and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked, vol.* 16, Article no. 100203, 2019, DOI: 10.1016/j.imu.2019.100203

[21] S. D. Desai, S. Giraddi, S. P. Narayankar, N. R. Pudakalakatti, and S. Sulegaon, "Back-propagation neural network versus logistic regression in heart disease classification," In: Mandal, J., Bhattacharyya, D., Auluck, N. (eds.), *Advanced Computing and Communication Technologies*. pp 133–144. Springer, Singapore, 2019, DOI: 10.1007/978-981-13-0680-8_13

[22] C. Tania, and O. Raymond, "Logistic regression prediction model for cardiovascular disease," *International Journal* of *New Media Technology, vol.* 7, pp. 33-38, 2020, DOI: 10.31937/ijnmt.v7i1.1340 .

[23] M. E. Hossain, S. Uddin, and A. Khan, "Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes," *Expert System with Applications,* vol. 164, Article no. 113918, 2021, DOI: 10.1016/j.eswa.2020.113918

[24] G. Ambrish, G. Bharathi, G. Anitha, S. Chetana, S., Dhanraj, and M. Kiran, "Logistic regression technique for prediction of cardiovascular disease," *Managing Global Transit,* vol. 3, no. 1, pp. 127-130, 2022, DOI: 10.1016/j.gltp.2022.04.008

[25] A. B. Majumder, S. Gupta, D. Singh, and S. Majumder, " An intelligent system for prediction of COVID-19 case using machine learning framework-logistic regression," *Journal of Physics: Conference Series*, vol. 1797, Article no. 012011, 2021, DOI: 10.1088/1742-6596/1797/1/012011

[26] A. Verghese, T. Sudalaimuthu, and S. Visalaxi, "Analysis and forecasting Covid-19 spread in India using logistic regression and prophet time series," in *International Conference on Computational Performance Evaluation - 2021*, pp. 928-932, 2021, DOI: 10.1109/ComPE53109.2021.9752218.

[27] A. Kulenovic, and A. Lagumdzija-Kulenovic, "Using logistic regression to predict long COVID conditions in chronic patients," *Studies in Health Technology and Informatics,* vol. 29, pp. 265-268, 2022, DOI: 10.3233/SHTI220713

[28] R. Nopour, M. Shanbehzadeh, and H. Kazemi-Arpanahi, "Using logistic regression to develop a diagnostic model for COVID-19: A single-center study," Journal of *Education* and *Health Promotion,* vol.11, Article no. 153, 2022, DOI: 10.4103/jehp.jehp_1017_21.

[29] C. Corinna, and V. Vladimir, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273–297, 1995, DOI: 10.1007/BF00994018 .

[30] Y. J. Son, H. G. Kim, E. H. Kim, S. Choi, and S. K. Lee, "Application of support vector machine for prediction of medication adherence in heart failure patients," *Journal of Healthcare Informatics Research,* vol. 16, no. 4, pp. 253-261, 2010, DOI: 10.4258/hir.2010.16.4.253

[31] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC Medical Informatics and Decision Making,* vol. 22, pp. 10-16, 2010, DOI: 10.1186/1472-6947-10-16

[32] S. Huang, N. Cai, P. P. Pachecom, S. Narrandesm, Y. Wang, and W. Xu, “Applications of support vector machine learning in cancer genomics,” *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41-51, 2018, DOI: 10.21873/cgp.20063

[33] Md. Ehtisham Farooqui, and Jameel Ahmad, “Disease prediction system using support vector machine and multilinear regression,” *International Journal* of *Innovative* Research in *Computer Science* and *Technology, vol.* 8, no. 4, pp. 331–336, 2020, DOI: 10.2139/ssrn.3673232

[34] E. Owusu, P. B. Sekyerehene, J. K. Appati, and J. Y. Ludu, “Computer-aided diagnostics of heart disease risk prediction using boosting support vector machine,” *Computational Intelligence* and *Neuroscience,* vol. 2021, Article no. 3152618, 2021, DOI: 10.1155/2021/3152618

[35] H. Puri, J. Chaudhary, K. R. Raghavendra, R. Mantri, and K. Bingi, “Prediction of heart stroke using support vector machine algorithm,” in *8th International Conference on Smart Computing and Communications - 2021*, pp. 21-26, 2021, DOI: 10.1109/ICSCC51209.2021.9528241.

