JOURNAL HEALTH NPEPS

http://dx.doi.org/10.30681/2526101014416 ORIGINAL ARTICLE

Application of machine learning on the prediction of diabetes screening outcomes using common symptoms in resource-poor settings

Aplicación del aprendizaje automático en la predicción de los resultados del cribado de diabetes mediante síntomas comunes en entornos con recursos limitados

Aplicação do aprendizado de máquina na predição dos resultados do rastreamento de diabetes utilizando sintomas comuns em contextos com recursos limitados

Yashik Singh[1]

ABSTRACT

Objective: to evaluate whether a machine learning algorithm could accurately predict diabetes screening outcomes using easily recognized symptoms rather than laboratory or physiological measurements. **Method:** de-identified patient data from Sylhet Diabetes Hospital in Bangladesh, were used. Fourteen symptoms, along with age and gender, were input into nine supervised machine learning models: Random Tree, C4.5, C-RT, CS-MC4, Linear Discriminant Analysis, Rule Induction, Decision List, ID3, and Partial Least Squares. **Results:** the models effectively predicted diabetes status based on symptoms, achieving an average accuracy of 94.2% ± 4, TPR of 93.4% ± 4, TNR of 95% ± 5, FPR of 4.6% ± 5, FNR of 6.6% ± 5, and an F-measure of 94.3% ± 4. **Conclusion**: the Random Tree algorithm performed best and shows strong potential for development into a user-friendly screening tool to encourage timely medical evaluation.

Descriptors: Diabetes Mellitus; eHealth; Machine Learning; Electronics, Medica; Biomedical Technology.

RESUMEN

***Objetivo**: evaluar si un algoritmo de aprendizaje automático podía predecir con precisión los resultados del cribado de diabetes utilizando síntomas fácilmente reconocibles en lugar de mediciones de laboratorio o fisiológicas. **Método**: se utilizaron datos anónimos de pacientes del Hospital de Diabetes Sylhet en Bangladesh. Se introdujeron catorce síntomas, junto con la edad y el sexo, en nueve modelos de aprendizaje automático supervisado: Árbol aleatorio, C4.5, C-RT, CS-MC4, Análisis discriminante lineal, Inducción de reglas, Lista de decisiones, ID3 y Mínimos cuadrados*

[1]Computer Scientist. PhD Medical Informatics. Senior Lecturer at the University of KwaZulu-Natal. Durban, KwaZulu-Natal, South Africa. E-mail: singhyashik@gmail.com ORCID ID: https://orcid.org/0000-0001-9676-8169 **Corresponding author** – Address: 238 Mazisi Kunene Rd, Glenwood, Durban, 4041, South Africa.

CITATION: Singh Y. Application of machine learning on the prediction of diabetes screening outcomes using common symptoms in resource-poor settings. J Health NPEPS. 2026; 11(1):e14416.

EDITOR IN CHIEF: Vagner Ferreira do Nascimento

*parciales. **Resultados:** los modelos predijeron eficazmente el estado de diabetes basándose en los síntomas, alcanzando una precisión promedio del 94,2 % ± 4, un TPR del 93,4 % ± 4, un TNR del 95 % ± 5, un FPR del 4,6 % ± 5, un FNR del 6,6 % ± 5 y una medida F del 94,3 % ± 4. **Conclusión:** el algoritmo de Árbol Aleatorio obtuvo el mejor rendimiento y muestra un gran potencial para convertirse en una herramienta de detección intuitiva que fomente la evaluación médica oportuna.*
***Descriptores:** Diabetes Mellitus; Salud Electrónica; Aprendizaje Automático; Electrónica Médica; Tecnología Biomédica.*

