Revolutionizing Healthcare: Unleashing the Power of Machine Learning in Remote Diagnosis and Health Management

Parth Puneet[a] , Prakhar Shukla[b] and Gaurav Parashar[c]

[a,b,c] KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

ARTICLE HISTORY
Compiled April 2, 2024

ABSTRACT
The medical industry has increasingly used artificial intelligence (AI) as technology progresses. The primary advantage of machine learning comes in its capacity to derive significant insights from the immense volumes of health data gathered daily. The progress in the Internet of Things (IoT), machine learning, big data, and high-performance computing has created exceptional possibilities to enhance public health by maximizing scarce human resources. This work presents a diabetes prediction model that categorizes diabetes based on common parameters such as blood sugar, body mass index (BMI), age, insulin, and other relevant factors, as well as external factors that contribute to the development of diabetes. Our analysis utilizes the PIMA India and Sylhet datasets. Machine learning techniques, such as the Support Vector Machine (SVM), Naive Bayes classifier (NB), and decision tree, are used in the study.

KEYWORDS
—Healthcare, Artificial Intelligence, Machine Learning, Diabetes Prediction

1. Introduction

A strong state of well-being lays the foundation for a satisfying life, unlocking pathways to joy and success. The health index of a nation serves as a reflection, indicating its economic strength, scientific achievements, defense capabilities ties, and social unity. While health is traditionally defined as the "absence of diseases," its true essence lies in the timely identification and anticipation of illnesses, followed by prompt and suitable interventions. This necessity becomes even more pronounced in a nation as diverse as India, where a mosaic of demographics, climate variations, and socio-cultural intricacies converge. Extending our focus to the national level, the health index becomes apparent as a complex indicator of a country's general resilience and vibrancy. A country's economic strength is reflected in the health index, which acts as a mirror to show patterns of wealth or inequality. A thriving economy makes investments in public health programs, healthcare infrastructure, and research possible, all enhancing well-being overall.

Diabetes is a prevalent, lifelong condition that affects individuals of all ages. It occurs when the level of glucose in the blood becomes too high. According to the

Email: parth.2024cse1011@kiet.edu

World Health Organization (WHO)[1], the number of people diagnosed with diabetes surged from 108 million in 1980 to 422 million in 2014, highlighting a significant global health issue. In 2019 alone, an estimated 2 million deaths were attributed to diabetes and related kidney diseases. For Type 1 Diabetes[2], risk factors include a family history of the condition and age, although currently, there's no known method for preventing Type 1 Diabetes. In the case of Type 2 Diabetes, factors primarily include being overweight, being aged 45 years or older, having prediabetes, and having a family member with Type 2 Diabetes.

2. Literature Review

The experiment's primary goal is to design and implement a model for diabetes prediction that utilizes machine learning techniques successfully. Soni et al. [1] suggested ensemble learning and the use of classification techniques SVM, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting classifiers (GBC), and K-nearest neighbor (KNN). Furthermore, 77% classification accuracy was attained.

In their research, Xue et al. [2] use the NB as a classification technique. SVM achieved 96.54% accuracy. Tigga et al. [3] in the paper, implement six machine-learning classification strategies and compare the results with independent metrics. The tests were conducted on a dataset collected through online and offline surveys of 18 questions. The same statistics were also linked to the PIMA dataset (PIMA). Research results show that irregular timberland has an accuracy of 94.10%. Thomas and others [4] propose DT for diabetes prediction. For their research, they used the PIMA dataset. The accuracy of DT algorithms is examined and assessed in this work. The test results demonstrated that the devised system was effective, with an 87%. At small sample numbers, DT scales quickly and yields inaccurate response predictions. You et al. [5] conducted a study using two-class SVM and two-class DT to search for important diabetic factors in the PIMA dataset. It was found, through correlation analysis, that examining only the potential patient's glucose levels, BMI, and age is more efficient than conducting a comprehensive medical examination, which is time-consuming. The experiment with these features using SVM resulted in an accuracy of 70.4 %.

