VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2026, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 14, Number 1, April-June 2026 pp:46-61

Diabetes Prediction using Machine Learning Algorithms with Performance metrics and Holdout method on Egyptian dataset

Saida O. Said [1]*, **Nurul Liyana Binti Mohamad Zulkufli** [1], **Asmarani Binti Ahmad Puzi** [1], **Asadullah Shah** [1]

[1]Department of Computer Science, International Islamic University Malaysia, Kuala Lumpur, Malaysia, Department of Information System, International Islamic University Malaysia, Kuala Lumpur, Malaysia

Keywords: *Diabetes, Machine learning algorithms, Holdout Method, Performance Metrics.*

Journal Info:
Submitted:
March 25, 2026
Accepted:
April 05, 2026
Published:
April 22, 2026

Abstract Diabetes is a non-communicable disease affecting people of all ages worldwide, therefore, early detection using machine learning techniques is crucial. This study aims to predict diabetes using multiple machine learning algorithms, performance metrics, and holdout validation on an Egyptian dataset. The dataset was divided into four age groups, including paediatric, early adulthood, middle age, and geriatric. Ten algorithms were applied and validated using 80:20, 70:30, and 60:40 split ratios with accuracy, precision, and recall as evaluation metrics. Results showed that Random Forest, Extra Trees, and Support Vector Machine performed best in the paediatric group, while Gradient Boosting, Random Forest, and Support Vector Machine achieved superior performance in early adulthood, middle age, and geriatric groups. In contrast, Decision Tree, K-Nearest Neighbors, and AdaBoost consistently demonstrated lower performance. Further analysis reveals that classification performance varies significantly across age groups, with the middle age and geriatric groups achieving the highest accuracy above 0.99, followed by the paediatric group 0.98–0.99, while early adulthood exhibits comparatively lower performance due to increased class overlap. Confusion matrix results indicate strong diagonal dominance in higher-performing groups, reflecting better class separability, whereas performance heatmaps confirm that top models maintain a balanced trade-off between accuracy, precision, and recall with minimal variation across different data splits. Feature importance analysis shows that higher performing models rely on a small number of dominant predictors, particularly in the middle age and geriatric groups, while more distributed feature contributions in early adulthood reduce predictive effectiveness. Therefore, the findings demonstrate that ensemble methods provide robust and consistent performance, and that age-based dataset segmentation enhances classification accuracy and model stability.

***Correspondence author email address:** saida.omar@live.iium.edu.my
DOI: 10.21015/vtse.v14i2.2375

1 INTRODUCTION

Globally, diabetes is among the non-communicable chronic diseases that affect the body to not regulate its blood glucose effectively. According to [1] approximately 463 million of people worldwide have been affected with diabetes and [2] has recorded that, about 1.5 million of death are caused by diabetes each year whilst the pediatrics age has been increase from 2 to 5

percent. However, in the Egyptian context, it had been shown that to increase its prevalence of diabetes for the Middle East and Africa. Moreover, the adults who were diagnosed with diabetes were around 8.9 million and the number is suggested to rise to 13.1 million by 2035. Hence, Globally, Egypt has ranked in top 10 in diabetes prevalence, as the study of [3] has mentioned that from the global list, the Egypt has placed as 8th position. There are different classifications of diabetes, [4] and [5] explained these two types of classification which include of Type 1 and Type 2 diabetes. These studies had explained that the Type 1 diabetes occurs due to extreme low or due to the absence of insulin. Whilst for the paediatric age, there are little signs and symptoms that start with the children of diabetes and later on increase those symptoms. However, Type 2 diabetes is caused due to the imbalance between insulin level and sensitivity which leads to insulin resistance, and it occurs to adults only.

Predicting diabetes can be easily done using machine learning algorithms and it can be categorized as supervised machine learning, unsupervised machine learning and reinforcement machine learning. The supervised machine learning is used to train the labelled data and the algorithms that are mostly used such as Naïve bayes, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, Logistic Regression and Decision Trees [6]. For the Unsupervised machine learning is used to train the unlabeled data that includes K-Means clustering [7]. While reinforcement machine learning is more effective in requiring sequential decision-making, such as robotics and game playing as suggested to [8] and [9].