[36] H. K. Madhu, and D. Ramesh, “Heart attack analysis and prediction using SVM,” *International Journal of Computer Application*, vol. 183, no. 27, pp. 35-39, 2021, DOI: 10.5120/ijca2021921658

[37] S. Guhathakurata, S. Kundu, A. Chakraborty, and J. S. Banerjee, “ A novel approach to predict COVID-19 using support vector machine,” *Data Science for COVID-19,* vol. 2021, pp. 351–364, 2021, DOI: 10.1016/B978-0-12-824536-1.00014-9.

[38] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, “Enhanced heart disease prediction based on machine learning and χ2 statistical optimal feature selection model,” *Designs*, vol. 6, no. 5, Article no. 87, 2022, DOI: 10.3390/designs6050087

[39] W. Ji, T. Yang, Q. Yuan, G. Cheng, and S. Yu, “Prediction model of accident vehicle speed based on artificial intelligence decision tree algorithm,” In: Long, S., Dhillon, B.S. (eds.) Man-Machine-Environment System Engineering. MMESE 2022, *Lecture Notes in Electrical Engineering,* pp 317–324 Springer, Singapore, 2023, DOI: 10.1007/978-981-19-4786-5_44

[40] J. R. Quinlan, “Induction of decision trees,” *Machine Learning,* vol. 1, pp. 81–106, 1986, DOI: 10.1007/BF00116251

[41] S. D. Howlader, T. Biswas, A. Roy, G. Mortuja, and D. Nandi, "A comparative analysis of algorithms for heart disease prediction using data mining", International Journal of Information Technology and Computer Science, vol.15, no.5, pp.45-54, 2023. DOI: 10.5815/ijitcs.2023.05.05

[42] Purushottam, K. Saxena. and R. Sharma, “Efficient heart disease prediction system using decision tree,” in *International Conference on Computing, Communication & Automation - 2015*, pp. 72-77, 2015, DOI: 10.1109/CCAA.2015.7148346

[43] B. Bahrami, and M. H. Shirvani, “Prediction and diagnosis of heart disease by data mining techniques,” *Journal of Multidisciplinary Engineering and Science*, vol. 2, no. 2, pp. 164-168, 2015.

[44] M. Subburaj, and P. Ramu, “Heart disease prediction system using decision tree and Naive Bayes algorithm,” *Current Medical Imaging* Reviews, vol. 15, no. 8, pp. 712 – 717, 2019, DOI: 10.2174/1573405614666180322141259

[45] A. Ahmad, O. Safi, S. Malebary, S. Alesawi, and E. Alkayal, “Decision tree ensembles to predict Coronavirus disease 2019 infection: A comparative study,” *Complexity*, vol. 2021, Article no. 5550344, 2021, DOI: 10.1155/2021/5550344

[46] N. Kapoor, and P. Singh, “Heart disease prediction using decision tree and random forest classification techniques,” in Goundar, S., Rayani, P. (eds.) *Applications of Big Data in Large- and Small-Scale Systems*, pp. 234-259. IGI Global, 2021. DOI: 10.4018/978-1-7998-6673-2.ch015

[47] V. S. K. Reddy, P. V. Meghana, N. V. S. Reddy, and B. A. Rao, “Prediction on cardiovascular disease using decision tree and Naïve Bayes classifiers,” *Journal of Physics: Conference Series,* vol. 2161, Article no. 012015, 2022, DOI: 10.1088/1742-6596/2161/1/012015

[48] C. Sateesh, and R. Balamanigandan, “Heart disease prediction using innovative decision tree technique for increasing the accuracy compared with convolutional neural networks,” in *2nd International Conference on Innovative Practices in Technology and Management - 2022*, pp. 583-587, 2022, DOI: 10.1109/ICIPTM54933.2022.9754196.