In the study, Rajni et. al [6] propose a procedure using the Bayes hypothesis for diabetes prediction utilizing machine learning methods to extract the necessary data so that the problem can be solved with high accuracy. Here, the mean is calculated to handle missing data and probability is used for calculating yes (positive) and no (negative) values. In comparison to previous controlled techniques, the suggested method enhances accuracy on the Pima dataset by 72.9%. In the study conducted by Sisodia D et al. [7], the Naive Bayes (NB), SVM, and DT algorithms are used among which The NB algorithm is considered the best machine learning method for this test because it has higher accuracy than other classification algorithms, with an accuracy of 76.30%. The study carried out by Mujumdar A et al. [8] suggests the use of numerous classification and common machine learning methods that yield the most precise results. Various classification methods include methods like LR, GBC, LDA, AdaBoost Classifier (ABC), Gaussian NB, Extra Trees Classifier (ETC), Bagging, RF, DT, Perceptron, SVC, and KNN on the PIMA dataset. Out of all, GBC, ETC, and ABC gave the highest accuracy of 77%.

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] https://www.cdc.gov/diabetes/basics/risk-factors.html

In the model, author Saru S et al. [9] employ ensemble methods by using various classifiers like DT, KNN, and NB on the PIMA dataset. Here, the highest prediction result of 78% was obtained using DT without bootstrap. Authors of [10] paper used seven machine-learning algorithms used on the Pima Dataset for prediction namely DT, KNN, RF, NB, AB, LR, and SVM. For some criteria, including precision, accuracy, recall, and F measure, all models exhibit good results. Each model has an accuracy of 70%. For the train/test and test phases, LR and SVM yielded accuracies of 77% and 78%, respectively. Another neural network approach was used to predict diabetes in which the hidden layers in the network model were 1, 2, and 3, and occasionally 200, 400, and 800. In 400 epochs, the accuracy of the second hidden layer was 88.6%, the highest of all the models used in PIDD. The research conducted by Zou et al.[11] uses two datasets for better diabetes prediction. The first dataset is from a hospital physical examination in Luzhou, China and the second dataset is Pima Indians. When RF is used, the accuracy for the Luzhou dataset is 80.84%, and for the Pima Indians dataset, the accuracy is 77.21%. Talha Mahboob and others [12] propose ANN for diabetes prediction. The accuracy of the ANN approach was 75.7%, the RF method was 74.7%, and the K-means clustering method was 73.6%. ANN outperforms other methods. Authors of paper [13] found that NB is better than the DT method J48. The results predicted that NB got 76.3021% accuracy followed by J48 with an accuracy of 73.8281%.

N.Gupta et al.[14] uses various classification techniques namely Multilayer Perception(MLP), J48, JRIP, and Bayes Network on Pima Indians dataset. The paper determined that J48 has the highest accuracy of 81.33%. A.Iyer et al. [15] proposed DT and NB algorithm for Diabetes Prediction on the PIMA dataset. Based on the cross-validation technique, the DT achieves an accuracy of 74.8698%. Based on the percentage spilt(70:30) technique, J48 gives 76.9565% accuracy and NB achieves an accuracy of 76.5652%. The paper [16] predicted diabetes in females by taking the PIMA dataset. Various supervised learning algorithms have been used such as SVM, KNN, NB, RF, CT, NN, AB, and LR. When 10-fold cross-validation is applied, LR outperforms other techniques. The AUC for LR is 0.825. Aziz Perdana and others [17], in their research paper have used the KNN method for classification. By employing KNN with a value of k=22, they achieved the highest accuracy of 83.12%. R. Sivanesan et al. [18] analyzed the performance of the J48 Decision Tree with different metrics. Based on the training set, the accuracy of correctly classified was 84.11%. After using 10-fold cross-validation, the accuracy of correctly classified was 73.82%.