Recent studies have demonstrated the effectiveness of machine learning techniques in diabetes prediction across different populations and age groups. Study [10] investigated AI-based diabetes prediction among older adults in South Korea using Extreme Gradient Boosting, Light Gradient Boosting, Random Forest, and Gradient Boosting with a 70:30 holdout split. Among these models, Extreme Gradient Boosting achieved the best performance, with an accuracy of 0.8488, precision of 0.7792, recall of 0.6691, an F1-score of 0.7200, and an AUC of 0.7957. Although the study applies machine learning techniques to predict diabetes using health data collected via a mobile application, and leverages advanced algorithms and interpretability techniques, it is limited to a specific population — individuals aged 60 years in Seoul — which restricts the generalizability of the findings to broader and more diverse populations. The non-stratified 70:30 data split may also influence model reliability and generalizability.

Moreover, the study targets only one age group without comparative analysis based on age thus limiting its clinical applicability in other stages of life. In a similar manner, study [11] investigated the development of diabetes prediction based on e-health records with supervised machine learning algorithms, such as Logistic Regression, Random Forest and K-Nearest Neighbors, and a 70:30 holdout split. K-Nearest Neighbors (KNN) was the most effective of these models, and its accuracy was 0.9609, sensitivity 0.9854 and specificity was 0.9363. The paper employs a large open dataset and performance is well predicted, especially using KNN. Nevertheless, it does not focus on structured demographic segmentation but mostly concentrates on the overall model accuracy and performance measures.

Though the sample is quite large, the age range is very great, so it is considered a homogenous population, which does not explicitly analyze age groups and restricts the possibility to identify age-specific disease dynamics. Also, the dataset is region-based (USA), which limits its use to more heterogenous populations. Contrary, the current paper includes age-group segmentation consistency and more detailed evaluation strategies, which allow gaining deeper clinical insights and becoming more personalized in predicting diabetes. Furthermore, research article [12] aimed at predicting gestational diabetes mellitus (GDM) in the first trimester with cross-validation of five-fold. Various machine learning models have been tested, such as Decision Tree, Multilayer Perceptron, K-Nearest Neighbors, Naive Bayes, Random Forest, and Extreme Gradient Boosting; however, the best performance was shown by the Random Forest.

The research uses these models to hospital based Iranian data and introduces clinically important characteristics, which suggest that machine learning is effective in predicting early GDM. The study, however, is only limited to a group of pregnant women, limiting its gener-

alizability to the prediction of diabetes in general populations. Moreover, the data is institution-based and region specific which may be restrictive to generalizability. The main emphasis of the study is on predictive performance with little emphasis on comparative demographic segmentation other than the targeted group. Data preprocessing is also stated, but the level of transparency is lacking as to how the preprocessing is done, and how it affects model performance.

In another study, Type 2 diabetes screening of an Iranian population was done using seven machine learning algorithms with an 80:20 holdout ratio with the Gradient Boosting Machine performing the best. [13]. The work uses various machine learning methods on the data of the Fasa Adult Cohort Study in Iran and has several advantages in the relatively large data size, as well as sex-specific analysis of men and women groups and the proper preprocessing steps. It is however restricted to adult population (35-70 years) only and does not cover pediatrics and broader age groups. Despite the presence of gender based stratification, the study does not delve into detailed age based stratification, thus limiting the information on various stages of life. Also, the five-fold cross-validation and one 80:20 split might not be robust enough to evaluate the models.

The analysis mainly revolves around predictive performance with little focus on the wider clinical interpretation. In addition, study [14] suggested a feature selection-based model with the use of the Random Forest algorithm on the Pima data, and classification with Support Vector Machine (SVM) with an accuracy of 0.8009. The experiment is concerned with the improvement of Type 2 diabetes prediction by making good use of features alone and shows the significance of interpretability in the enhancement of the model performance. It is however constrained by use of only one public dataset that is Pima having relatively small number of attributes besides, the study is restricted to binary classification of Type 2 diabetes only and does not take into account other types of diabetes and demographic diversity like age groups.

Despite employing preprocessing and feature selection methods, the method is mostly algorithm-driven, and few attempts are made to explore subgroup-specific trends or clinical segmentation. Also, research article [15] analyzed the prediction of diabetes onset in adults in Taiwan based on 80:20 holdout split where decision jungle and boosted Decision Tree models performed best. The experiment utilizes the machine learning method on outpatient examination data of a Taipei Municipal hospital environment, which includes a sample of about 15,000 female patients aged 20 to 80 years old. The dataset includes eight clinical features, such as pregnancies, glucose level, blood pressure, insulin, BMI, and age, collected over the periods 2018–2020 and 2021–2022. The study has several advantages such as a fairly large dataset and an examination of a range of models with good predictive capabilities; however, it is confined to a demographic group of female patients only and thus limiting the generalizability.