RESUMO
***Objetivo:** avaliar como um algoritmo de aprendizado de máquina poderia prever com precisão os resultados da triagem de diabetes usando sintomas facilmente reconhecidos em vez de medições laboratoriais ou fisiológicas. **Método:** foram utilizados dados de pacientes anonimizados do Hospital de Diabetes de Sylhet, em Bangladesh. Quatorze sintomas, juntamente com idade e sexo, foram inseridos em nove modelos de aprendizado de máquina supervisionado: Random Tree, C4.5, C-RT, CS-MC4, Análise Discriminante Linear, Indução de Regras, Lista de Decisão, ID3 e Mínimos Quadrados Parciais. **Resultados:** os modelos previram eficazmente o estado da diabetes com base nos sintomas, atingindo uma precisão média de 94,2% ± 4, TPR de 93,4% ± 4, TNR de 95% ± 5, FPR de 4,6% ± 5, FNR de 6,6% ± 5 e uma medida F de 94,3% ± 4. **Conclusão:** o algoritmo Random Tree apresentou o melhor desempenho e demonstra um forte potencial para ser desenvolvido como uma ferramenta de triagem fácil de usar, incentivando a avaliação médica oportuna.*
***Descritores:** Diabetes Mellitus; eSaúde; Aprendizado de Máquina; Eletrônica Médica; Tecnologia Biomédica.*

INTRODUCTION

Diabetes is a chronic medical condition that affects how the body processes glucose, which is the primary source of energy for cells. It occurs when the body has difficulty in regulating the levels of glucose in the bloodstream[1].

There are two main types of diabetes. Type-1 diabetes occurs due to an autoimmune condition where the body's immune system mistakenly attacks and destroys insulin-producing cells in the pancreas. Type-2 diabetes is the most common and is often associated with lifestyle factors like obesity, physical inactivity, and poor dietary choices. In Type-2 diabetes, the body becomes resistant to the effects of insulin, and the pancreas can not produce enough insulin to maintain normal blood sugar levels[2].

The prevalence of diabetes on a global scale has reached alarming proportions, with the number of affected individuals steadily rising. It is globally estimated that 734 million people are living with diabetes and this is projected to increase significantly to 822 million people by 2040[3]. The Type-2 lifestyle disease is increasing in prevalence in resource poor countries.

Projections of the burden of diabetes indicate a 170% increase in the resource poor countries by 2025. More than 75% of people with diabetes will reside in developing countries, as compared with 62% in 1995[4]. Thus, this epidemic knows no borders, affecting people from all walks of life and regions. One of the major challenges associated with diabetes is the often silent and gradual onset of the condition.

This means that compounded with the diagnosed diabetes population, there is the added burden of the large number of undiagnosed patients with diabetes. Worldwide, 44% of people living with diabetes are not diagnosed. The highest proportions of undiagnosed diabetes is found in resource poor countries like Africa (54%), Western Pacific (53%) and South-East Asia regions (51%)[5].

It was reported that 35% of undiagnosed patients had complications at the time of diagnosis which included retinopathy, neuropathy,nephropathy, is chemic heart disease, atherosclerotic cerebrovascular disease and peripheral arterial disease[6]. Thus, early detection and diagnosis play a critical role in mitigating the disease's progression and preventing complications.

Studies have shown that individuals from poor resource countries or low socioeconomic standing tend to have poorer glycemic control, more diabetic complications, and higher mortality[7]. This has been attributed to poor health seeking behavior due to lack of financial stability[8]. Many rural patients are hesitant to take leave from work and travel long distances to healthcare facilities for consults with practitioners. This is due to the fact that most rural patients will have to dock a day's pay, incur extra expenses with regard to transport, find someone to look after children at home, and spend an entire day at the healthcare facility etc. Because of large numbers of patients at the public healthcare facilities, those seeking HbA1c tests often fast the entire day.

The use of machine learning algorithms to predict if a patient is diabetic may be of benefit to both the patient and healthcare facility. Patients will only visit the healthcare facility for tests if they are flagged as possibly diabetic by the machine learning algorithm. This reduces the financial and time burden on the patients, as well as saves resources in the rural health care facility. Given the rising resource poor country prevalence of diabetes, raising

awareness about the importance of early detection, coupled with accessible healthcare resources and education, is crucial in curbing the impact of this widespread health concern.

An Adaboost machine learning algorithm that used mostly physiological data such at blood glucose level, blood pressure, skin fold thickness etc., was developed to predict the diagnoses of diabetes. The algorithm produced an accuracy of 80.7%[9]. Given a patients' demographic characteristics, lifestyle habits, and medical history as input into a back propagation algorithm, the onset of diabetes was predicted with 83.1% accuracy[10].

