ZANCO Journal of Pure and Applied Sciences

TYPE Original article
DOI http://dx.doi.org/10.21271/ZJPAS.37.6.7

OPEN ACCESS

*Corresponding author

Salar Ameen Raheem
salar.raheem@epu.edu.iq

RECEIVED :19 /01 /2025
ACCEPTED :26/06/ 2025
PUBLISHED :31/ 12/ 2025

KEYWORDS:

Diabetes, Dataset, Machine Learning, Healthcare Systems, Diseases Diagnosis.

Machine learning for diabetes diagnosis: insights from the Erbil Diabetes Dataset and algorithmic performance

Salar Ameen Raheem[1*], Amal Taha Mawlood[2] , Ibrahim Ismael Hamarash[3]

[1]Department of Information and Communication Technology Eng., Erbil Polytechnic University, Erbil, Kurdistan Region, Iraq
[2]Department of Computer Science, Knowledge University, Erbil, Kurdistan Region, Iraq
[3]Department of Electrical Eng., College of Engineering, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq

ABSTRACT

Machine learning technologies have brought significant operational improvements for disease diagnosis-related healthcare activities. Among various conditions, diabetes is particularly suited for prediction through historical and personalized data, which serve as a cornerstone of many machine learning applications. In this study, we present, for the first time, a newly developed, domain-specific diabetes research dataset, called Erbil Diabetes Dataset. The data were collected under the supervision of a medical professional at a laboratory in Erbil, Kurdistan Region of Iraq. The dataset contains twelve key characteristics which were captured from 662 people who visited the laboratory for check-ups. A standard procedure was employed to preprocess the features before presenting them to the diabetes research community for use. The performance evaluation of these algorithms on the specified dataset utilized five algorithms that included Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Xtreme Gradient Boost (XGBoost), and Decision Tree (DT). The evaluation process analyzed the performance results of algorithms through accuracy-recall measurements in addition to precision and F1-score metrics. The result shows KNN and XGBoost reaching outstanding performance values and predictive accuracy measures at 99.25% and 98.80% respectively. The accuracy levels of SVM decrease to 73.68% caused by their sensitivity to hyperparameter optimization. A statistical analysis using a one-way ANOVA test on F1-scores revealed a significant outcome ($F = 558.51$, $p < 0.001$), confirming that the differences in model performance were meaningful rather than due to random variation.

1.Introduction

The Artificial Intelligence (AI) together with its subsystem Machine Learning (ML) has delivered substantial enhancements to medical sciences and healthcare systems throughout the past several years. The AI classification task supports disease diagnosis by processing broad clinical datasets extensively (Polevikov, 2023, Esmaeilzadeh, 2024).

Diabetes represents one of the medical conditions positively affected by the advancements in AI and ML systems. The World Health Organization reports that 10 percent of global population suffers from diabetes ((WHO), 2023). Diabetes is a chronic metabolic illness that causes abnormal blood glucose control which stems from either inadequate insulin production or insensitivity to insulin by the body. Scientists remain uncertain about diabetes' exact origin yet causes environmental factors and lifestyles along with genetics seem may have some influences (Holt et al., 2017). Diabetes consists of three distinct forms which include Type-1 and Type-2 along with gestational diabetes. Type-1 diabetes occurs when the pancreas fails to produce insulin while Type-2 diabetes develops due to the body's cells becoming resistant to insulin and gestational diabetes appears during pregnancy and resolves after the birth of the child (Holt et al., 2017).

The successful care of diabetes along with prevention of diabetes-related complications relies on early and accurate identification of the condition (Çakmak and Özdemir, 2024). Traditional diagnostic techniques encompass assessments such as HbA1c, oral glucose tolerance, and fasting blood glucose. While the traditional diagnostic approach confirms the disease's existence, it is unable to identify future risk factors. This new strategy aims to improve the efficiency and accuracy of future forecasts and diagnosis by utilizing AI and Machine Learning (ML) techniques (Nissar et al., 2024).

In this paper, we present a new dataset collected from a local microbiology clinic in Erbil, Kurdistan Region of Iraq, referred to as the Erbil Diabetes Dataset (EDD) (Raheem et al., 2024). Our objective is to utilize machine learning algorithms and our dataset to build a model for identifying the patterns related to diabetes.

Twelve identified features were used in the analysis to predict diabetes based on clinical and demographic patient data. The selected features demonstrated both high relevance to diabetes risk factors and availability during routine clinical evaluations. These attributes include: ‘age’, residence type (city or village), ‘diastolic blood pressure (BP)’, ‘plasma glucose level (GL)’, ‘cholesterol (CHOL)’, ‘high-density lipoprotein (HDL)’, ‘low-density lipoprotein (LDL)’, ‘triglyceride’, ‘body mass index (BMI)’, and ‘family history’ (covering both first- and second-degree relatives). The HbA1C test result was used as the target variable for diagnosis, as it is a standard measurement for detecting diabetes in patients. The model features were strategically selected because they are medically relevant and commonly recorded in real clinical settings, making them practical for day-to-day healthcare use.

The structure of the paper is as follows: Section 2 reviews the related studies, section 3 outlines the collection and preprocessing of the dataset, section 4 describes the machine learning models we used and discusses their strong and weak points in Diabetes analysis using ML, section 5 presents the results and discussion of the findings and finally, the last section concludes the findings and offers suggestions for further research.