Moreover, even though age is an attribute, the data is assumed to be a homogeneous population where no direct division by age-group is taken, which restricts the capacity to identify age-specific trends. Study [16] in Bangladesh used machine learning methods and feature selection to determine risk factors of diabetes with an accuracy of more than 0.822 and AUC of 0.872, and glucose was found to be the most important predictor. This research applies the machine learning algorithms and feature selection techniques such as the Principal Component Analysis (PCA) and the information gain on the survey data.

It combines both clinical and non-clinical attributes and implements the right preprocessing procedures, including deletion of outliers, missing values, and normalization. Nevertheless, it has certain shortcomings due to the rather limited number of 738 cases of the research, which can have an impact on the generalizability of the results. Moreover, despite the comprehensive age distribution of the dataset, it is not analyzed as a stratified population, and the age groups are not explicitly divided, which restricts the information about age-specific trends. The research is also limited to binary classification diabetic versus non-diabetic and is more about determining noteworthy features as opposed to providing in-depth sub-group or comparative analysis.

Study [17] in the Egyptian context compared the performance of WEKA and Google Colab in predicting Type 1 and Type 2 diabetes in relation to various age groups with the aid of six machine learning algorithms and cross-validation. These findings revealed that Google Colab

was better at all ages and the accuracy of Random Forest was the highest. The analysis is conducted on an Egyptian diabetes dataset to assess the performance of the machine learning based on its main measures which include accuracy, precision and recall. In spite of its use of age-based grouping, it can only apply it to adult groups and not that of pediatrics, hence limiting the complete age-spectrum information. The research also uses only the 10-fold cross-validation and mainly focuses on the comparison of the tools instead of the clinical interpretation of results with minimal explanations regarding the effects of preprocessing and in-depth subgroup analysis.

In addition, the research is binary classification of Type 1 and Type 2 diabetes with no wider classification. Lastly, research study 18 tested diabetes prediction on two datasets of varying sizes, and age group composition, and utilized 10-fold cross-validation. The findings proved that the performance of algorithms can greatly differ according to the characteristics of the dataset used, and Naive Bayes and Random Forest present good overall results.

The researchers implement several machine learning methods on two publicly available data sets but only one of these data sets includes the age-based segmentation, which creates an indication of inconsistency in the research. Despite the evaluation of multiple algorithms, the datasets include few and rather homogeneous attributes, limiting the possibility to obtain detailed age-specific insights. Moreover, there are no well-established preprocessing steps to use with WEKA, making it less transparent and reproducible. The analysis is still more inclined towards predictive performance, and less on clinical interpretation.

The majority of the available studies [10],[18] show high accuracy in diabetes prediction, but they have multiple limitations, such as the use of population-specific data, inability to perform a full age-group segmentation, and generalizability. Most research assumes the heterogeneous population is homogeneous despite big age differences and uses narrow validation methods and does not provide enough detailed preprocessing, which influences the reliability.

Moreover, no significant focus on the subgroup analysis of the various population groups is made. Therefore, there is a need for a more comprehensive approach that incorporates structured age-group segmentation, diverse datasets, and robust evaluation approach, including multiple validation strategies, to produce more generalizable and meaningful insights into diabetes prediction.

2 MATERIAL AND METHODS

The proposed approach used in this study is shown in Figure 1. The experiment was conducted using Google Colaboratory, and the dataset was obtained from Egypt and categorized into four different age groups [18].

Figure 1 illustrates the proposed Approach for diabetes prediction using machine learning techniques across different age groups. Initially, the raw dataset is divided into four age categories, namely Pediatrics, Early Adulthood, Middle Age, and Geriatrics. Each subgroup undergoes a data preprocessing stage, which includes handling categorical variables, managing missing values, removing duplicates, feature scaling, and addressing class imbalance using Random Over Sampling.

The processed datasets are then subjected to three different data split ratios, namely 80:20, 70:30, and 60:40, to evaluate model robustness and generalization performance. Each split ratio splits the dataset into two part, training and testing set, with the models being trained on the former and tested on the latter. This parallel experimental design guarantees an overall evaluation of model performance with varying training conditions. After model training and testing, standard metrics are used to evaluate its performance. Confusion matrices, feature importance analysis, and performance heatmaps are also used to analyze the results to determine the strongest and most accurate models. This method allows a comparative study of machine learning algorithms in detail and provides an insight into performance differences between age groups and validation strategies.

2.1 DATASET COLLECTION

The study used the dataset that was collected from the Egypt and then was categorized from four different age groups starts from 1 year to 80 years with the total cases of 13000 records. This dataset used both genders which are male and female. Then the Egyptian dataset was able to be categorized into four age groups as shown in the Table 1.