An Xboost classifier applied to an input of pregnancy, glucose, blood pressure, skin thickness, BMI, and age, produced an accuracy of 81%[11]. Dietary information like the number of calories in daily meals, physiological test results and examination data was used to predict diabetes with an accuracy of 82.1%[12].

Given the context of a resource poor countries, the aim of this study is to evaluate whether a machine learning algorithm could accurately predict diabetes screening outcomes using easily recognized symptoms rather than laboratory or physiological measurements. This algorithm can be converted into an easy to use app that rural patients can use from the comfort of their own homes.

METHOD

Machine learning is a subset of artificial intelligence that empowers computers to learn and make predictions or decisions without being explicitly programmed for each task. At its core, machine learning relies on data and algorithms to detect patterns and make sense of complex information.

The process typically begins with collecting and preprocessing data, followed by the selection of an appropriate machine learning algorithm. During the training phase, the algorithm iteratively processes the data, adjusting its internal parameters to minimize the difference between its predictions and the actual outcomes. Once the model is trained, it can be used to make predictions or classifications on new, unseen data.

The key strength of machine learning lies in its ability to handle vast amounts of data and discover intricate relationships, enabling it to solve a wide range of tasks, from image recognition and natural language processing to

medical diagnosis and financial forecasting, revolutionizing various industries.

Publicly available de-identified data was obtained from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. This data is freely available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at the UCI data repository (http://archive.ics.uci.edu/). The data was collected and verified by Islam[13].

All records contained within the publicly available Sylhet Diabetes Hospital dataset were eligible for inclusion if they met the following criteria:

1. Individuals with complete data for all predictor variables, including age, sex, and the 14 diabetes-related symptoms (polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity).
2. Records with a clearly defined outcome variable indicating diabetes status (diabetic or non-diabetic).
3. De-identified patient data available through the UCI Machine Learning Repository under an open-access license.

Records were excluded from analysis if any of the following conditions were present:

1. Missing, incomplete, or inconsistent data in any of the predictor or outcome variables.
2. Duplicate entries identified during data preprocessing.
3. Records lacking a definitive classification of diabetes status.

The data consisted of 520 instances with 17 attributes. Each data instance consisted of the following input attributes: age (numeric), sex (Male, Female), polyuria (Yes, No), polydipsia (Yes, No), sudden weight loss (Yes, No), weakness (Yes, No), polyphagia (Yes, No), genital thrush (Yes, No), visual blurring (Yes, No), itching (Yes, No), irritability (Yes, No), delayed healing (Yes, No), partial paresis (Yes, No), muscle stiffness (Yes, No), alopecia (Yes, No), and obesity (Yes, No). The output predicted was whether the patients are Diabetic or not Diabetic. Five-fold cross validation was used to split the data into training and testing datasets.

Five-fold cross-validation is a technique used in machine learning and statistical modeling to assess the performance and generalization of a

predictive model. It is a process that helps ensure that a model's performance evaluation is reliable and representative of its ability to make accurate predictions on new, unseen data. The dataset is divided into five roughly equal subsets, or folds. The model is trained and evaluated five times, with a different fold used as the test set in each iteration and the remaining four as the training set. This means that each fold serves as the test set once, while the others are used for training.

This investigation constitutes a retrospective diagnostic prediction model study using publicly available de-identified data. The objective was to develop and internally validate supervised machine learning models to predict diabetes screening outcomes based on symptomatology. The study adhered to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline for prediction model development and validation studies.

Tanagra Version 1.4.50 (2003) was used to build nine supervised machine learning algorithms to predict diabetes diagnosis given 14 symptoms, age and gender as input. These were random tree, C4.5 decision tree algorithm, C-RT classification tree algorithm, CS-MC4 cost sensitive decision tree algorithm, linear discriminant analysis (LDA), rule induction, decision list, ID3 tree learning, and partial least squares (LPS). Tanagra is a free, open-source data mining and machine learning software suite developed for academic instruction and research applications, by Ricco Rakotomalala at Lumière University Lyon.