[49] R. Vijaya Saraswathi, K. Gajavelly, A. Kousar Nikath, R. Vasavi, and R. Reddy Anumasula, “Heart disease prediction using decision tree and SVM,”in A. B. Reddy, B. Kiranmayee, R. R. Mukkamala, and K. Srujan Raju, (eds.), *Algorithms for Intelligent Systems*, Springer, Singapore. pp 69–78, 2022, DOI: 10.1007/978-981-16-7389-4_7

[50] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: from early developments to recent advancements,” *Systems* Science & *Control Engineering, vol.* 2, no. 1, pp. 602-609, 2014, DOI: 10.1080/21642583.2014.956265

[51] G. Szűcs, “Random response forest for privacy-preserving classification,” *Journal of Computer Engineering*, vol. 2013, pp. 1–6, 2013, DOI: 10.1155/2013/397096

[52] T. Ooka, H. Johno, K. Nakamoto, Y. Yoshioki, H. Yokomichi, and Z. Yamagata, “Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: Large-scale health check-up data in Japan,” *BMJ Nutrition, Prevention* & *Health*, vol. 4, Article no. e000200, pp. 140-148, 2021, DOI: 10.1136/bmjnph-2020-000200

[53] S. Sharma, A. Aggarwal, and T. Choudhury, “Breast cancer detection using machine learning algorithms,” in *International Conference on Computational Techniques, Electronics and Mechanical Systems - 2018*, pp. 114-118, 2018, DOI: 10.1109/CTEMS.2018.8769187

[54] S. Bharati, M. A. Rahman, and P. Podder, “Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA,” in *4th International Conference on Electrical Engineering, Information and Communication Technology 2018*, pp. 581-584, 2018, DOI: 10.1109/CEEICT.2018.8628084.

[55] C. C. Wu, W. C. Yeh, W. D. Hsu, M. M. Islam, P. A. A. Nguyen, T. N. Poly, Y. C. Wang, H. C. Yang, and Y. C. Jack Li, “Prediction of fatty liver disease using machine learning algorithms,” *Computer Methods* and *Programs in Biomedicine,* vol. 170, pp. 23-29, 2019, DOI: 10.1016/j.cmpb.2018.12.032.

[56] E. D. H. Gates, J. S. Lin, J.S. Weinberg, S. S. Prabhu, J. Hamilton, J. D. Hazle, G. N. Fuller, V. Baladandayuthapani, D. T. Fuentes, and D. Schellingerhout, “Imaging-based algorithm for the local grading of glioma,” *American Journal* of *Neuroradiology,* vol. 41, no. 3, pp. 400-407, 2020, DOI: 10.3174/ajnr.A6405.

[57] Z. Ma, J. Ma, Y. Miao, and X. Liu, “Privacy-preserving and high- accurate outsourced disease predictor on random forest,” *Information Science,* vol. 496, pp. 225–241, 2019, DOI: 10.1016/j.ins.2019.05.025

[58] S. Mohan, T. Chandrasegar, and S. Gautam, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE Access*, vol. 7, pp. 81542-81554, 2019, DOI: 10.1109/ACCESS.2019.2923707.

[59] A. K. G. Escamilla, A. H. E. L. Hassani, and E. Andres, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked, vol.* 19, Article no. 100330, 2020, DOI: 10.1016/j.imu.2020.100330

[60] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques", *SN Computer Science*, vol. 1, Article no. 345, 2020, DOI: 10.1007/s42979-020-00365-y

[61] M. Pal, and S. Parija, "Prediction of heart diseases using random forest," *Journal of Physics: Conference Series*, vol. 1817, Article no. 012009, 2021, DOI: 10.1088/1742-6596/1817/1/012009

[62] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *Journal of Supercomputing*, vol. 77, pp. 5198–5219, 2021, DOI: 10.1007/s11227-020-03481-x

[63] R. Shanthakumari, C. Nalini, S. Vinothkumar, E. M. Roopadevi, and B. Govindaraj, "Multi disease prediction system using random forest algorithm in healthcare system," in *International Mobile and Embedded Technology Conference - 2022*, pp. 242-247, 2022, DOI: 10.1109/MECON53876.2022.9752432.

[64] W. Wei, S. Visweswaran, and G. F. Cooper, "The application of Naive Bayes model averaging to predict Alzheimer's disease from genome-wide data," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 370– 375, 2011, DOI: 10.1136/amiajnl-2011-000101

[65] K. J. Wang, B. Makond, and K. M. Wang, "Modeling and predicting the occurrence of brain metastasis from lung cancer by Bayesian network: a case study of Taiwan," Computers in Biology and Medicine, vol. 47, pp. 147–160, 2014, DOI: 10.1016/j.compbiomed.2014.02.002.