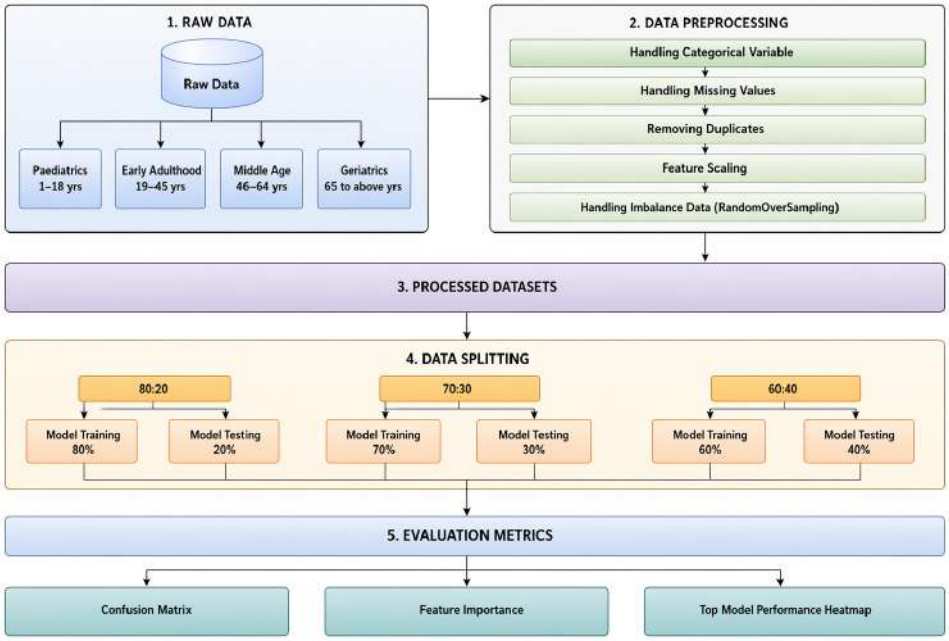

Figure 1. The proposed Approach

Table 1. Categorization of the Age Group

S/N	Name of the Age Group	Age (Years)
1	Pediatrics	1–18
2	Early Adulthood	19–45
3	Middle Age	46–64
4	Geriatric	65 and above

Then, Table 2 showed the 12 attributes for the pediatrics patient's group whilst Table 3 showed the three groups of adults with 23 attributes.

Table 2. Pediatrics Dataset Description

S/N	Attributes	Description
1.	Age	1–18
2.	Sex	Male and Female
3.	BMI	Body Mass Index (Weight and Height)
4.	Glucose Level	Glucose level present in the patients
5.	Insulin Level	Indicates the amount of insulin in the blood
6.	HbA1c	Person's average sugar level for the past 2 to 3 months
7.	Blood Pressure	Amount of the patient's blood pressure as it flows through arteries
8.	Cholesterol (mg/dL)	Measure of the amount of cholesterol in the blood
9.	Physical Activity Level	1. High 2. Moderate 3. Low
10.	Family History	1. Present 2. Absent
11.	Diet Score	Ranges from 1 to 10
12.	Type of Diabetes	0 – No diabetes 1 – Type 1 diabetes

2.2 ANALYZING TOOLS

The software tool that was used was Google Colaboratory that uses python programming language and Jupyter notebook which is free cloud-based platform developed by google. As depicted by the age groups in Figure 1 both datasets were cleaned in pre-processing stage.

2.3 DATA PREPROCESSING

In this study, data preprocessing phase was applied to improve data quality and ensure suitability for machine learning models. This includes handling categorical variables through appropriate encoding, managing missing values, removing duplicate records, performing feature scaling, and addressing class imbalance using techniques such as random oversampling. These steps result in processed datasets that are clean, structured, and ready for analysis which are described as follows:

2.3.1 HANDLING CATEGORICAL VARIABLES

In this study, preprocessing was applied to both paediatric and adult datasets containing numerical and categorical variables related to clinical, demographic, and lifestyle factors. For handling Categorical variables, it was transformed into numerical representations