2. Related Works

Many supervised machine learning algorithms have been applied in diabetes prediction in research studies. Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT), Naive Bayes (NB), Logistic Regression (LR) and Random Forest (RF) applied to the PIMA Indians diabetes dataset by (Tigga and Garg, 2020). In their study, the Random Forest provided the maximum accuracy at 75%. In another study, Support Vector Machines, Decision Trees, and Naive Bayes algorithms have been tested, with SVM achieving the highest prediction accuracy at 80% (Vijayan and Anjali, 2015, Panwar et al., 2016). k-Nearest Neighbors applied in another study with accuracy of 85% (Vijayan and Anjali, 2015, Panwar et al., 2016). The identification of risk factors has been evaluated using artificial neural network, logistic regression, and decision

tree algorithms through existing research data from 1487 Guangzhou participants. Among the three algorithms decision tree demonstrated the highest success rate of 77.87% (Meng et al., 2013).

Any AI model's accuracy depends on the dataset quality since medical sector datasets contain problems with incomplete patient dossiers that create missing values and random patient clinic visits that generate class imbalances alongside domain expert labeling. The performance of algorithms alongside their accuracy levels change because of these characteristics (Gong et al., 2023). Researchers conducted evaluations of ML algorithms using a variety of dataset types in the literature. When performance was compared, the PIMA diabetes dataset showed that the DT algorithm worked better than KNN. It had an accuracy rate of 90.43% when tested on that dataset (Hashi et al. 2017). Further tests and research using data from a specialist hospital showed that random forest algorithms excel in this setting with 88.76% accuracy among several different tested algorithms such as logistic regression, random forest, KNN, SVM, NB, and gradient boosting (Muhammad et al., 2020). Another assessment of the SVM and NB and KNN and DT algorithms on adult population data showed that DT achieved higher accuracy among the algorithms (Hashi et al., 2017, Muhammad et al., 2020). In these applications of Machine Learning, only limited features have been used, while diabetics is a complex disease and the more features lead to more realistic models and improved accuracies.

An examination shows that the Decision Tree algorithm delivered the best classification accuracy for predicting diabetes among common risk factors, while SVM followed closely, then logistic regression followed by NB. Research performed on KNN, random forest and NB and DT classes shows both NB and random forest models outperforming others when dealing with datasets containing various features among multiple points (Varma and Panda, 2019, Pranto et al., 2020).

Multiple statistical measurements were applied to compare KNN, LR, NB, SVM, RF, and DT across the analysis of offline and online data and the PIMA dataset. Random forest achieved the best accuracy rate of 94.10% during these performance tests (Tigga and Garg, 2020). Another assessment framework for diabetes diagnosis employed decision trees with NB, SVM and logistic regression but demonstrated that decision trees achieved the highest accuracy rate of 79% (Tigga and Garg, 2020, Llaha and Rista, 2021).

This survey demonstrates the important role of collected data using local dataset information leads to improved diagnostic and forecasting capabilities for diabetes throughout different communities. The training of machine learning algorithms requires local data acquisition because variations in distinct health and demographic traits exist between different communities. The algorithms attain better accuracy and reliability through the use of data collection which represents the unique attributes of local communities. Medical prediction and diagnostic development heavily relies on data that specifically reflects context because it directly impacts the performance of machine learning models.

To contribute to both local and global diabetes health research efforts, this paper presents Erbil Diabetes Dataset, a new diabetes dataset from Erbil, Kurdistan Region of Iraq. Five well-known machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF), Extreme Gradient Boost (XGBoost) are used for the evaluation of diabetes prediction modeling based on local data observations and risk factors.

3. Data Collection and Preprocessing

Throughout history, medical research into diabetes discovery and treatment approaches has revealed that blood sugar levels and the likelihood of diabetes prevalence depend on multiple key factors. Age causes insulin sensitivity, which directly increases the risk of developing diabetes throughout adulthood. Changes in situation and a busy city environment will likely affect diabetes risk levels through environmental and lifestyle factors involving diet and physical activity. Blood pressure (BP) serves as an indicator of elevated diabetes susceptibility. The illness risk increases

when individuals have elevated levels of high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides, and cholesterol (CHOL). Family history, person with diabetic family members is more likely to develop the disease because of their genetics. Obesity and diabetes have a well-established relationship based on BMI levels because high BMI values indicate obesity which directly affects insulin resistance and causes adverse effects on general metabolic health (Barbieri et al., 2024, Fatah and Alkaki, 2021).

In this study, data were collected from a private Pharma Laboratory in Erbil, Kurdistan Region of Iraq. The data collection was carried out by an experienced, trained nurse under the supervision of a licensed expert physician, directly during patient visits, mostly on the advice of their physicians. Eight attributes were selected: age, city (indicating whether the person lives in a city or a village), diastolic blood pressure (BP), plasma glucose level (GL), cholesterol (CHOL), high-density lipoprotein (HDL), low-density lipoprotein (LDL), triglycerides, family history (first and second degree), and body mass index (BMI). The HbA1C test results were recorded as the target variable for the dataset. The dataset is called Erbil Diabetes Dataset (EDD) and publicly available at (Raheem et al., 2024).

The established dataset contains 662 records, with 420 males and 242 females, ensuring low bias in the dataset. Among these records, 105 were labeled as positive for diabetes, while 557 were labeled as negative (492 were pre-diabetes and 65 were classified as having no diabetes). The target labels have three classes: 1 for non-diabetic, 2 for pre-diabetic, and 3 for diabetic, as shown in Figure 1.