The Random tree algorithm is an ensemble machine learning algorithm that utilizes decision trees. It constructs a multitude of decision trees during training, with each tree independently making a prediction. The final prediction is determined by aggregating these individual tree outputs, often using a majority vote for classification or averaging for regression. Random tree algorithm reduces overfitting, improves model accuracy, and handles high-dimensional or noisy data effectively through the incorporation of diverse decision trees, making them a robust choice for various classification and regression tasks.

C4.5 is a classic machine learning algorithm for decision tree construction. It recursively splits data based on attribute values to create a tree. It selects attributes using

information gain and generates rules for classification. Pruning is employed to avoid overfitting, resulting in interpretable decision trees used for classification and regression tasks.

Classification and Regression Trees (C-RT) is a machine learning algorithm that builds binary trees for classification and regression. It recursively partitions data by choosing the optimal attribute and split point. For classification, Gini impurity measures the quality of splits, and for regression, mean squared error is used.

CS-MC4 is an extension of the C4.5 decision tree algorithm designed to handle imbalanced datasets and cost-sensitive classification problems. It incorporates multiple objectives, including class distribution and misclassification costs, to create decision trees that prioritize rare or costly misclassifications while maintaining good overall accuracy. This makes it well-suited for applications where certain errors are costlier or critical than others.

Linear Discriminant Analysis (LDA) is a dimensionality reduction and classification algorithm. It seeks to maximize the distance between class means while minimizing within-class variance. LDA transforms data into a lower-dimensional space, emphasizing class separability. It's used for classification, feature selection, and data preprocessing in various machine learning applications.

Rule induction is a machine learning technique that generates decision rules from data. It employs algorithms to find concise, human-readable rules to classify or predict outcomes. These rules are constructed based on feature values, and they assist in understanding the logic behind a model's predictions or classifications.

Decision list learning is a machine learning approach where a series of if-then rules are generated to make decisions. It's typically used for classification tasks, where each rule tests a specific condition and leads to a particular class assignment. Decision lists are interpretable, simple models suitable for various domains.

Iterative Dichotomiser 3 (ID3) is a classic decision tree learning algorithm. It builds a tree by recursively selecting attributes based on information gain and entropy measures to maximize information gain. It's used for classification tasks, creating interpretable trees by iteratively splitting data into subsets until a stopping criterion is met.

Partial Least Squares finds latent variables by maximizing the covariance between input features and the target variable. PLS is beneficial when dealing with high-dimensional data and multicollinearity, making it useful in fields like chemometrics, quantitative modeling and big data biinfrmatics.

Accuracy, specificity, false-positive rate (FPR), false-negative rate (FNR), specificity or true-negative rate (TNR), sensitivity or true-positive rate (TPR), and $F_{measure}$ were the statistical measures that were used to calculate the efficacy of the machine learning algorithms in predicting a positive or negative diabetes diagnosis. Equations 1 – 6 show how these statistical measures were calculated.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$TPR = \frac{TP}{TP+FN} \quad (2)$$

$$TNR = \frac{TN}{TN+FP} \quad (3)$$

$$FPR = \frac{FP}{FP+TN} \quad (4)$$

$$FNR = \frac{FN}{FN+TP} \quad (5)$$

$$F_{measure} = 2\,X\,\frac{Sensitivity\ X\ Specificity}{Sensitivity + Specificity} \quad (6)$$

Where TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives values respectively.

Z-score and *p*-score were calculated in order to perform a statistically significant proportion test as per equation 7 and 8. This was calculated to determine if the statistical measures (accuracy, TPR, TNR, FPR, FNR, $F_{measure}$) obtained where statistically different from chance, and also if it is statistically different from each other. ANOVA test was also performed to determine if there were statistically significant differences between the various classification models, with alpha set at 5%. Where *p* refers to the proportion of the measure tested and P is the proportion of the average measure.

$$\Omega = \sqrt{\frac{P\times(1-P)}{n}} \quad (7)$$

$$Z - Score = \frac{p-P}{\Omega} \quad (8)$$

RESULTS

A power analysis was conducted to determine the minimum sample size required to ensure that the results are statistically sound. With alpha = 0.05, Beta = 0.2, and power = 0.8, the minimum sample size required is 274. The data size of 520 exceeds this, thus indicating that statistical tests performed can detect a significant difference. A Table 1 shows the results of the accuracy, TPR, TNR, FPR, FNR and $F_{measure}$ calculations for each of the eight machine learning algorithms. In order to determine if the algorithm produced results that were statistically different from pure chance, Z-scores was calculated according to equations 7 and 8, with P = 50% (binary chance) for accuracy and $F_{measure}$. Z-scores for each individual algorithm associated with accuracy and $F_{measure}$ produced a *p*-value < 0.001.

