Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Disease Diagnosis and Prediction Framework using an Ensemble Classifier

Vimala Devi P[1] [0000-0001-8422-6266] , Kalaichelvi V[2*] [0000-0003-1374-2065]

[1,2]Department of Computer Science and Engineering, Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam, Tamil Nadu, India

Abstract. The significance of digitization in healthcare has increased due to the rapid advancement of technology. Individuals with chronic illnesses need to be monitored closely and treated right away. The possibility of major consequences is decreased by early identification of these disorders. Algorithms for machine learning are frequently employed to forecast this kind of illness. In order to forecast diabetics, we provide a novel ensemble-based machine learning technique in this work. A range of machine learning methods, such as Naïve Bayes, Support Vector Machine, Decision Tree, k-NN, and Random Forest, were employed to analyse 16769 records in total. With a 97% accuracy rate, the best ensemble based SNDT classifier is selected through voting method. Recall, accuracy, and F1-score are among the other performance metrics that are assessed.

Keywords: Healthcare, Machine learning, SVM, Decision Tree, Naïve Bayes.

1 INTRODUCTION

Recent advancements in healthcare include the digitization of medical records, health status monitoring, analysis, and prediction. The Tremendous growth of data in the medical field requires special attention to process and formulates machine learning algorithms that help medical experts make decisions. Early detection of chronic diseases such as diabetes, heart failure, etc., helps to take proper treatment and reduces the death rate. To achieve this, machine learning techniques are used. Machine Learning is a part of Artificial Intelligence, which processes huge volume of data with accuracy and minimum time, especially in healthcare. Abdullah Alanazi discussed the overview of various machine learning algorithms, applications, and challenges [1]. The machine learning models are classified into reinforcement learning, unsupervised learning, and supervised learning [2, 3]. Mohammad Shehab et al. explained various machine learning classifiers used for medical diagnosis [4]. The medical applications are classified into five categories, medical chemistry, brain, cancer, wearable sensors, and medical imaging. The detailed study of different types of cancer and the classifiers used to predict the cancer are also analysed.

Diabetics is the type chronic illness that can be identified using variety of machine learning approaches [5]. They employed Random Forest, AdaBoost, Neural Network, Support Vector Machine, Naive Bayes, and Decision Tree classifiers. The results of machine learning algorithms and neural network classifiers were separately analysed and compared. Rahat Ullah et al. applied machine learning methods along with artificial neural networks to predict the risk of asthma patients [6]. The clinical characteristics are extracted from patients' blood samples using Raman spectra and the blood samples from healthy patients were given as input to the classifiers. The outcome of these classifiers proves that SVM is the suitable classifier. A model to forecast cardiac illness utilising cloud and machine learning techniques was presented by Forum Desai et al. [7]. They checked the patients' status using machine learning and cloud computing algorithms. This model was validated using the measures such as recall, accuracy, precision, AUC curve, etc. the results show that the Logistic Regression is the best model with an accuracy of 85.96.Ekta Maini et al. developed a system using Python programming for the early diagnosis of heart disease [8]. Five algorithms were applied for the prediction model and finally, Random forest method was determined to be the most accurate model and deployed in cloud.

2 LITERATURE SURVEY

Ensemble algorithms are collections of several machine learning algorithms, which are used to produce better performance [9]. Rakesh Chandra Joshi et al. used a two-stage ensemble algorithm to detect glioma [10]. The data were divided into glicoma cases and non-glicoma cases in the first stage and grades were calculated in the second stage. This method was evaluated using measures such as accuracy, recall, precision, etc. From the results, it was

observed that a hard voting-based classifier achieves higher accuracy when compared to other classifiers. Chronic diseases are a major cause of death in today's world. Conventional methods have some drawbacks to predict the disease. To overcome this machine learning algorithms are used. Vardhan Shorewala described a method to detect coronary heart disease in an earlier stage [11]. He compared the base classifiers against ensemble techniques such as stacking, bagging, and boosting. The results show that the stacking methods are effective methods when compared to others. Diabetics is a chronic disease and is categorized into Type I, Type II, and gestational diabetics. Type II diabetes commonly affects all age people. Shahid Mohammad Ganie and Majid Bashir Malik used ensemble classifiers to predict type II diabetics [12]. The bagging classifier achieves the maximum accuracy, according to the data. Tao Zheng et al. developed a framework to identify type 2 diabetics [13]. The machine learning models are adopted in this framework and analysed using various metrics known as accuracy, recall, precision, sensitivity and specificity. Support vector machines, random forests, and logistic regression exhibit strong prediction performances with a 95% accuracy rate, according to the findings.