[66] E. Miranda, E. Irwansyah, A. Y. Amelga, M. M. Maribondang, and M. Salim, "Detection of cardiovascular disease risk's level for adults using Naive Bayes classifier," *Healthcare Informatics Research*, vol. 22, no. 3, pp. 196-205, 2016, DOI: 10.4258/hir.2016.22.3.196

[67] J. Singh, A. Kamra, and H. Singh, "Prediction of heart diseases using associative classification," in *5th International Conference on Wireless Networks and Embedded Systems - 2016*, pp. 1-7, 2016, DOI: 10.1109/WECON.2016.7993480.

[68] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Natural Methods*, vol. 15, pp. 233–234, 2018, DOI: 10.1038/nmeth.4642

[69] U. N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization," *Biomedical Research*, vol. 29, no. 12, pp. 2646-2649, 2018, DOI: 10.4066/biomedicalresearch.29-18-620

[70] S. Anitha, and N. Sridevi, "Heart disease prediction using data mining techniques," *Journal of Applied Analysis and Computation*, vol. XIII, no. II, pp. 48-55, 2019.

[71] A. Hemanth Kumar, and G. Pavani, "Naive Bayes classifier for predicting Corona virus," *Journal of Engineering Science*, vol. 13, no. 09, pp. 144 – 155, 2022, DOI: 10.1109/ICICV50876.2021.9388410.

[72] R. Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert System with Applications*, vol. 37, no. 2, pp. 1568-1572, 2010, DOI: 10.1016/j.eswa.2009.06.040

[73] S. Wu, and J. Guo, "A data mining analysis of the Parkinson's disease," *iBusiness*, vol. 3, pp. 71-75, 2011, DOI: 10.4236/ib.2011.31012

[74] P. Naraei, A. Abhari, and A. Sadeghian, "Application of multilayer perception neural networks and support vector machines in classification of healthcare data," in *Future Technologies Conference – 2016*. pp. 848-852, 2016, DOI: 10.1109/FTC.2016.7821702.

[75] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clinical Epidemiology* and *Global Health*, vol. 7, no. 3, pp. 293–299, 2019, DOI: 10.1016/j.cegh.2018.10.003

[76] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal* of *Clinical Epidemiology*, vol. 110, pp. 12-22, 2019, DOI: 0.1016/j.jclinepi.2019.02.004

[77] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1-18, 2020, DOI: 10.1186/s12859-020-03626-y

[78] Y. L. Chiu, M. J. Jhou, T. S. Lee, C. J. Lu, and M. S. Chen, "Health data-driven machine learning algorithms applied to risk indicators assessment for chronic kidney disease," *Risk Management and Healthcare Policy, vol.* 2021, no. 14, pp. 4401-4412, 2021, DOI: 10.2147/RMHP.S319405

[79] N. S. Hassan, A. Mohsin Abdulazeez, J. N. Saeed, D. Qader Zeebaree, A. Al-Zebari, and F. Y. H. Ahmed, "A compassion of three data mining algorithms for heart disease prediction," in *IEEE Symposium on Industrial Electronics & Applications 2021*, pp. 1-6, 2021, DOI: 10.1109/ISIEA51897.2021.9509985

[80] J. D. Acheme, and O. R. Vincent, "Machine-learning models for predicting survivability in COVID-19 patients," *Data Science for COVID-19*, vol. 2021, pp. 317-336, 2021, DOI: 10.1016/B978-0-12-824536-1.00011-3

[81] M. Shanbehzadeh, R. Nopour, and U. Kazemi-Arpanahi, "Using decision tree algorithms for estimating ICU admission of COVID-19 patients," *Informatics in Medicine Unlocked, vol.* 30, Article no. 100919, 2022, DOI: 10.1016/j.imu.2022.100919

[82] Cevik, S. Angco, E. Heydarigharaei, H. Jahanshahi and N. Prayogo, "N. Active learning for multi-way sensitivity analysis with application to disease screening modeling', *Journal of Healthcare Informatics*, vol. 6, pp. 317–343, 2022, DOI: 10.1007/s41666-022-00117-y