This indicated that there is a statistical difference between the results produced by the algorithms and binary chance. This thus indicates, that the machine learning algorithms built have the ability to predict if a patient is diabetic or not given the input of the 14 symptoms, gender and age.

Table 1 - Shows the accuracy, False-positive rate (FPR), false-negative rate (FNR), specificity or true-negative rate (TNR), sensitivity or true-positive rate (TPR) and $F_{measure}$ of each algorithm. *p < 0.001.

Algorithm	Accuracy	TPR	TNR	FPR	FNR	F Measure
Random Tree	99,8*	100,0*	99,5*	0,5	0,0*	99,7*
C4.5	96,5	95,0	99,0*	1,0	5,0	97,0
C-RT	96,5	96,6*	96,5	3,5	3,4	96,5
CS-MC4	96,5	95,0	96,5	3,5	5,0	95,7
LDA	91,3	88,1	96,5	3,5	11,9*	92,1
Rule Induction	96,1	96,2*	96,0	4,0	3,8	96,1
ID3	86,1*	89,7*	83,0*	17,0*	10,3*	86,2*
Decision List	94,3	92,5	97,0	3,0	7,5	94,7
PLS	90,4	87,8	94,5	5,5	12,2*	91,0

The machine learning algorithms together produced an average accuracy of 94.2% ± 4, TPR of 93.4% ± 4, TNR of 95.0% ± 5, FPR of 4.6% ± 5, FNR of 6.6% ± 5 and $F_{measure}$ of 94.3% ± 4. The high average, TPR, TNR and $F_{measure}$, coupled with the low values for FPR and FNR shows that the machine learning

algorithms in general are efficient and effective in predicting if a patient is diabetic or not.

Z-scores were calculated for comparing each algorithm's statistical measures (accuracy, TPR, TNR etc.) and the associated average statistical measures of all eight machine learning algorithms. The associated *p*-value for the Random tree and ID3 algorithms were less than 0.001. This indicates that the computational efficiency to predict diabetes diagnosis was statistically significant for Random tree and ID3 when compared to the average statistical measures. Random tree performed better than the average and ID3 performed less effectively. C4.5, C-RT, CS-MC4, LDA, Rule Induction, Decision list, and PLS showed not statistical difference when compared to the average of all eight machine learning algorithms.

ANOVA tests produced a *p*-value <0.001 for each statistical measure. This further indicates that there is a difference in the ability of each algorithm to predict the diabetic outcome. Table 2 shows the percentage improvement of the Random tree algorithm in predicting the accuracy, TPR, TNR, FPR, FNR and $F_{measure}$ calculations when compared to each of the other machine learning algorithms. On average there was in improvement of 7% ± 5 in accuracy, 8% ± 4 in TPR, 5% ± 6 in TNR, 83% ± 14 in FPR, 112% ± 55 in FNR and 7% ± 4 in $F_{measure}$ calculations. Thus, it may be concluded that Random tree outperformed all other algorithms in predicting the diabetic outcome.

Table 2 - showing the percentage improvement of Random tree in predicting the diabetic outcome compared to the other algorithms.

Algorithm	Accuracy	TPR	TNR	FPR	FNR	$F_{Measure}$
C4.5	3	5	1	50	76	3
C-RT	3	4	3	86	52	3
CS-MC4	3	5	3	86	76	4
LDA	9	13	3	86	180	8
Rule Induction	4	4	4	88	58	4
ID3	16	11	20	97	156	16
Decision List	6	8	3	83	114	5
PLS	10	14	5	91	185	10

The performance of the Random Tree algorithm was further compared with previously published machine learning models. The comparative analysis is presented in Table 3.

Table 3 - Demonstrating the comparison of the Random Tree algorithm with literature.