Table 3. Early Adulthood, Middle Age and Geriatric Dataset Description

S/N	Attributes	Description
1.	Age	19 and above
2.	Sex	Male and Female
3.	BMI	Body Mass Index (Weight and Height)
4.	Fasting Blood Glucose (mg/dL)	Blood sample to be taken before patients start to eat
5.	Insulin Level	Indicates the amount of insulin in the blood
6.	Family History	1. Present 2. Absent
7.	HbA1c	Person's average sugar level for the past 2 to 3 months
8.	Environmental Factors	1. Present 2. Absent
9.	Systolic Blood Pressure	The maximum blood pressure during contraction of the ventricles
10.	Cholesterol Levels (mg/dL)	Measure of the amount of cholesterol in the blood
11.	Physical Activity	1. High 2. Moderate 3. Low
12.	Socioeconomic Factors	1. High 2. Moderate 3. Low
13.	Smoking Status	1. Smoker 2. Non-Smoker
14.	Glucose Tolerance Test	1. Normal 2. Abnormal
15.	Pancreatic Health	Maintaining a weight that is healthy for the patients
16.	Pulmonary Function	1. Normal 2. Abnormal
17.	Cystic Fibrosis Diagnosis	1. Present 2. Absent
18.	Steroid Use History	1. Yes 2. No
19.	Genetic Testing	1. Positive 2. Negative
20.	Urine Test	1. Ketones Present 2. Normal 3. Protein Present
21.	Neurological Assessments	1. Severe 2. Moderate 3. Mild
22.	Dietary Habits	1. Healthy 2. Unhealthy
23.	Type of Diabetes	0 – No diabetes 1 – Type 1 diabetes 2 – Type 2 diabetes

using appropriate encoding techniques, while numerical attributes were retained and standardized where necessary. Additionally, age was utilized to segment the data into paediatric and adult groups to support comparative analysis. These preprocessing steps ensured a structured and consistent dataset, thereby improving model performance and reliability.

2.3.2 HANDLING MISSING VALUES

Missing values within the dataset were addressed using appropriate imputation techniques to ensure data completeness and consistency. Numerical attributes were handled using statistical measures such as mean imputation, while categorical variables were treated using mode imputation. This process helps prevent bias and ensures that the dataset remains suitable for model training without introducing inconsistencies or reducing the dataset size unnecessarily.

2.3.3 REMOVING DUPLICATES

Duplicate records were identified and removed from the dataset to eliminate redundancy and ensure data integrity. The presence of duplicate entries can lead to biased model training and overfitting, as the model may give undue importance to repeated observations.

2.3.4 FEATURE SCALING

Feature scaling was applied to continuous numerical attributes, including age, body mass index (BMI), glucose level, insulin level, HbA1c, blood pressure, cholesterol levels, and diet score, to normalize their ranges and improve model performance.

2.3.5 HANDLING IMBALANCED DATA

To address class imbalance within the dataset, Random Oversampling was employed to increase the number of instances in the minority class. In medical datasets, imbalanced class distribution can lead to biased predictions, where the model favours the majority class. By balancing the dataset, the model's ability to correctly identify minority class instances, such as less frequent diabetes cases, is significantly improved.

2.4 HOLDOUT METHOD

In the preprocessing of the dataset, the random oversampling was used, then proceeding with the holdout method. The holdout method is the validation method that is used to validate the dataset for splitting the dataset into Training set and Testing tests. The training set is utilized to fit or train the model, whereas the testing set is reserved for assessing the model's generalization ability on previously unseen data. In this study the dataset is split into the ratios of 80:20, 70:30 and 60:40 of all age groups.

2.5 MACHINE LEARNING ALGORITHM USED IN PREDICTION

A diverse set of machine learning algorithms, including Random Forest (RF), Decision Tree (DT), Extra Trees (ET), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, Bagging, Logistic Regression (LR), Naïve Bayes (NB), and Gradient Boosting (GB), was selected to ensure a comprehensive evaluation of diabetes prediction performance across different modeling paradigms. These algorithms were selected to reflect a balanced representation of tree-based models, ensemble learning algorithms, distance-based algorithms, probabilistic algorithms and linear classifiers. Their inclusion allows for identifying the most effective model for diabetes prediction across different age groups and are discussed as follows:

- **Random Forest**: This algorithm is used to enhance the high accuracy, that is used to build and integrate multiple decision trees.

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_n(x)\}$$

This algorithm was selected due to its robustness in handling nonlinear relationships and its ability to reduce variance through aggregation. In medical datasets, where features such as glucose level, BMI, and age may exhibit complex interactions, Random Forest provides improved generalization and resistance to overfitting.

- **Decision Tree**: It is a structure that follows like a flowchart like model that splits the data in branches based on the condition, and the path that ends is a prediction or decision.

$$\text{Entropy} = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

The model is highly interpretable, making it suitable for medical applications where decision transparency is critical. However, it is prone to overfitting, particularly when dealing with noisy or small datasets.

- **Extra tree**: This is an algorithm which is also known as Extremely Randomized Trees that combines multiple decision trees but with additional Randomization.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

This approach enhances generalization and reduces variance compared to standard decision trees. It is particularly effective for medical datasets with noisy features, as the increased randomness prevents overfitting.