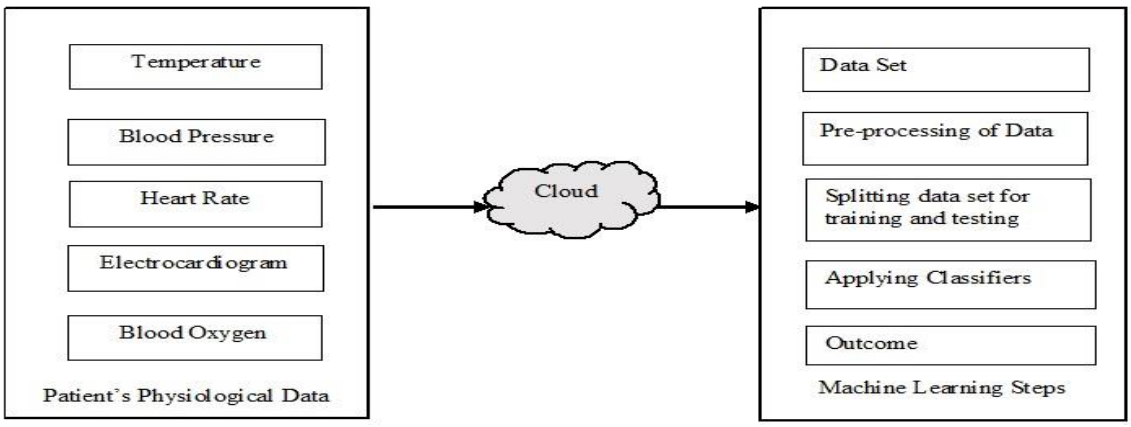
Advancements in technology increase the use of smart devices, connected by the Internet of Things (IoT). It consists of wearable sensors to collect data, transmit the data using wireless standards and store the data in the cloud. Continuous monitoring of patients is possible through IoT devices and generates a huge volume of data. Mohammad Saeid Mahdavinejad et al. discussed the machine learning algorithms used to process these data [14]. They also discussed the characteristics of data and applications of IoT. Priyan Malarvizhi Kumar and Usha Devi Gandhi designed a three-tier architecture to process data coming from IoT devices [15]. Tier-1 was responsible for data collection, Apache HBase was used to store data in Tier-2 and a prediction model was developed using logistic regression in Tier-3.Kashif Naseer Qureshi et al. developed an M-health model to handle emergencies [16]. There are four stages in this model, data collection, analysis, storage, and prediction using machine learning algorithms. The proposed model is compared against the existing algorithms in terms of Recall, specificity, and accuracy. Heart disease and stroke are the critical diseases and main causes of death [17]. Early detection of these diseases increases the lifetime of the patients. Zafer Al-Makhadmeh and Amr Tolba utilized IoT devices to monitor patients continuously [18]. The higher-order Boltzmann deep belief neural network is used to process the gathered data. This model has the highest accuracy and least amount of time to identify the patient. The machine-learning techniques used to detect heart failure are reviewed [19, 20]. This survey focused on symptoms, prediction models, the severity of heart failure, mortality, etc.

3 PROPOSED METHODOLOGY

3.1 Architecture Diagram

The proposed model consists of three phases, data collection, storage and analysis using machine learning algorithms. The physiological data known as blood pressure, temperature, blood glucose, heart rate etc are collected from patient's body and transmitted to cloud. The machine learning techniques are implemented to construct predictive model. Fig 1. describes the architecture of the proposed model.

Fig. 1. Architecture Diagram



3.2 Machine Learning Techniques

3.2.1 Support Vector Machine

SVM is a supervised machine learning technique that works well with linear data and is capable of both regression and classification. To determine the hyperplane dividing the two classes, a statistical method is employed. In order to make it simple to assign new points to the appropriate category, this method establishes the optimum decision boundary that can divide the n-dimensional space into classes. The operation of SVM is illustrated in the following diagram.