[83] R. Priyanka, and L. Shahram, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods* and *Programs in Biomedicine*, vol. 1, Article no. 100032, 2021, DOI: 10.1016/j.cmpbup.2021.100032

[84] S. Deepti, and S. Dilip Singh, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018, DOI: 10.1016/j.procs.2018.05.122

[85] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, Article no. 515, 2018, DOI: 10.3389/fgene.2018.00515

[86] S. Jayakumar, V. Muthukumaran, J. Thimmia Raja, R. B. Joseph, M. Munirathanam, J. B. Balajee, "Diabetes disease prediction using decision tree for feature selection," *Journal of Physics: Conference Series*, vol. 1964, Article no. 062116, 2021, DOI: 10.1088/1742-6596/1964/6/062116

[87] N. Nai-arun and R. Moungmai, "Diagnostic Prediction Models for Cardiovascular Disease Risk using Data Mining Techniques", *ECTI-CIT Transactions*, vol. 14, no. 2, pp. 113–121, 2020, DOI: 10.37936/ecti-cit.2020142.199897M.

[88] T. M. Alam, M. F. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar, and Z. Abbas, “A model for early prediction of diabetes,” *Informatics in Medicine Unlocked, vol.* 16, Article no. 100204, 2019, DOI: 10.1016/j.imu.2019.100204.

[89] K. Vijiya Kumar, B. Lavanya, I Nirmala, and S. S. Caroline, “Random forest algorithm for the prediction of diabetes,” in *IEEE International Conference on System, Computation, Automation and Networking - 2019*, pp. 1-5, 2019, DOI: 10.1109/ICSCAN.2019.8878802.

[90] G. Shanmugasundar, M. Vanitha, R. Cep, V. Kumar, K. Kalita, and M. A. Ramachandran, “Comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining,” *Processes*, vol. 9, Article no. 2015, 2021, DOI: 10.3390/pr9112015S.

[91] B. Kotsiantis, “Supervised machine learning: a review of classification techniques,” *Informatica*, vol. 31, pp. 249-268, 2007, DOI: 10.1115/1.1559160.

[92] L. Smith, and C. Lamprecht, "Identifying the limitations associated with machine learning techniques in performing accounting tasks," *Journal of Financial Reporting and Accounting*, vol. 22, no. 2, pp. 227-253, 2024, DOI: 10.1108/JFRA-05-2023-0280

[93] P. Pitchipoo, P. Venkumar, and S. Rajakarunakaran, “Grey decision model for supplier evaluation and selection in process industry: a comparative perspective,” *The International Journal of Advanced Manufacturing Technology*, vol.76, pp. 2059–2069, 2015, DOI: 10.1007/s00170-014-6406-2

[94] J. L. Deng, “Introduction to grey system,” Journal of Grey Systems, vol. 1, no. 1, pp. 1-24, 1989

Authors’ Profiles

Shunmuga Priya Subramanian received her bachelor's degree in Computer Science from Madurai Kamaraj University, Madurai, India, in 2001, and a Master’s in Computer Applications from Bharathiar University, Coimbatore, India, in 2004. She also holds a Master’s degree in Computer Science and Engineering from Anna University, Chennai, India, which she completed in 2010. She is currently pursuing a Ph.D. in Computer Applications at Kalasalingam Academy of Research and Education, Krishnankoil, India. Her research interests include machine learning and image processing. She is presently working as an Assistant Professor in the Department of Commerce with Computer Applications at Rajapalayam Rajus College, Rajapalayam, Tamilnadu, India.

Dr. Amuthaguka Duraipandian holds a Ph.D. in Computer Science and has extensive research experience in Network Security and Soft Computing techniques. She has 20 years of teaching experience and 10 years of research experience. She has 15 publications to her credit in reputed international journals and IEEE conferences. Her research interests include Predictive Analysis in Healthcare, Network Security, Data Analytics, and Cybercrime. She is currently serving as an Associate Professor in the Department of Computer Applications at Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil-626126, Tamil Nadu, India.

How to cite this paper: Shunmuga Priya Subramanian, Amuthaguka Duraipandian, "Ranking of Machine Learning Algorithms Used in Disease Prediction: A Review-based Approach", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.6, pp.74-88, 2024. DOI:10.5815/ijitcs.2024.06.06