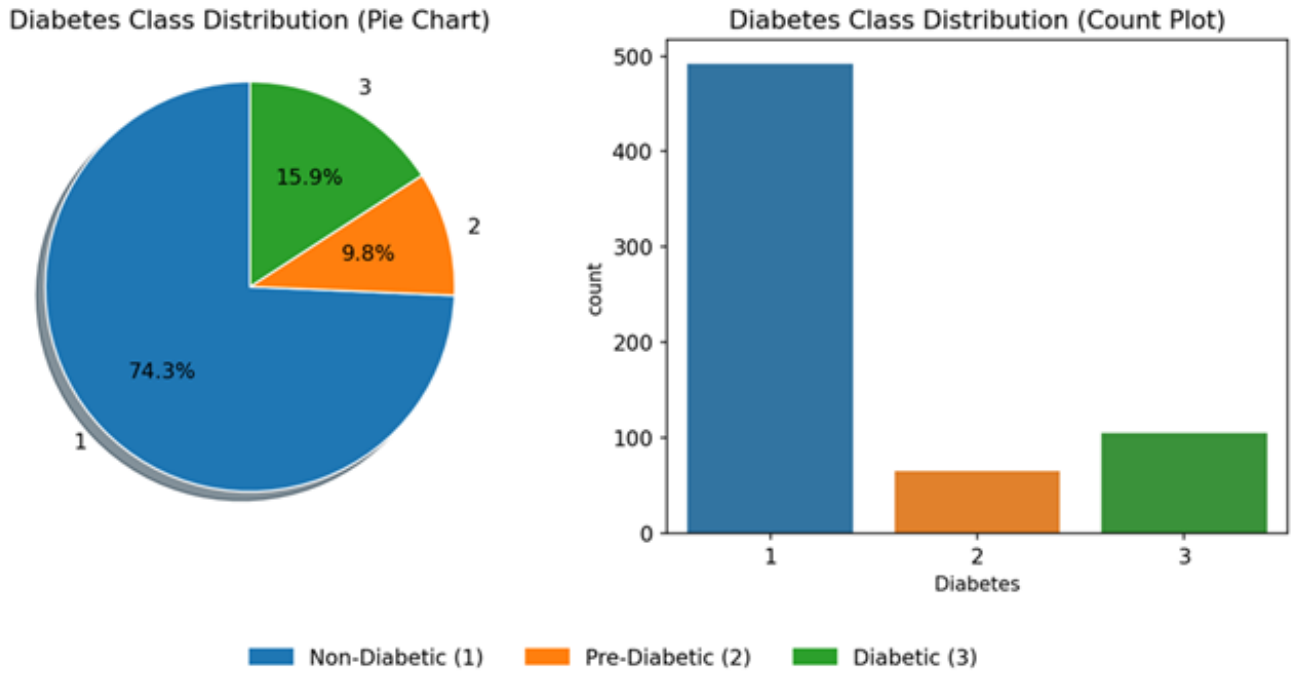

Figure 1: Ratio of Non-Diabetic, Pre-Diabetic, and Diabetic Patients (1: non-diabetic, 2: pre-diabetic, 3: diabetic)

The majority of healthcare-related datasets contain missing values and other contaminants that can affect their usefulness (Soni and Varma 2020). Our initial dataset also included categorical data and several missing items within records. The dataset underwent preprocessing, where categorical data were converted to numerical types, and records with missing attribute values were removed. The type of preprocessed is binary. A part of the dataset is shown in Table 1.

Table 1: Sample of the dataset

	No	Visit_ID	Systolic_BP	Diastolic_BP	Patient_Age	Sex	Cholesterol	Triglyceride	HDL	LDL	VLDL	Family	BMI	Hba1c	Diabetes
0	1	24002367.0	12	8.0	51	2.0	174.0	92.0	35.0	125.0	18.0	0.0	34.3	4.9	1.0
1	2	24002365.0	11	7.5	42	1.0	153.0	72.0	34.0	105.0	14.0	0.0	28.2	5.1	1.0
2	3	24002356.0	12	8.0	34	2.0	144.0	99.0	36.0	108.0	19.0	23.0	23.03	5.0	1.0
3	4	24002401.0	12	9.0	53	1.0	159.0	297.0	27.0	99.0	59.0	6.0	20.8	6.7	3.0
4	5	24002401.0	11	7.5	37	1.0	262.0	234.0	30.0	192.0	47.0	0.0	29.7	5.2	1.0

4. ML Model Development and Evaluation

In machine learning applications, the size and quality of the dataset, the hyperparameter tuning, the feature selection as well as the algorithm, affect the accuracy of the results. Default value Scikit-learn was used. To determine an algorithm with satisfactory accuracy, five ML classification algorithms were applied to the new dataset: K-Nearest Neighbors (KNN), Random Forest, Decision Tree, Support Vector Machine (SVM) and XGBoost. The results were evaluated using accuracy, recall, precision, and F1-score metrics.

4.1 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a fundamental algorithm in machine learning. The algorithm operates within the instance-based learning framework of lazy learning by deploying local approximation and delaying computational operations until evaluation time. This algorithm classifies a data point based on the classification of its neighbors. KNN uses two steps to classify data points through neighbor proximity by measuring the distance to K closest neighbors followed by assigning the prevalent class of these neighbors. In our study, the Euclidean distance is used for this purpose (see Equation 1) (Katarya and Jain, 2020, Salh and Ali, 2022):

$$D(x,y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2} \quad (1)$$

We selected KNN for our dataset because of its non-parametric nature, meaning it makes no explicit assumptions about the form of the mapping function. This is particularly useful for medical data, where the relationships between features and outcomes can be complex and nonlinear. The algorithm is illustrated in Algorithm 1. (Murphy, 2012, Bishop, 2006, Mohammed and Yousif, 2019)