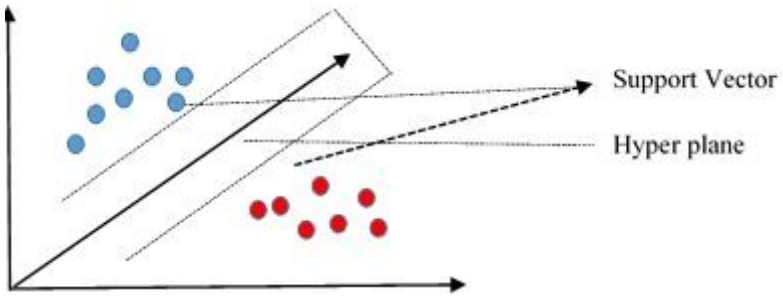


Fig. 2. Working of SVM

The data points that are closest to the hyperplane are the support vectors. The optimal hyper plane is chosen from among a variety of options using either the maximum distance or the maximum margin.

3.2.2 Random Forest

Random Forest is a supervised machine learning technique that is primarily used for classification but also supports regression.

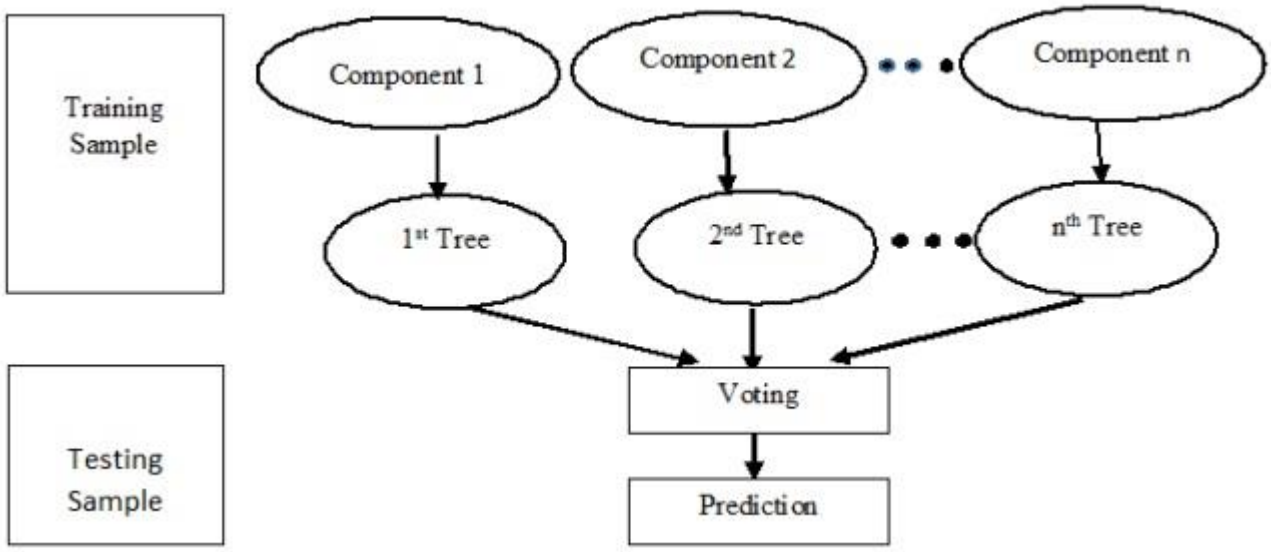


Fig. 3. Block Diagram of Random Forest

K-NN

The K-nearest neighbour method is a straightforward supervised machine learning technique that utilises neighbour distances as its basis. It calculates similarity between the new data and the training data by storing the training data. The similarity is determined using the formula for Euclidean distance.

$$\text{Euclidean distance}=\sqrt{(xn-x0)^2+(yn-y0)^2}$$

(1)

Naïve Bayes

The quick and supervised machine learning algorithm, Bayes, makes predictions about the result using conditional probability. It utilises the Bayes Theorem and is also known as a probabilistic classifier. With prior knowledge, the Bayes Theorem is utilised to calculate a hypothesis's probability.