- **K Nearest Neighbors**: It is the algorithm that uses proximity to compare one point of dataset that was trained and make the prediction of what it had memorized.

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

While KNN is simple and intuitive, it suffers from limitations in high-dimensional spaces. The presence of irrelevant or redundant medical features reduces the effectiveness of distance metrics, leading to lower classification accuracy.

- **Logistic Regression**: It is the supervised machine learning algorithm that uses its probabilistic technique that estimates of the occurring to the event.

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

This algorithm was included as a baseline model due to its interpretability and effectiveness in binary classification. However, its assumption of linear relationships between features and the outcome may limit its performance in complex medical datasets.

- **Support Vector Machine**: It is the algorithm for classification and regression tasks.

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1$$

It is particularly effective in high-dimensional spaces and can model nonlinear relationships using kernel functions. This makes it suitable for medical datasets with complex feature interactions. However, its performance is sensitive to parameter selection.

- **Naïve Bayes**: It used Bayes theorem that based on probability, and it is called “Naïve” Due to the independent of the attributes of the given class.

$$P(y \mid x) = \frac{P(x \mid y)\,P(y)}{P(x)}$$

It assumes independence among features, which simplifies computation and enables fast training. However, this assumption is often violated in medical datasets, where features such as BMI and glucose levels may be correlated, potentially reducing classification performance.

- **Adaboost**: It is also known as Adaptive boosting that combines multiple weak leaners in order to create strong classifiers thus improving accuracy.

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

Although AdaBoost improves classification performance by focusing on difficult samples, it is sensitive to noise. In medical datasets, where measurement errors and inconsistencies are common, this sensitivity may lead to overfitting and reduced performance.

- **Bagging**: It is among the techniques that are also known as Bootstrap Aggregating. It is used to reduce overfitting by training many models on random parts of the data and finally averages their results.

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{(*b)}(x)$$

This method improves model stability and is particularly effective when applied to high-variance models such as decision trees. However, it does not address model bias.

- **Gradient Boosting**: It is the technique that is used for machine learning to combine weak prediction model into a single model, and target label encoding, depending on variable type.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

This method achieves high predictive accuracy by correcting errors iteratively. It is well-suited for capturing complex patterns in medical data but requires careful tuning to prevent overfitting.

2.6 EVALUATION PERFORMACE METRICS

There are numbers of metrics that evaluate the datasets in the machine learning algorithm. In this study, the researchers used the following metrics with their formulae whereby TP is regarded as True Positive, TN is represented as True Negative whilst FP is False Positive and FN is False Negative.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

3 RESULTS

The Egyptian dataset of the diabetes has been tested by categorizing the dataset into four age groups with the ten machine learning algorithms. Table 4, Table 5, Table 6 and Table 7 showed the results of each age group across different splitting ratios.

Table 4. Results of Accuracy, Precision and Recall for Pediatric with Different Splitting

	80:20			70:30			60:40		
Model	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**
RF	0.98	0.99	0.96	0.97	0.98	0.95	0.98	0.98	0.96
DT	0.92	0.88	0.94	0.94	0.92	0.92	0.95	0.93	0.92
ET	0.98	0.98	0.96	0.97	0.97	0.92	0.97	0.98	0.95
KNN	0.96	0.97	0.95	0.96	0.96	0.93	0.97	0.97	0.95
LR	0.97	0.97	0.96	0.96	0.96	0.94	0.97	0.98	0.95
SVM	0.98	0.97	0.97	0.96	0.97	0.94	0.97	0.96	0.94
NB	0.97	0.96	0.97	0.96	0.95	0.95	0.97	0.96	0.94
AB	0.97	0.96	0.97	0.96	0.96	0.95	0.97	0.97	0.94
BAG	0.97	0.97	0.95	0.95	0.94	0.93	0.96	0.98	0.92
GB	0.96	0.96	0.95	0.96	0.96	0.94	0.97	0.98	0.94

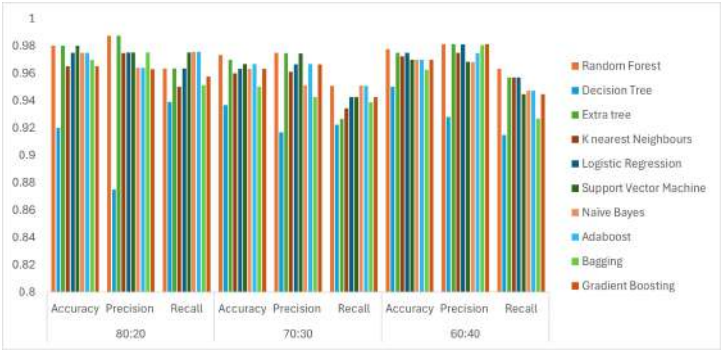

Figure 2. The bar graph of the split 60:40,70:30 and 80:20 of the ML for Pediatrics