3. Methodology

In our study, we use two diabetes-related datasets from UCI and Kaggle. These datasets contained different numbers of features and multiple inconsistencies. We use the following steps to ascertain the best performance of the models.

(1) Load and preprocess the data. Remove inconsistencies like missing values, outliers, inconsistent data, and duplicate rows. Perform feature scaling and normalization. Split the data into training and test sets.
(2) Select the models and perform hyperparameter tuning to achieve the best results.
(3) Finalize the model based on the performance metric.

- Dataset Collection The module covers gathering data and preprocessing it for further analysis. For the study, two datasets were identified from the UCI repos-

itory.

Dataset 1: Dataset-1[3] was gathered by doctors from patients from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh. The dataset contains attributes namely age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class. The dataset contains 520 entries.

Dataset 2: For the research, dataset-2[4] is the Pima Indians Diabetes dataset from kaggle. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Based on specific diagnostic metrics included in the collection, the dataset aims to predict diagnostically whether or not a patient has diabetes. These examples were chosen from a larger database under several restrictions. Specifically, all of the patients in this facility are Pima Indian women who are at least 21 years old. The dataset contains attributes: pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function, age, and outcome. This dataset has 768 entries, of which 268 are positive and the remaining 500 are negative for diabetes.

- Data Pre-Processing: The data pre-processing method improves results by handling superfluous data in the dataset in an efficient manner. But dealing with missing information is a major problem, especially regarding important variables like age, blood pressure, skinfold thickness, insulin, bmi, and blood sugar level. Since the values of these attributes must not have zero, therefore we used the imputation technique to remedy and preserve the integrity of the dataset. Our goal is to maintain data consistency and correctness by expanding the dataset to include all values, that enable more precise and significant analysis and forecasts.
- Model Selection The choice of a suitable machine-learning technique is crucial during this stage of creating a diabetes prediction model. Used different machine-learning methods, each with unique advantages and disadvantages, to accurately predict diabetes. SVM, DT, and NB are a few of the techniques used. The goal is to optimize both predictive accuracy and robustness while maintaining the consistency of the model's predictions by experimenting with different approaches.
  (1) Support vector classifier: SVM is a supervised machine learning algorithm that is used for both classification and regression tasks. For learning instances, the maximal reserve hyperplane serves as its crucial limit. SVM uses the hinge loss function to calculate the empirical risk and adds a solution system regularization term to optimize the structural risk. SVM can do non-linear classification thanks to the kernel method, one of the most often used kernel learning strategies.
  (2) Decision Trees: A non-parametric supervised learning technique for regression and classification is called a DT. The goal is to create a model that can forecast the value of a target variable using fundamental decision rules inferred from the data attributes.
  (3) Naïve Bayes: NB uses a set of algorithms based upon Bayes' theorem where the probability of previously occurred events is used to calculate the probability of posterior events. It is called naive as it takes the assumptions of features to be independent of each other. The classification model assigns class labels, represented by feature values, and the class labels are extracted

[3]https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset
[4]https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

from the dataset. For an item to be classified, it finds the maximum probability of each occurrence under the conditions that the item is considered as the item to be classified.

Table 1. Distinctive Paper Examination

Author	Algorithm	Dataset	Performance Metric
Soni [1]	RF	PIMA	77% classification accuracy
Xue [2]	SVM	Sylhet	Accuracy of 96.54%
Tigga [3]	RF	PIMA	Accuracy of 94%
Thomas [4]	Decision Tree	PIMA	Accuracy of 94%
S.You [5]	SVM	PIMA	Accuracy of 70.4%
Rajni [6]	RB-Bayes	PIMA	Accuracy of 72.9%
Sisodia [7]	Naive Bayes	PIMA	Accuracy of 76.30
Mujumdar[8]	Random Forest	PIMA	77% classification accuracy
Saru [9]	Decision Tree	PIMA	Accuracy of 78%
Khanam [10]	SVM	PIMA	Accuracy 77%-78%

The table 1 summarizes several studies using different machine learning algorithms to predict diabetes. It includes the author's name, the algorithm used, the dataset used, and its performance metrics, expressed as a percentage of accuracy. Algorithms exhibiting varying accuracy degrees on datasets include RF, SVM, DT, and NB. This highlights the significance of choosing algorithms depending on certain attributes. The table provides a quick comparative overview of the prediction performance of various algorithms in different papers.