Decision Tree

A hierarchical supervised machine learning technique with a tree topology, decision trees are employed in both regression and classification models. There are four nodes in its organisation: the root, decision, branch, and end nodes

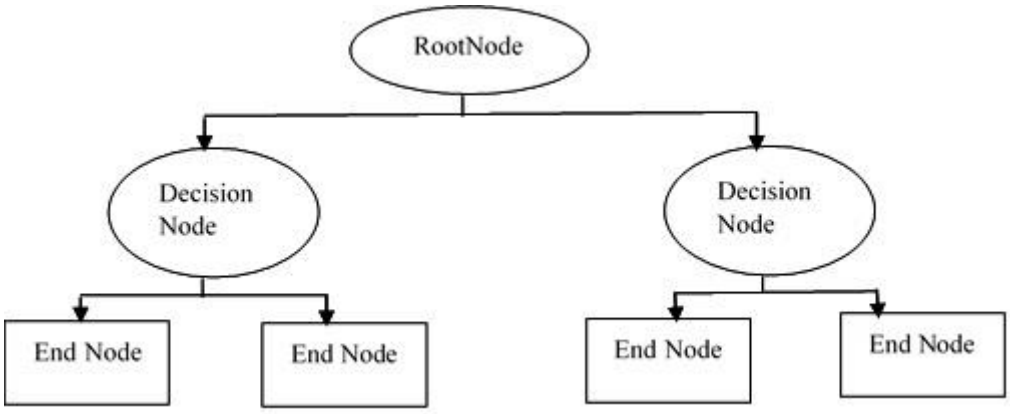


Fig. 4. Principle of Decision Tree

Decision nodes are used to perform evaluations based on the features and the terminal nodes represent the possible outcomes. Gini impurity and information gain are methods to select best feature at each dataset. The formula is,

$$EntropyA=-c\in CPc\log2(c)$$

(2)

$$Information\ gain = Entropy(A) - \sum \frac{|A_v|}{|A|}\ Entropy(A_v)$$

(3)

$$Gini\ Impurity = 1 - \sum P_i^2$$

(4)

Where

- S –calculated entropy of the dataset S
- c - classes in S
- P(c) - proportion of class c data points to all data points in the set, S
- Pi - percentage of the set's elements that fall into the ith category.
- *Entropy(A)* - entropy of dataset A
- $|A_v|/|A|$ -ratio of the values in A_v to the number of values in dataset A
- *Entropy(A_v)* entropy of dataset, X_v

3.3 Study Population Description

The data set was gathered from IEEE data port. The dataset consists of 16969 records and 10 features, among this nine features are input variables and these are independent with each other. The aim of the dataset is to identify the effect of blood glucose level on the human body along with other superficial body features. The table 1 describes the characteristics of the study population:

Feature	Description	Type of data	Range of Values
Age	Person's Age	Numerical value	Between 9 to 77
Blood Glucose Level(BGL)	Parameter used to measure the glucose level in blood.	Numerical value	100-125 mg/dl indicates prediabetics and >=126 mg/dl indicates diabetics
Diastolic Blood Pressure	Arteries pressure level when the heart rests between beats	Numerical value	Normal - <80 mm/hg Prehypertension -80-89 mm/hg High blood pressure ->=90mm/hg
Systolic Blood Pressure	Arteries pressure level when the heart beats	Numerical value	Normal -<90 mm/hg Prehypertension - 120-139 mm/hg High blood pressure >=140 mm/hg
Heart Rate	Number of heart beats per second.	Numerical value	60-100 beats per minute-Normal
Body Temperature	Body temperature in celcius or Fahrenheit	Numerical value	Normal-37 C or 97 F to 99 F
SPO2	Oxygen Saturation	Numerical value	-
Sweating (Y/N)	-	Numerical value	-
Shivering (Y/N)	-	Numerical value	-
Diabetic/Non-Diabetic	Presence or absence of diabetics		0-No diabetics 1-Diabetics

Table 1. Study Population Description

The machine learning algorithms are used to predict the patient's status based on accuracy. The steps involved in the machine learning techniques are preprocessing, feature selection and applying the model. The functions of Data preprocessing are finding missing values, outliers and removal, normalization and feature selection. Table 2 demonstrates the statistics summary of study population: This summary represents total number of records, values of mean, standard deviation, minimum and maximum. The observation from this study is that there is no outliers and missing values in the data.