Paper	Accuracy	$F_{measure}$	TPR	TNR
(Vijayan and Anjali 2015)[9]	80,7*			
(Woldemichael and Menaria 2018)[10]	83,1*			
(Tasin, Nabil et al. 2023)[11]	83*			
(Qin, Wu et al. 2022)[12]	82*			
(Pranto, Mehnaz et al. 2020)[14]	78*	84*	89*	81*
(Ahmed, Ahammed et al. 2021) [15]	96*			
(Islam and Jahan 2017)[16]	78*		89*	80*
(Varma and Panda 2019)[17]	74*			
(Rathore, Chauhan et al. 2017) [18]	82*			
(Radja and Emanuel 2019)[19]	77*	76*	90*	79*
Alghamdi, T (2023)[20]	89*			

The Random Tree algorithm generated a series of classification rules during model training. The extracted rules are summarized in Table 4.

Table 4 - Shows the rules created by the Random tree algorithm during the learning process.

Polydipsia	Itching	Gender	Alopecia	Polyuria	Irritability	Sudden Weight Loss	partial paresis	Genital thrush	age	Polyphagia	Obesity	Muscle Stiffness	Delayed Healing	Blurd vision	Outcome
Yes	Yes	Male		No	No		No			No	Yes				Diabetic
Yes	Yes	Male		No	No		Yes			No	Yes				Not Diabetic
Yes	Yes	Male		Yes	No					No					Diabetic
Yes	Yes	Male			Yes					No					Diabetic
Yes	Yes	Male	Yes					No	<57	Yes	Yes	No/ Yes			Diabetic
Yes	Yes	Male	Yes					No	>57	Yes					Diabetic
Yes	Yes	Male								Yes					Diabetic
Yes	Yes	Female													Diabetic
Yes															Diabetic
No		Male	No			No				No/ Yes		Yes	Yes	No /Yes	Not Diabetic
No	Yes	Male	No			No						Yes	No		Diabetic
No	Yes	Male	No		No	No						Yes	No		Not Diabetic
No	Yes	Male	No		Yes	No						Yes	No		Diabetic
No		Male	Yes			No			<41			No		No	Diabetic
No		Male	No			No		No	<41			No		No	Not Diabetic
No		Male	No			No		Yes	<41			No		No	Diabetic
No		Male		No	No	Yes	No		>40	No					Not Diabetic
No		Male		No	Yes	Yes	No		>40	No					Diabetic

No		Male		No	No	Yes	Yes		>40	No					Diabetic
No		Female	Yes							No					Not Diabetic
No		Female	Yes							Yes					Not Diabetic
No		Female	No										Yes	No	Diabetic
No		Female	No										Yes	Yes	Diabetic
No		Female	Yes												Diabetic

Table 4 shows the rules created by the Random tree algorithm during the learning process. Of all the symptoms in the input space of the algorithm, polydipsia, itching, alopecia, sudden weight loss, polyphagia, polyuria, irritability, muscle stiffness, delayed healing and gender were most prevalent in all the rules. Weakness contributed the least to the rules in predicting if a patient was diabetic or not. Polydipsia appeared in all the rules to predict if a patient is diabetic or not.

DISCUSSION

It has been shown that it is possible to use machine learning to predict a diabetic diagnosis using the presence of common symptoms, gender and age as input. Not only is it possible, but it was further shown, that the machine learning algorithms are very effective and have high accuracies, TPR, TNR and $F_{measures}$. This coupled with low FNR and FPR shows that the algorithms created can predict the diabetes outcome with high validity. Off all the algorithms created, the Random tree algorithm produced the best results.

The random tree algorithm produced an error rate of 0.2%. This means that during the cross-validation testing, the algorithm incorrectly predicated the diabetic outcome only 1 out 500 attempts. The algorithm correctly predicated the all the patients that have diabetes 100% of the time, and 99.5% of the time correctly predicted when the patient was not diabetic.

The $F_{measure}$ combines both precision and recall to provide a more comprehensive evaluation of the model's performance. It provides a better measure of the model's performance and is used when the input/output distribution are imbalanced or when the cost of false positives and false negatives is important for that particular domain. The high $F_{measure}$ of the random tree algorithm indicated that the

performance of predicting the diabetic outcome is very high.