4. Results & Conclusion

Among the various machine learning algorithms evaluated in our study, DT emerged as the most accurate model, exhibiting a diabetes prediction accuracy of 98%, refer table 2. We utilized various machine learning algorithms such as SVM, DT, and NB on the two datasets. The dataset from the Pima Indian dataset contains fewer parameters. Consequently, due to this limitation, the model's performance is not satisfactory. Conversely, when dealing with the Sylhet dataset, which encompasses more parameters and accounts for both genders, the model achieved higher accuracies. The highest accuracy defines the effectiveness of DT in handling the intricacies of datasets and making precise predictions. The study compared the accuracy of a machine learning algorithm using two different datasets. This comparative analysis revealed that the developed model significantly enhanced the accuracy and precision of diabetes prediction across various datasets.

Table 2. Comparison between accuracy of PIMA Dataset and Sylhet Dataset

Algorithm	PIMA Dataset	Sylhet Dataset
SVM	77.2%	93.2%
Decision Tree	74%	98%
Naive Bayes	77.2%	90.3%

The model's adaptability to different data sources underscores its relevance to real-

world situations. Additionally, the study lays the groundwork for further extensions. The natural progression involves examining the risk of mortality within the next few years among individuals with diabetes. Expanding the analysis to include predictors of diabetes in individuals without the condition, can furnish valuable insights for health interventions. This broadening of the research scope holds promise for advancing diabetes prediction and prevention strategies.

References

[1] Soni M, Varma S. Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (Ijert) Volume. 2020;9.

[2] Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. In: Journal of Physics: Conference Series; Vol. 1684; IOP Publishing; 2020. p. 012062.

[3] Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science. 2020;167:706–716.

[4] Thomas J, Joseph A, Johnson I, et al. Machine learning approach for diabetes prediction. International Journal of Information. 2019;8(2).

[5] YOU S, KANG M. A study on methods to prevent pima indians diabetes using svm. Korean Journal of Artificial Intelligence. 2020;8(2):7–10.

[6] Rajni R, Amandeep A. Rb-bayesalgorithm for the predictionof diabetic in pima indian dataset. International Journal of Electrical and Computer Engineering. 2019;9(6):4866.

[7] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018;132:1578–1585.

[8] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019;165:292–299.

[9] Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering. 2019;5(4).

[10] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. Ict Express. 2021;7(4):432–439.

[11] Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics. 2018;9:515.

[12] Alam TM, Iqbal MA, Ali Y, et al. A model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019;16:100204.

[13] Diwani SA, Sam A. Diabetes forecasting using supervised learning techniques. Adv Comput Sci an Int J. 2014;3:10–18.

[14] Gupta N, Rawal A, Narasimhan V, et al. Accuracy, sensitivity and specificity measurement of various classification techniques on healthcare data. IOSR Journal of Computer Engineering (IOSR-JCE). 2013;11(5):70–73.

[15] Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:150203774. 2015;.

[16] Bhoi SK, et al. Prediction of diabetes in females of pima indian heritage: a complete supervised learning approach. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021;12(10):3074–3084.

[17] Perdana A, Hermawan A, Avianto D. Analyze important features of pima indian database for diabetes prediction using knn. Jurnal Sisfokom (Sistem Informasi dan Komputer). 2023;12(1):70–75.

[18] Sivanesan R, Dhivya KDR. A review on diabetes mellitus diagnoses using classification on pima indian diabetes data set. International Journal of Advance Research in Computer Science and Management Studies. 2017;5(1).