	Age	Blood Glu-cose Level	Dias-tolic Blood Pres-sure	Systolic Blood Pressure	Heart Rate	Body Tem-pera-ture	SpO2	Sweat-ing(Y/N)	Shiver-ing(Y/N)
Count	16769	16769	16769	16769	16769	16769	16769	16769	16769
Mean	30.98	95.71	77.17	118.18	91.52	97.35	97.38	0.12	0.145
Standard Deviation	25.58	42.99	7.24	7.7	10.4	0.81	0.84	0.32	0.35
Minimum	9	50	60	95	78	96	93	0	0
25%	9	68	71	113	84	96.6	97	0	0
50%	14	83	76	119	89	97.3	98	0	0
75%	55	108	83	124	95	97.9	98	0	0
Maximum	77	250	98	145	130	98.9	99	1	1

Table 2. Statistics of Study Population

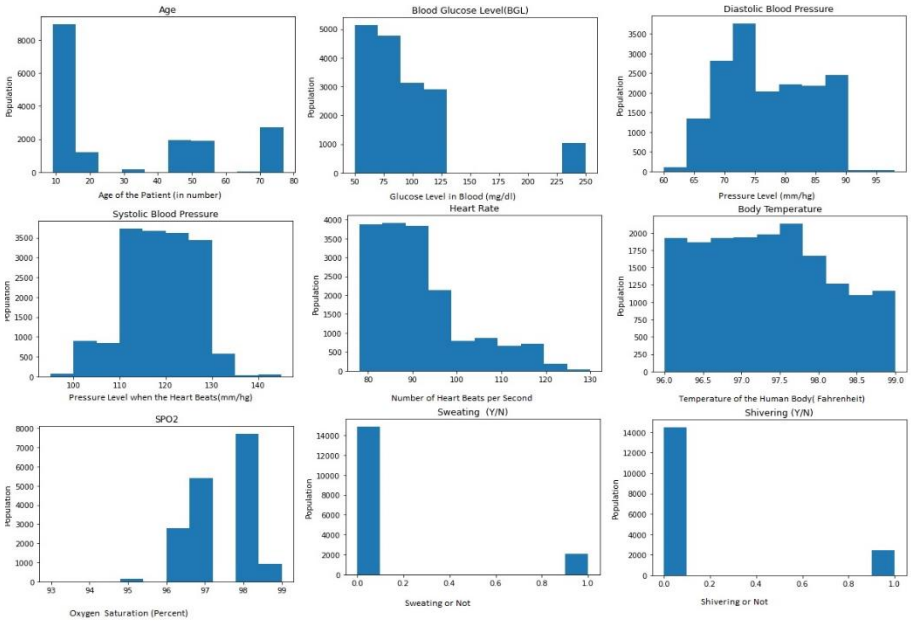


Fig. 5. Distribution of Study Population

The fig. 5 explains the distribution of the data with respect to the outcome. The number of patients, in terms of population is mentioned in y-axis and the corresponding features are defined in x-axis.

3.4 Correlation Matrix

The strength of linear relationship between the two variables is identified by Correlation matrix. The correlation matrix displayed in fig 6. is created using Pearson's correlation coefficient. The values range from -1 to 1, where 0 denotes no correlation, -1 indicates negative correlation, and 1 indicates positive correlation. As seen in fig. 6, it seems that there is high correlation between the features Sweating(Y/N), and Shivering(Y/N) i.e. 0.83, so that the column is dropped for further analysis. The other features are not having the strong correlation, hence all of these features are considered indivually for analysis. The matrix defines the negative correlation also.

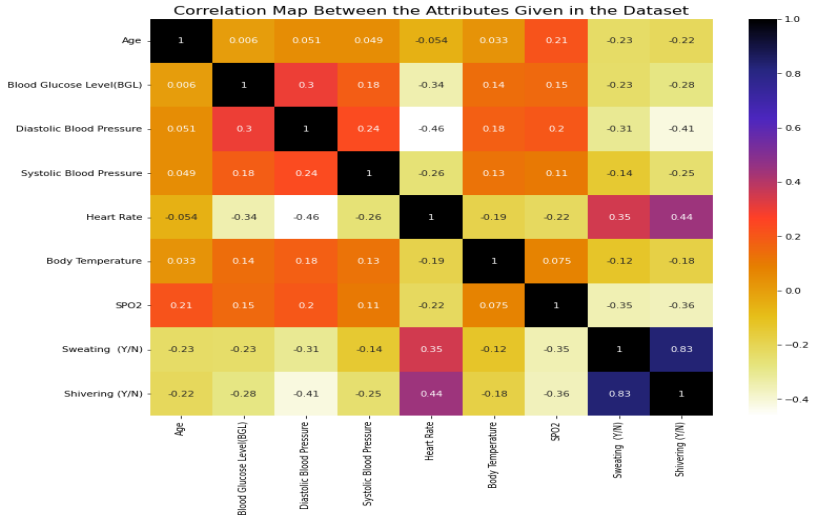


Fig. 6. Correlation Map

The dataset is divided in an 80:20 ratio into train and test models. The test data defines 3394 rows and eight columns, while the training set has 13575 rows and eight columns. Table 3. displays the example test and training sets.