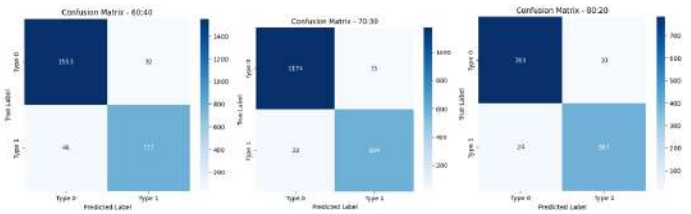

Figure 3. Confusion Matrix of the split 60:40,70:30 and 80:20 for the best model of Random Forest for Pediatrics

The Table 4 showed the pediatrics results that indicate that the Random Forest has the highest accuracy of 0.9800,0.9733 and 0.9775 with the split of 80:20, 70:30 and 60:40 respectively with the precision of 0.9750 and recall above 0.9630. While Extra Trees and Support Vector Machine had performed well with the accuracy 0.9800 of the ratio 80:20. Naïve Bayes records a recall of 0.9756 in the spilt of 80:20, and Logistic Regression achieves a precision of 0.9810 in 60:40. Furthermore, The confusion matrices for the 60:40, 70:30, and 80:20 train-test splits demonstrate that the Random Forest model performs consistently well in classifying Type 0 and Type 1 diabetes. The 60:40 split correctly classified 1553 non-diabetic and 777 diabetic cases with an overall accuracy of 97.8%. The 70:30 split achieved the highest performance with an accuracy of 97.3%, correctly predicting 1174 non-diabetic and 584 diabetic instances.

Similarly, the 80:20 split achieved an accuracy of 98%. Across all splits, the number of false positives and false negatives remained low, indicating strong predictive capability of the model as indicated in Figure 2.Moreover, in the Figure 3 has shown the feature importance of the pediatric group using the Random Forest model as the best model which reveals that BMI, insulin levels, age, and HbA1c are the most influential predictors of diabetes classification across all data splits of 60:40, 70:30, and 80:20. BMI consistently showed the highest importance score, indicating a strong relationship between body mass index and diabetes risk. Insulin concentration and HbA1c were also significant, as they show the clinical implications of both in the regulation of glucose and long-term monitoring of blood sugar levels. Contrastingly, other characteristics like sex, diet, and physical activity had relatively low importance scores indicating that there is a weak direct relationship with diabetes prediction in the dataset.

Table 5. Results of Accuracy, Precision and Recall for Early Adulthood with Different Splitting

	80:20			70:30			60:40		
Model	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**
RF	0.80	0.80	0.80	0.81	0.82	0.81	0.84	0.86	0.83
DT	0.69	0.70	0.69	0.69	0.71	0.69	0.70	0.71	0.70
ET	0.76	0.75	0.76	0.76	0.75	0.76	0.76	0.77	0.75
KNN	0.62	0.60	0.62	0.63	0.60	0.62	0.63	0.61	0.63
LR	0.77	0.76	0.77	0.77	0.81	0.77	0.78	0.77	0.77
SVM	0.80	0.81	0.80	0.80	0.81	0.80	0.79	0.81	0.80
NB	0.78	0.82	0.78	0.78	0.81	0.78	0.77	0.81	0.77
AB	0.61	0.58	0.62	0.62	0.59	0.62	0.62	0.60	0.62
BAG	0.74	0.73	0.74	0.74	0.74	0.75	0.76	0.75	0.76
GB	0.83	0.86	0.83	0.84	0.86	0.84	0.83	0.86	0.84

On the other hand, in the Early adulthood as indicated in Table 5 showed that the best performing algorithm is Gradient Boosting whereby it has high results across all metrics in all the ratios, closely followed by Random Forest, which also performs strongly reaching up to 0.8366 in accuracy and 0.8611 in precision and recall in the ratio split of 60:40. Moreover, Support Vector Machine also shows the constant high performance, with an average of 0.8000 in all three metrics in all splits. Naïve Bayes shows strong precision scores, reaching up to 0.8252 in 80:20 and 0.8162 in 60:40, although its accuracy slightly drops in the 60:40 split to 0.7768. Logistic Regression performs consistently, with accuracy improving from 0.7716 at 80:20 to 0.7839 at 60:40 and precision peaking at 0.8119 in the 70:30 split. On the lower end, K-Nearest Neighbors, Adaboost, and Decision Tree perform significantly worse.