Algorithm 1: K-Nearest Neighbors

1. Select the value of "K" (number of neighbors)
2. Select the distance metric (e.g., Euclidean, Manhattan)
3. Get the training data with features and set labels for it
4. KNN Model Initialization with training data
5. $i \leftarrow 1$
6. While each Remaining in (Instances To Classify do):
    7. $x\_test \leftarrow$ *Get Next Instance To Classify ()*
    8. distances = []
    9. for each instance (x, y) in Training Data:
        10. distance = Compute Distance (x, x_test*, Metric)*
        11.Append (distance, y) to distances
    12. end for
    13. Order distances by increasing distance
    14. Nearest Neighbors = Select Top K (distances, k)
    15. Labels = Extract Labels (Nearest Neighbors)
    16. Predicted Label = Most Common Label (Labels)
    17. Assign Predicted Label to x_test
    18. $i \leftarrow i + 1$
End While
19. Return KNN Model with training data

4.2 Random Forest

Random Forest (RF) serves as an ensemble learning algorithm which combines several

decision trees for enhanced classification accuracy performance and control over-fitting. It constructs a multitude of decision trees throughout training which use a class prediction from individual trees to determine the output class making the selection.

We selected Random Forest for the diabetes prediction problem because it effectively handles data imbalances by generating multiple decision trees from different subsets of the data and features, ensuring that minority classes are adequately represented. In the other side, random forest provides insights into feature importance, which is important in medical applications for understanding which factors are most influential in predicting diabetes. The RF algorithm is shown in Algorithm 2 (Pranto et al., 2020).

Algorithm 2: Random Forest

1. Select T: number of trees to be grown.
2. Select m: number of attributes should be used in each tree's learning
3. $i \leftarrow 1$
4. while $i \leq T$ *do*
5. Use random sampling with replacement to extract a training subset from the available training set.
6. Randomly select m features to create a feature subset
7. Train a decision tree which utilizes the chosen features from the training sample subset.
8. $i \leftarrow i + 1$
end
9. The system predicts outcomes by employing majority vote across all generated trees.

4.3 Decision Tree

The Decision Tree (DT) algorithm splits data into subsets through input feature value analysis to construct a tree structure-like model of decisions. The structural design of DT features internal nodes as input elements while branches show chosen rules that lead to final solution clusters at leaf nodes. The algorithm uses the entropy equation (see Equation 2) to measure the impurity of the data and determine the best feature splits (Singh et al., 2021)

$$\text{Entropy}\big(\text{E(s)}\big) = -\sum_{j=1}^{k} P_j * (\log_2 (P_j)) \quad (2)$$

There are several reasons behind the inclusion of Decision Tree in our diabetes prediction study. Firstly, tree structure enables exceptional visual understanding of decision pathways so physicians can explain and be transparent when making choices. Clinicians are able to comprehend how forecasts are made and follow the decision-making process with ease. Second, Decision Trees do not assume any particular distribution for the data because they are a non-parametric technique. The model embraces flexibility to analyze intricate relationships that exist between features and outcomes without data preparation needs. Finally, by calculation of impurity reduction at each split provides Decision Trees with valuable information regarding feature importance. This capacity facilitates to the study of important risk factors and their effects on diabetes by assisting in the identification of the most influential factors in diabetes prediction. The DT algorithm is shown in algorithm 3 (Alpaydin, 2020, Bishop, 2006).

Algorithm 3: Decision Tree

1. Select stopping criteria (maximum depth, minimum samples to split)
2. Prepare training set
3. Root Node Initialization
4. $i \leftarrow 1$
5. While not stopping criteria are met do:
6. Best Feature, Best Threshold ← Find Best Split (Data, Labels)
7. If Best Feature is None then:
8. Return Leaf Node with most common label
9. Left Data, Left Labels, Right Data, Right Labels ← Split Data (Data, Labels, Best Feature, Best Threshold)
10. Left Child ← Build Decision Tree (Left Data, Left Labels, Stopping Criteria)
11. Right Child ← Build Decision Tree (Right Data, Right Labels, Stopping Criteria)
12. Create Decision Tree Node with Best Feature, Best Threshold, Left Child, Right Child
13. $i \leftarrow i + 1$
14. Return Decision Tree

4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) functions as a supervised learning algorithm that medical research commonly uses. SVM identifies the optimal hyperplane which maximizes data separation among different classes within high-dimensional space. SVM achieves its objective by enlarging the gap between opposing data points through support vector optimization which produces solid predictions for previously unknown data.

The selection of SVM for our study is motivated by its high accuracy, especially in high-dimensional spaces, and its robustness to overfitting and outliers. The kernel trick allows SVM to perform classification in a higher-dimensional feature space without explicit coordinate computation of the data within the space. This flexibility is advantageous when dealing with non-linearly separable data, common in medical diagnostics. The algorithm is illustrated in Algorithm 4 (Murphy, 2012, Abe, 2005).