Index	Age	Blood Glucose Level	Dias-tolic Blood Pressure	Systolic Blood Pres-sure	Heart Rate	Body Tem-perature	SpO2	Sweat-ing(Y/N)
9616	9	102	70	117	91	96.9	97	0
3473	9	68	75	106	104	96.39	97	1
12	9	77	75	115	79	98.37	95	0
…			….				…	….
1281	55	61	68	120	101	97.35	96	0
6322	55	80	88	119	91	97.14	98	0
Index	Age	Blood Glucose Level	Dias-tolic Blood Pressure	Systolic Blood Pressure	Heart Rate	Body Tempera-ture	SpO2	Sweat-ing(Y/N)
11975	55	116	78	111	93	98.32	96	0
1542	55	62	69	102	109	96.5	78	1
2345	46	65	65	107	85	96.01	95	1
			….				…	….
7886	76	90	76	125	83	96.95709661	98	0
13298	9	127	76	115	83	96.80787239	98	0

Table 3. Test and Training Set

4 Results and Discussion

4.1 Ensemble Classifier

A collection of different machine learning classifiers is called an ensemble classifier. These classifiers are used to efficiently tackle real-world situations. The machine techniques used in this suggested work are support vector machines, random forests, logistic regression, decision trees, and K-nearest neighbors. The accuracy table lists these algorithms, which are coupled to create several ensemble classifiers. After analysing these algorithms, the optimal one is selected using performance metrics. To carry out the suggested task, the following configurations are used: Windows 8.1, a Jupyter notebook, and an Intel Core i5 processor with 4 GB RAM.

4.2 Measures of Performance Assessment

The performance of the prescribed work is evaluated using the metrics listed below.

Accuracy defines the total number of correct predictions.

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{Total\ Samples\ used\ for\ Testing} \quad (5)$$

Where True positive= Number of patients predicted correctly to have diabetics

True Negative=Number of persons predicted correctly as healthy persons.

False Positive= Number of healthy persons predicted incorrectly to have diabetics

False Negative= Number of patients predicted incorrectly as healthy

Total Samples include the sum of true positives, false positives, false negatives and true negatives.

Sensitivity-It is defined as the ratio of predicted values with diabetics to the actual number of patients affected by diabetics.

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ Negative} \quad (6)$$

$$Prescision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (7)$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{(Prescision + Recall)} \quad (8)$$

Fig 7 shows the Accuracy chart for various classifiers. The graph demonstrates, that the suggested ensemble-based SNDT classifier is better than other classifiers in terms of accuracy.

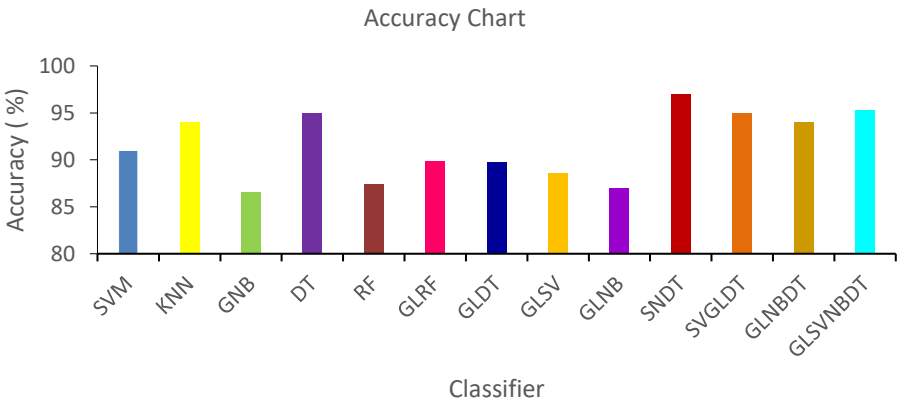


Fig. 7. Accuracy Chart

The table 4. also demonstrates the performance of SNDT classifier.