Z-tests were used to compare the performance of the Random tree algorithm and results produced from literature. Table 3 shows the results of each comparison. For all the statistical measures, a *p*-value <0.001 was obtained. This indicates that the proposed random tree algorithm outperforms all the reported models described in the reported literature.

Earlier investigations using physiological and laboratory variables have reported moderate predictive performance. For example, Vijayan and Anjali[9] reported an accuracy of 80.7%, while Woldemichael and Menaria[10] achieved 83.1% using demographic and clinical variables. Similarly, Tasin et al[11] reported an accuracy of 83% using an XGBoost classifier, and Qin et al[12] achieved 82.1% using lifestyle and physiological data. These studies primarily relied on biochemical or clinical measurements such as glucose levels, blood pressure, or body mass index.

In contrast, the present study utilized only self-reported symptoms, age, and gender. Despite the absence of laboratory parameters, the Random Tree algorithm achieved higher predictive performance than most previously reported models. This suggests that symptom-based screening tools may have substantial discriminative potential, particularly in resource-constrained settings where laboratory testing is not readily accessible.

Ahmed et al[15] reported a high accuracy of 96% using machine learning for diabetes prediction. However, their study incorporated clinical and laboratory variables and developed a web-based application. Similarly, Pranto et al[14] reported an F-measure of 84% and sensitivity of 89% in a female Bangladeshi population using structured clinical data. Compared to these studies, the current findings demonstrate comparable or superior classification performance while relying exclusively on easily recognizable symptoms.

Additionally, prior investigations frequently evaluated a limited number of algorithms. The present study compared nine supervised machine learning techniques, enabling a broader comparative assessment of classifier performance within the same dataset. The consistent superiority of the Random Tree model over alternative classifiers suggests that ensemble-based decision tree approaches may better capture

nonlinear interactions among symptom variables.

From a contextual perspective, most previously published models were developed using datasets that included laboratory-confirmed glucose measurements. Such models, although clinically informative, require access to healthcare facilities. In contrast, symptom-based models may serve as preliminary screening tools that encourage timely medical consultation, particularly in low-resource environments where undiagnosed diabetes remains prevalent.

Despite the promising findings, several limitations must be acknowledged. First, this study was based on secondary analysis of a publicly available dataset, which limits control over data collection procedures, measurement accuracy, and verification of symptom reporting. The high accuracy achieved by the Random Tree algorithm should be interpreted cautiously. One possible explanation is that the dataset was relatively clean and structured, with complete symptom data and limited variability, making classification easier than in real-world clinical settings.

The temporal context of data collection and clinical confirmation procedures were not directly accessible, potentially introducing information bias. Second, the sample size (n=520) is relatively modest for prediction model development, and although five-fold cross-validation was performed, no external validation using an independent dataset was conducted. The absence of external validation limits generalizability to other populations and healthcare settings. Third, the dataset originated from a single geographic region (Bangladesh), which may restrict applicability to populations with different demographic, socioeconomic, or epidemiological characteristics.

Fourth, symptom variables were self-reported, which may introduce recall bias or subjective interpretation variability. Finally, although multiple classifiers were compared, hyperparameter optimization was not exhaustively explored, which may influence relative performance estimates.

CONCLUSION

It is thus concluded, that the efficient performance of the Random tree algorithm can produce a valuable app for many patients who do not know if they are diabetic or need an informed push to visit healthcare professionals for testing.

The study offers several novel contributions to the scientific literature. Unlike many previous diabetes prediction models that rely heavily on laboratory or physiological measurements, this study demonstrates that high predictive performance can be achieved using only easily recognizable symptoms, age, and gender. This approach directly addresses screening challenges in low-resource settings where access to biochemical testing may be limited.

Furthermore, the simultaneous comparison of nine supervised machine learning algorithms within the same dataset provides a comprehensive methodological evaluation rarely presented in prior studies. The consistent superiority of the Random Tree model highlights the potential utility of ensemble-based decision tree methods in symptom-driven diagnostic prediction. Importantly, the study proposes a framework that can be translated into a low-cost, user-centered digital screening application, thereby bridging computational modeling with practical public health implementation.