Algorithm 4: Support Vector Machine (SVM)

1. Select the C regularization parameter
2. Select kernel function (e.g., linear, polynomial etc.)
3. Build the training dataset with features and labels
4. Initialization weights w and bias b
5. $i \leftarrow 1$
6. While Not Converged do:
7. for every (x, y) in Training Data:
8. $Decision\ Value = w * Kernel(x) + b$
9. if $y * Decision\ Value < 1$ *then:*
10. Minimizing a Hinge Loss by updating learnable weights and bias:
11. $w = w - \eta * (w - C * y * x)$
12. $b = b + \eta * C * y$
13. end if
end for
13. Check convergence criteria
14. $i \leftarrow i + 1$
End While
15. Return (w, b)
16. function Predict SVM (x_test*, w, b, Kernel):*
17. $Decision\ Value = w * Kernel(x\_test) + b$
18. if $Decision\ Value > 0$:
19. return +1
20. else:
21. return -1

4.5 Extreme Gradient Boosting (XGBoost)

XGBoost represents a high-performance ensemble learning method that uses the gradient boosting framework for its operation. XGBoost creates decision trees in a sequential order that focuses on fixing prediction mistakes of previous trees. XGBoost achieves high-performance results with the application of regularization techniques which combine with shrinkage and column subsampling methods and efficient handling of missing values. The built-in enhancements make the model highly effective on structured data while preventing overfitting to deliver better results for unseen instances. in this paper XGBoost implemented for predicting diabetes by tuning its hyperparameters to assess its capability in processing clinical and demographic characteristics. XGBoost delivers widespread application for classification problems and especially healthcare applications because of its scalable design and accurate predictions. The XGBoost algorithm is shown in algorithm 5 (Chen and Guestrin, 2016).

Algorithm 5: Extreme Gradient Boosting

1. Select T: the number of trees to be grown.
2. Select regularization parameters (λ and γ), learning rate η (eta), and other hyperparameters.
3. The first prediction should use the target variable mean as an initial starting point.
4. $i \leftarrow 1$ *// Counter for the number of trees*
5. While $i \leq T$ *do*
6. Compute the residuals (errors) by comparing the current predictions to the actual target values.
7. Construct a decision tree to predict the residuals from the current predictions.
8. Regularization techniques (L1, L2) should be used to stop overfitting and enhance general performance.
9. The model update requires the new tree predictions weighted with learning rate η to be added.

10. $i \leftarrow i + 1$ // Increment the tree counter
12. End while
13. For each new data point, predict the outcome by summing the predictions from all trees.
End

5. Results and Discussion

The Python library scikit-learn was used to develop these algorithms. Initially, the dataset was split into 80% for training and 20% for testing. This method guarantees that the models are trained on a substantial portion of the data while still reserving a significant amount for evaluating their performance.

Different methods were employed to reduce class bias because the dataset contains 105 positive instances when compared to 557 negative instances. All models (SVM, Random Forest, Decision Tree, and XGBoost) employed cost-sensitive learning since this approach applied stricter prediction penalties targeted at minimizing mistakes for minority class examples. The KNN-based model delivered better generalization capabilities by creating synthetic examples of minority class through Synthetic Minority Over-sampling Technique (SMOTE).

Using (ML) in health care system assist doctors for accurate diagnosis disease. Support vector machine, K-nearest neighbor, XGBoost, random forests, and decision tree algorithms have been used to build the models. The performance of the tests was evaluated using accuracy, recall, precision, F1-score. The following hypothesis are applied:

True Positives (Tp): a person who has diabetes and was expected to have the disease.
False Positives (Fp): a person who does not have diabetes but was expected to have the disease.
False Negatives (Fn): a person who has diabetes and was not expected to have the disease.
True Negatives (Tn): a person who does not have diabetes and was not expected to have the disease.

The definitions for the evaluation metrics are as follows (Varoquaux and Colliot, 2023):
Accuracy is the percentage of correct (i.e true) predictions metrics which is a major index for ML model evaluation, see equation 3 and 4.

$$Accuracy\ (Acc) = \frac{\sum Correct\ predictions\ number}{\sum Total\ Predictions\ number} \quad (3)$$

or,

$$Accuracy\ (Acc) = (Tp + Tn)/N \quad (4)$$

Where N is the total number of predictions.

Precision is the ratio of accurately anticipated positive observations to all positively expected observations, see equation 5 and 6.

$$\text{Precision}(P) = \frac{True\ positive}{\sum Postive\ predictions\ number} \quad (5)$$

$$\text{Precision}(P) = Tp/Tp + Fp \quad (6)$$

Recall measures the ability of a ML model to correctly identify all relevant instances in the dataset. Mathematically, it is the ratio of correctly predicted positive observations to all observations in the actual positive class, equation 7 and 8.

$$\text{Recall}(R) = \frac{True\ positive}{True\ positive + False\ negative} \quad (7)$$

$$\text{Recall}(R) = Tp/Tp + Fn \quad (8)$$

The F1-score is another metric which have been used in our evaluation. It is the weighted average of recall and precision, equation 9 and 10. This index provides a balanced measure of a model's performance, especially where there is an uneven class distribution.

$$\text{F1 score} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (9)$$

$$\text{F1 score} = (2 * (P * R))/(P + R) \quad (10)$$

Figure 2 presents the Pearson correlation heatmap for the features in the diabetes dataset. The data reveals that HbA1c measurement has the highest connection to diabetic conditions (r = 0.82) because diabetic patients consistently show elevated HbA1c statistics. The association between older patient age and diabetes risk appears moderate with a correlation value of r = 0.35 which indicates seniors tend more often to develop diabetes. The lipid profile variables of Cholesterol share a significant correlation with LDL (r = 0.88) while VLDL exhibits a strong

relationship with Triglyceride (r = 0.89) due to their natural physiological associations. The identified relationships between features assist in showing important risk factors which aid model feature selection procedures. The weak relationships between the target variable Diabetes and most other features indicates that machine learning algorithms might benefit from complex patterns and interactions which makes them suitable for this prediction task.