Algorithm	Accuracy	Precision	Recall	F1-score
SVM	91	94	96	95
KNN	94	98	96	97
GNB	86.6	98	87	92
DT	95	96	96	97
RF	87.4	87	97	93
GLRF	89.8	90	97	93
GLDT	89.7	90	97	94
GLSV	88.6	90	96	94
GLNB	87	90	97	93
SNDT	97	98	98	98
SVGLDT	95	96	96	97
GLNBDT	94	97	96	97
GLSVNBDT	95.3	96	96	97

Table 4. Accuracy Table

The analysis of Mean absolute error (MAE) and Root Mean Square (RMSE) metrics are shown in fig 7. According to the diagram the proposed algorithm has the lower RMSE value than the classifiers. It indicates the superior performance of the proposed classifier.

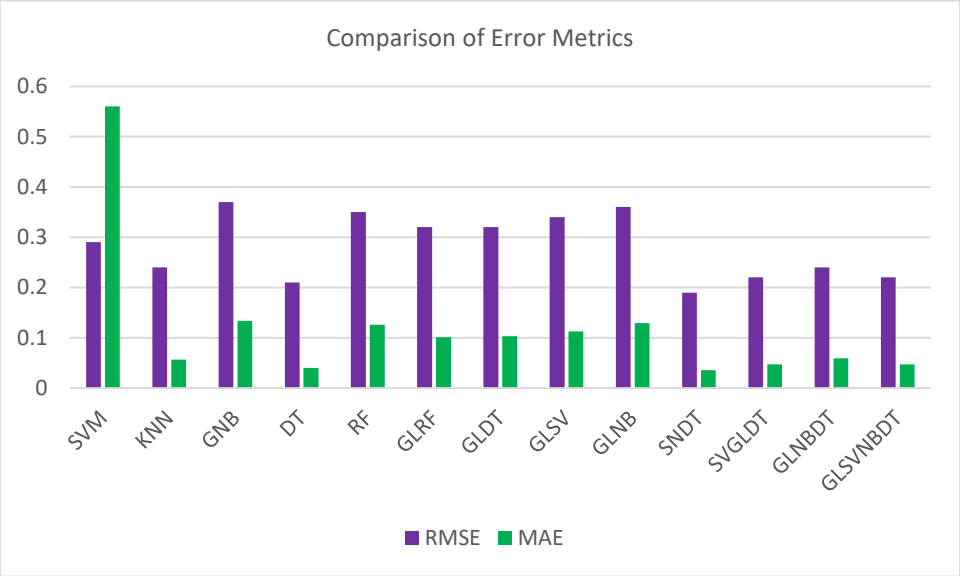


Fig. 8. Comparison of Error Metrics

5 CONCLUSION

The proposed work, uses machine learning algorithms to create a new healthcare application. It was evaluated and compared to determine how accurately the different ensemble classifiers predicted the existence of diabetes. When compared to other classifiers, the suggested model accurately predicts the disease with a 97% accuracy rate. To

strengthen the suggested model, other performance metrics including recall, precision, and F1-score are also assessed. This model demonstrates the efficiency of the algorithms for developing automated tool for clinical applications.