Collectively, these considerations position the study as a meaningful contribution to the growing body of research on artificial intelligence–assisted diabetes screening, particularly within the context of resource-constrained healthcare environments. The current study may be furthered to include building the mobile app with the rules produced from this study and extending the testing into the field.

REFERENCES

1. Polonsky KS. The past 200 years in diabetes. N Engl J Med. 2012; 367(14):1332-40.
2. Unnikrishnan R, Pradeepa R, Joshi SR, Mohan V. Type 2 Diabetes: Demystifying the Global Epidemic. Diabetes. 2017; 66(6):1432-42.
3. Sifunda S, Mbewu AD, Mabaso M, Manyaapelo T, Sewpaul R, Morgan JW, et al. Prevalence and Psychosocial Correlates of Diabetes Mellitus in South Africa: Results from the South African National Health and Nutrition Examination Survey (SANHANES-1). Int J Environ Res Public Health. 2023; 20(10).
4. King H, Aubert RE, Herman WH. Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diabetes Care. 1998; 21(9):1414-31.
5. Ogurtsova K, Guariguata L, Barengo NC, Ruiz PL-D, Sacre JW, Karuranga S,

et al. IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. Diabetes Res Clin Pract. 2022; 183:109118.

6. Gedebjerg A, Almdal TP, Berencsi K, Rungby J, Nielsen JS, Witte DR, et al. Prevalence of micro- and macrovascular diabetes complications at time of type 2 diabetes diagnosis and associated clinical characteristics: A cross-sectional baseline study of 6958 patients in the Danish DD2 cohort. JDC. 2018; 32(1):34-40.
7. Walker JJ, Livingstone SJ, Colhoun HM, Lindsay RS, McKnight JA, Morris AD, et al. Effect of socioeconomic status on mortality among people with type 2 diabetes: a study from the Scottish Diabetes Research Network Epidemiology Group. Diabetes Care. 2011; 34(5):1127-32.
8. Weng C, Coppini DV, Sönksen PH. Geographic and social factors are related to increased morbidity and mortality rates in diabetic patients. Diabet Med. 2000; 17(8):612-7.
9. Vijayan VV, Anjali C. Prediction and diagnosis of diabetes mellitus—A machine learning approach. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS); 2015: IEEE.
10. Woldemichael FG, Menaria S, editors. Prediction of diabetes using data mining techniques. 2018 2nd international conference on trends in electronics and informatics (ICOEI); 2018: IEEE.
11. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2023; 10(1-2):1-10.
12. Qin Y, Wu J, Xiao W, Wang K, Huang A, Liu B, et al. Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. Int J Environ Res Public Health. 2022; 19(22).
13. Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. In: Gupta M, Konar D, Bhattacharyya S, Biswas S. Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing. 2020; 992.
14. Pranto B, Mehnaz SM, Mahid EB, Sadman IM, Rahman A, Momen S. Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. Info. 2020; 11(8):374.

15. Ahmed N, Ahammed R, Islam MM, Uddin MA, Akhter A, Talukder MA, et al. Machine learning based diabetes prediction and development of smart web application. IJCCE. 2021; 2:229-41.
16. Islam MA, Jahan N. Prediction of onset diabetes using machine learning techniques. Int J Comput Appl. 2017; 180(5):7-11.
17. Varma KM, Panda DB. Comparative analysis of Predicting Diabetes Using Machine Learning Techniques. J Emerg Technol Innov Res. 2019; 6:522-30.
18. Rathore A, Chauhan S, Gujral S. Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women. Int J Adv Comput Sci. 2017; 8(5).
19. Radja M, Emanuel AWR, editors. Performance evaluation of supervised machine learning algorithms using different data set sizes for diabetes prediction. 2019 5th international conference on science in information technology (ICSITech); 2019: IEEE.
20. Alghamdi T. Prediction of Diabetes Complications Using Computational Intelligence Techniques. Appl Sc. 2023; 13(5):3030.

Funding: The author declare that there was no funding.

Conflict of interest: The authors declare no conflict of interest.

Authors' participation:
- **Design:** Singh Y.
- **Development:** Singh Y.
- **Writing and proofreading:** Singh Y.

Submited: 19/12/2025
Accepted: 25/04/2026