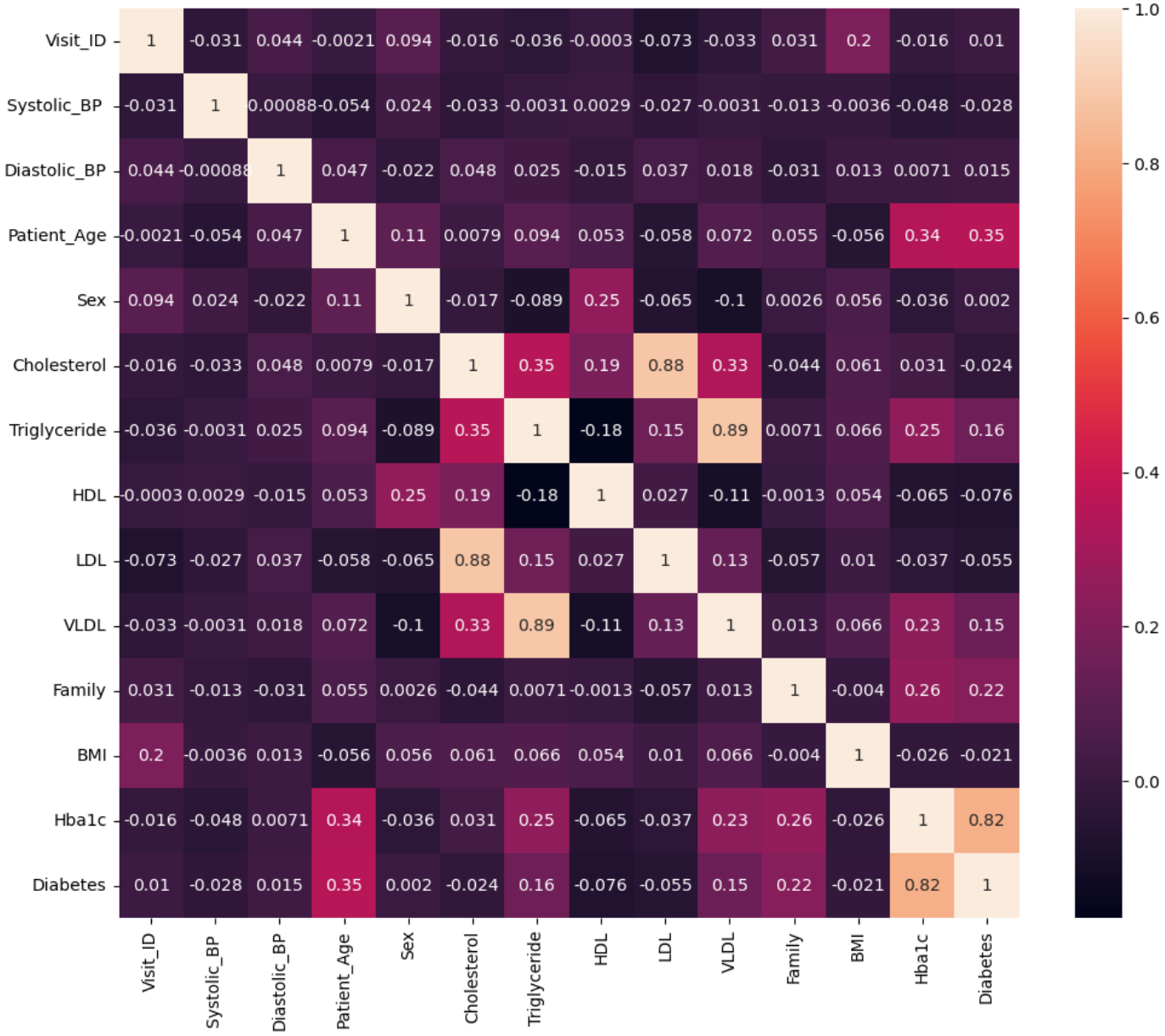

Figure 2: Pearson correlation heatmap of the features in the diabetes dataset.

Table 2 presents the results of all models with their evaluation metrics. The table is converted into a diagram of Figure 3 for better visualization. Table 2 shows that the KNN classifier produces better results than other classifiers for predicting diabetes mellitus. According to Table 2 and Figure 3, KNN achieves 100% precision, 99% recall, and an F1-score of 99% on this dataset. The accuracy of KNN is 99.25% and XGBoost is 98.80%. Followed by the Random Forest model with an accuracy of 96.99%, the Decision Tree model with 96.24%, and finally, the Support Vector Machine model with 73.68% accuracy. These experimental results demonstrate that KNN is effective in predicting diabetes mellitus from medical datasets using various risk factors.

Table 2: Performance evaluation result of the model

Algorithms	Accuracy (%)	Precision	Recall	F1-score
KNN	99.25%	1	0.99	0.99
Random Forest	96.99%	0.94	0.91	0.92
Decision Tree	96.24%	0.93	0.94	0.93
SVM	73.68%	0.55	0.74	0.64
XGBoost	98.80%	0.99	0.98	0.98