6 REFERENCES

1. Abdullah Alanazi: Using machine learning for healthcare challenges and opportunities, Informatics in Medicine Unlocked. pp. 1-5 (2022).
2. Pradeep Kumar Kushwaha, M. Kumaresan: Machine learning algorithm in healthcare system: A Review.2021 International Conference on Technological Advancements and Innovations (ICTAI), IEEE.
3. Pushpanjali Singh, Kumari Suniti Singh, Dr. Harsh Vikram Singh: Machine Learning for Healthcare: A Survey & its Algorithm for the Security of Medical Images.IEEE, (2021).
4. Mohammad Shahab, Laith Abualigah, Qusai Shambour , Muhannad A. Abu-Hashem ,Mohd Khaled Yousef Shambour , Ahmed Izzat Alsalibi , Amir H. Gandomi:Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine 145, pp.1-21, (2022).
5. Jobeda Jamal Khanam, Simon Y. Foo: A comparison of machine learning algorithms for diabetes prediction. ICT Express 7, pp.432–439, (2021).
6. Rahat Ullaha, Saranjam Khanb, Hina Alia, Iqra Ishtiaq Chaudhary, Muhammad Bilala, Iftikhar Ahmad: A comparative study of machine learning classifiers for risk prediction of asthma disease. Photodiagnosis and Photodynamic Therapy, 28, pp.292–296, (2019).
7. Forum Desai, Deepraj Chowdhury, Rupinder Kaur, Marloes Peeters, Rajesh Chand Arya , Gurpreet Singh Wander, Sukhpal Singh Gill , Rajkumar Buyya, Health Cloud: A system for monitoring health status of heart patients using machine learning and cloud computing, Internet of Things 17, 100485, (2022).
8. Ekta Maini, Bondu Venkateswarlu, Baljeet Maini, Dheeraj Marwaha: Machine learning based heart disease prediction system for Indian population: An exploratory study done in South India, Medical journal armed forces, India, 77,pp. 302-311,(2021)
9. Rajkamal Rajendran, Anitha Karthi: Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers. Expert Systems With Applications, 207 (2022)
10. Rakesh Chandra Joshi, Rashmi Mishra, Puneet Gandhi, Vinay Kumar Pathak, Radim Burget, Malay Kishore Dutta: Ensemble based machine learning approach for prediction of glioma and multi-grade classification. Computers in Biology and Medicine, 137,104829, (2021).
11. Vardhan Shorewala: Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26 100655, pp.1-8, (2021).
12. Shahid Mohammad Ganie, Majid Bashir Malik: An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. Healthcare Analytics, 21, 00092, pp. 1-14, (2022).
13. Tao Zheng, Wei Xie , Liling Xu , Xiaoying He , Ya Zhang , Mingrong You , Gong Yang , You Chen: A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, pp.120–127, (2017).
14. Saravanakumar, S., & Thangaraj, P. (2019). A computer aided diagnosis system for identifying Alzheimer's from MRI scan using improved Adaboost. Journal of medical systems, 43(3), 76.
15. Kumaresan, T., Saravanakumar, S., & Balamurugan, R. (2019). Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel support vector machine. Cluster Computing, 22(Suppl 1), 33-46.
16. Saravanakumar, S., & Saravanan, T. (2023). Secure personal authentication in fog devices via multimodal rank-level fusion. Concurrency and Computation: Practice and Experience, 35(10), e7673.
17. Thangavel, S., & Selvaraj, S. (2023). Machine Learning Model and Cuckoo Search in a modular system to identify Alzheimer's disease from MRI scan images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11(5), 1753-1761.
18. Saravanakumar, S. (2020). Certain analysis of authentic user behavioral and opinion pattern mining using classification techniques. Solid State Technology, 63(6), 9220-9234.

19. Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, Amit P. Sheth: Machine learning for internet of things data analysis: a survey. Digital Communications and Networks, 4, pp. 161–175, (2018).
20. Priyan Malarvizhi Kumar, Usha Devi Gandhi: A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases, Computers and Electrical Engineering. 65, pp.222–235, (2018).
21. Kashif Naseer Qureshi, Sadia Din, Gwanggil Jeon, Francesco Piccialli: An accurate and dynamic predictive model for a smart M-Health system using machine learning. Information Sciences, 538, pp.486–502, (2020).
22. Govindaraj Ramkumar, J. Seetha, R. Priyadarshini, M. Gopila, G. Saranya: IoT-based patient monitoring system for predicting heart disease using deep learning. Measurement,(2023).
23. Sekaran, R., Munnangi, A. K., Ramachandran, M., & Gandomi, A. H. (2022). 3D brain slice classification and feature extraction using Deformable Hierarchical Heuristic Model. Computers in Biology and Medicine, 149, 105990-105990.
24. Ramesh, S. (2017). An efficient secure routing for intermittently connected mobile networks. Wireless Personal Communications, 94, 2705-2718.
25. Sekaran, R., Al-Turjman, F., Patan, R., & Ramasamy, V. (2023). Tripartite transmitting methodology for intermittently connected mobile network (ICMN). ACM Transactions on Internet Technology, 22(4), 1-18.
26. Zafer Al-Makhadmeh, Amr Tolba: Utilizing IoT wearable medical device for heart disease prediction using higher order Boltzmann model: A classification approach. Measurement 147, 106815, pp. 1-9, (2019).
27. Evanthia E. Tripoliti , Theofilos G. Papadopoulos, Georgia S. Karanasiou , Katerina K. Naka, Dimitrios I. Fotiadis; Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Computational and Structural Biotechnology Journal 15, pp.26–47, (2017).
28. Nabila Sabrin Sworna, A.K.M. Muzahidul Islam, Swakkhar Shatabda, Salekul Islam: Towards development of IoT-ML driven healthcare systems: A survey. Journal of Network and Computer Applications, 196, (2021).