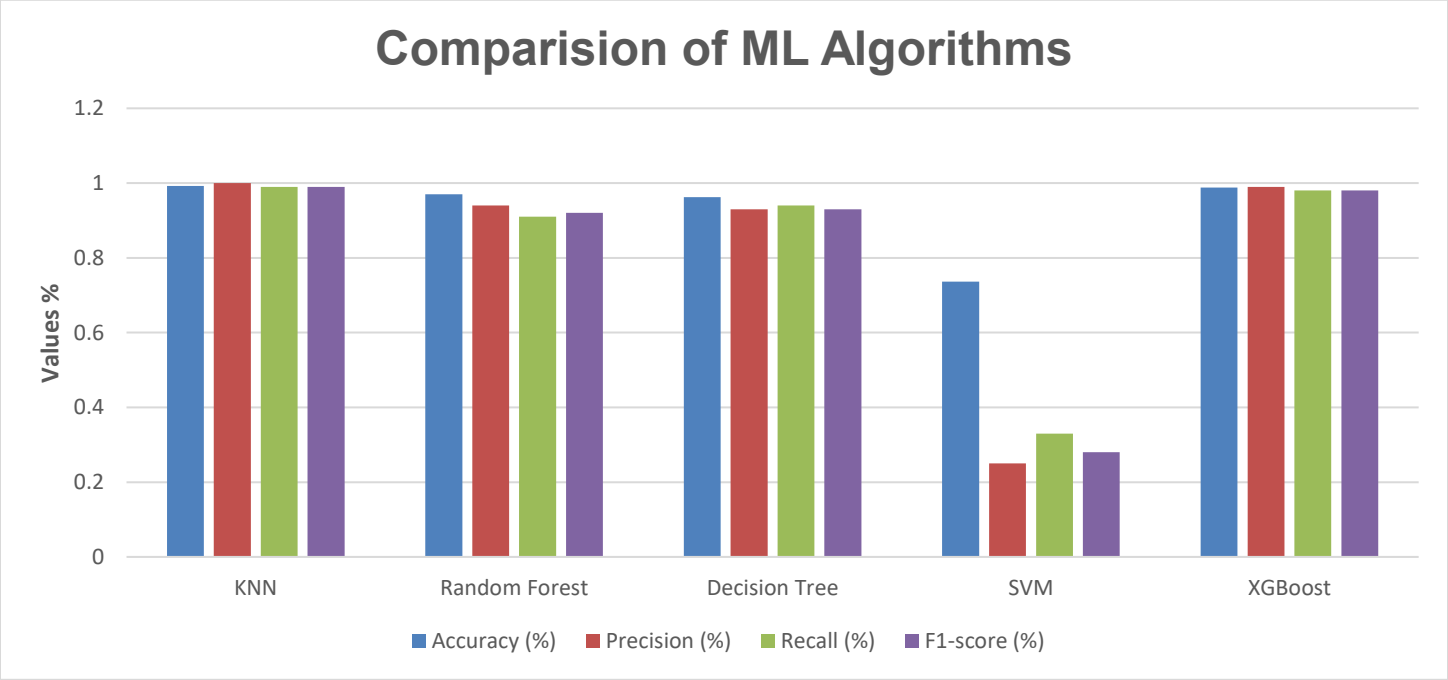

Figure 3: Comparative analysis of machine learning techniques

To further evaluate the statistical significance regarding performance differences between models, a one-way ANOVA testing approach evaluated the statistical importance of performance variability between all five tested algorithms based on their F1-scores. Table 3 shows the result of a one-way ANOVA analysis of F1-scores performed on five machine learning algorithms. Statistical analysis reveals that there is a significant difference in model performance because the F-statistic has a value of 558.51 with a p-value below 0.001. The KNN and XGBoost achieved the greatest F1-scores that secured their position as top performers for this assignment. Statistical analysis confirmed the low F1-score of the SVM model which indicates its relatively weaker performance in the task. The importance of selecting perfect machine learning algorithms for diabetes prediction becomes clear because minor performance distinctions affect model effectiveness.

Table 3: One-way ANOVA result on F1-scores of ML Algorithms

Source	F-value	p-value	Conclusion
Model Comparison	558.51	< 0.001	Significant difference

6. Conclusion

In this study, we established a new dataset for diabetes, which we have named Erbil Diabetes Dataset (EDD). The data has been collected in a private lab in Erbil, Kurdistan Region of Iraq. It contained records with 662 participants with twelve attributes. The attributes have been selected and labeled with the help of an expert physician. Five Machine Learning algorithms have been tested to model for the prediction, named KNN, SVM, DT, RF, and XGBoost. The performance of the algorithms have been compared through the evaluation metrics accuracy, precision, recall, and f1-score.

The results demonstrated that KNN is the most accurate model for predicting diabetes, with an accuracy rate of 99.25%. The XGBoost algorithm has the second-highest accuracy of 98.80%. The RF algorithm has the third highest accuracy of 96.99%, followed by the DT with an accuracy of 96.24%, and lastly the SVM-based model with an accuracy of 73.68%.

KNN is a non-parametric algorithm, works well with datasets where the boundaries between classes are nonlinear or where the relationship between features and the target variable is complex. The Random Forest and Decision Tree models exhibit a high degree of accuracy (96.99% and 96.24%, respectively), indicating that ensemble methods and tree-based models are suitable for handling the dataset since they can capture interactions and non-linear correlations between attributes. However, the low accuracy of the Support Vector Machine (SVM) with an accuracy of 73.68%, can be attributed to its sensitivity to the selection of kernel and hyperparameters, which may mean that it is less able than the other models to handle class imbalances.

Building a dataset for other locations in Iraq is recommended to show a good generalization of the module of Iraq. Furthermore, one-way ANOVA test function was used to evaluate statistically the differences between the predictive outcomes of all five models through their F1-scores. A statistical analysis of the models showed a significant performance variation ($F = 558.51$, $p < 0.001$) which indicates the observed performance differences cannot result from randomness. Statistical tests confirm KNN and XGBoost provide the best options for this classification problem because they outperform SVM in the analysis. ANOVA integration brings statistical foundation strength to the model selection process which highlights the general importance of testing statistics for medical prediction problems using machine learning algorithms.

As for future work, the fields of diabetes and AI should continue to be researched. To attain 100% accuracy in diabetes prediction at a reasonable cost, various combinations of machine learning algorithms and hyperparameter tuning are needed, in addition to new, massive datasets from around the globe to improve modeling performance.

Acknowledgment: Not applicable.

Financial support: No financial support.

Potential conflicts of interest. All authors report no conflicts of interest relevant to this article.

References

(WHO), W. H. O. 2023. Diabetes [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed 18 July 2024].

ABE, S. 2005. Support vector machines for pattern classification (Vol. 2, p. 4). London: Springer.

ALPAYDIN, E. 2020. Introduction to machine learning, MIT press.

BARBIERI, M., PRATTICHIZZO, F., LA GROTTA, R., MATACCHIONE, G., SCISCIOLA, L., FONTANELLA, R. A., TORTORELLA, G., BENEDETTI, R., CARAFA, V. & MARFELLA, R. 2024. Is it time to revise the fighting strategy toward Type 2 Diabetes? Sex and Pollution as New Risk Factors. Ageing Research Reviews, p.102405.

BISHOP, C. M. 2006. Pattern recognition and machine learning by Christopher M. Bishop, Springer Science+ Business Media, LLC.

ÇAKMAK, V. S. & ÖZDEMIR, S. Ç. 2024. Patients with diabetic foot ulcers: A qualitative study of patient knowledge, experience, and encountered obstacles. Journal of Tissue Viability, 33(4), pp.571-578.

CHEN, T. & GUESTRIN, C. 2016. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

ESMAEILZADEH, P. 2024. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. Artificial Intelligence in Medicine, 151, p.102861.

FATAH, K. S. & ALKAKI, Z. R. A. 2021. Application of Binary Logistic Regression Model to Cancer Patients: a case study of data from Erbil in Kurdistan region of Iraq. Zanco

Journal of Pure and Applied Sciences, 33(4), pp.117-128.
GONG, Y., LIU, G., XUE, Y., LI, R. & MENG, L. 2023. A survey on dataset quality in machine learning. Information and Software Technology, 162, p.107268.
HASHI, E. K., ZAMAN, M. S. U. & HASAN, M. R. 2017. An expert clinical decision support system to predict disease using classification techniques. International conference on electrical, computer and communication engineering (ECCE) (pp. 396-400). IEEE.
HOLT, R., COCKRAM, C., FLYVBJERG, A. & GOLDSTEIN, B. 2017. Textbook of Diabetes-Preface to the Fifth Edition. Textbook of Diabetes, 5th Edition (pp. xiv-xv). Wiley-Blackwell.
KATARYA, R. & JAIN, S. 2020. Comparison of different machine learning models for diabetes detection. IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE) (pp. 1-5). IEEE.
LLAHA, O. & RISTA, A. 2021. Prediction and Detection of Diabetes using Machine Learning. RTA-CSIT (pp. 94-102).
MENG, X.-H., HUANG, Y.-X., RAO, D.-P., ZHANG, Q. & LIU, Q. 2013. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), pp.93-99.
MOHAMMED, B. & YOUSIF, R. Z. 2019. Intelligent system for screening diabetic retinopathy by using neutrosophic and statistical fundus image features. ZANCO Journal of Pure and Applied Sciences, 31, pp.30-39.
MUHAMMAD, L., ALGEHYNE, E. A. & USMAN, S. S. 2020. Predictive supervised machine learning models for diabetes mellitus. SN Computer Science, 1(5), p.240.
MURPHY, K. P. 2012. Machine learning: a probabilistic perspective, MIT press.
NISSAR, I., MIR, W. A., SHAIKH, T. A., AREEN, T., KASHIF, M., KHIANI, S. & HUSSAIN, A. 2024. An Intelligent Healthcare System for Automated Diabetes Diagnosis and Prediction using Machine Learning. Procedia Computer Science, 235, pp.2476-2485.
PANWAR, M., ACHARYYA, A., SHAFIK, R. A. & BISWAS, D. 2016. K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus. sixth international symposium on embedded computing and system design (ISED) (pp. 132-136). IEEE.
POLEVIKOV, S. 2023. Advancing AI in healthcare: a comprehensive review of best practices. Clinica Chimica Acta, 548, p.117519.
PRANTO, B., MEHNAZ, S. M., MAHID, E. B., SADMAN, I. M., RAHMAN, A. & MOMEN, S. 2020. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information, 11(8): 374.
RAHEEM, S. A., TAHA, A. & HAMARASH, I. I. 2024. Erbil Diabetes Dataset. V1 ed. Mendeley Data.
SALH, C. H. & ALI, A. M. 2022. Comprehensive study for breast cancer using deep learning and traditional machine learning. Zanco Journal of Pure and Applied Sciences, 34(2), pp.22-36.
SINGH, A., DHILLON, A., KUMAR, N., HOSSAIN, M. S., MUHAMMAD, G. & KUMAR, M. 2021. eDiaPredict: an ensemble-based framework for diabetes prediction. ACM Transactions on Multimidia Computing Communications and Applications, 17(2s), pp.1-26.
TIGGA, N. P. & GARG, S. 2020. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167, pp.706-716.
VARMA, K. M. & PANDA, B. 2019. Comparative analysis of predicting diabetes using machine learning techniques. J. Emerg. Technol. Innov. Res, 6(6), pp.522-530.
VAROQUAUX, G. & COLLIOT, O. 2023. Evaluating machine learning models and their diagnostic value. Machine learning for brain disorders (pp. 601-630).
VIJAYAN, V. V. & ANJALI, C. 2015. Prediction and diagnosis of diabetes mellitus—A machine learning approach. IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.