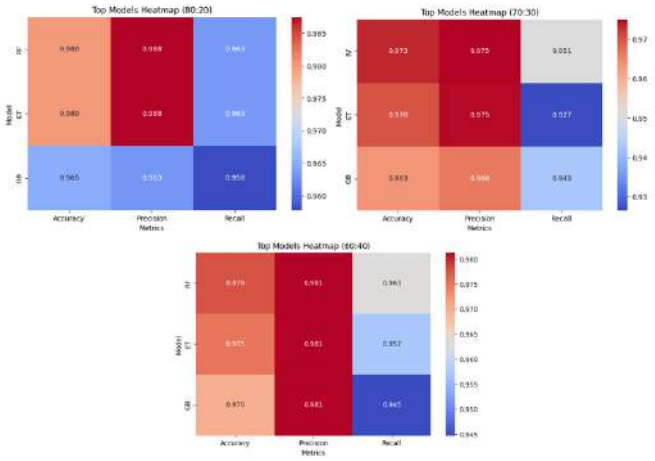

Figure 4. Top 3 Model performance Heatmap of the split 60:40,70:30 and 80:20 for Pediatrics Group

However, in Figure 4, the confusion matrices for the Gradient Boosting classifier demonstrate strong performance in distinguishing between Type 0, Type 1, and Type 2 diabetes across all train-test splits. In the 60:40 split, the model correctly predicted 851 Type 0 cases, 958 Type 1 cases and 667 Type 2 cases with a total accuracy of about 83.5. The same was also found in the 70:30 and 80:20 splits with both splits recording accuracy of approximately 83.6. The few cases of misclassification suggest that the model is very successful at determining the right type of diabetes in case of the chosen clinical and lifestyle characteristics.

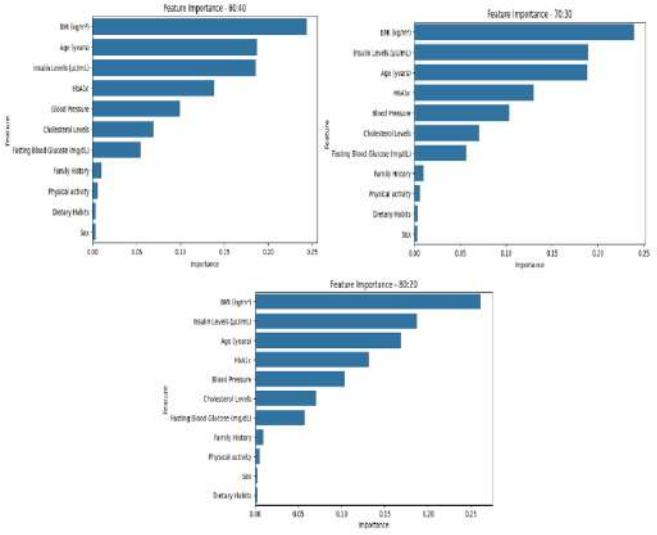

Figure 5. Feature importance of attributes of the split 60:40,70:30 and 80:20 for the best model of Random Forest for Pediatrics

The results of the feature importances in the Gradient Boosting model of the 60:40, 70:30 and 80:20 train-test splits are shown in Figure 5. The analysis shows that HbA1c is an outcome that seems the most powerful predictor in all splits followed by the insulin level, BMI, and age. The other variables like the environmental factors, smoking status, and socioeconomic factors have low scores in importance. The fact that the rankings of features stay the same across various data splits shows that the Gradient Boosting model is stable and able to find the most useful predictors of early adulthood diabetes.

For the Middle age group, the highest performing algorithms in this test are Random Forest, Support Vector Machine, Bagging, and Gradient Boosting, with each consistently achieving near perfect scores. As shown in the Table 6, Random Forest has accuracy of 0.99607, precision of 0.99610, and recall of 0.99609 at the split of 80:20, with only a marginal drop in other splits. Support Vector Machine and Gradient Boosting mirror these results, both hitting 0.99609 accuracy and precision. Bagging also performs outstandingly, with 0.99609 accuracy in 80:20 and still strong at 0.99119 in the 60:40 split.

Table 6. Results of Accuracy, Precision and Recall for Middle Age with Different Splitting

	80:20			70:30			60:40		
Model	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**
RF	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DT	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
ET	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
KNN	0.95	0.94	0.94	0.95	0.94	0.94	0.95	0.94	0.94
LR	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
SVM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
NB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
AB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
BAG	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
GB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

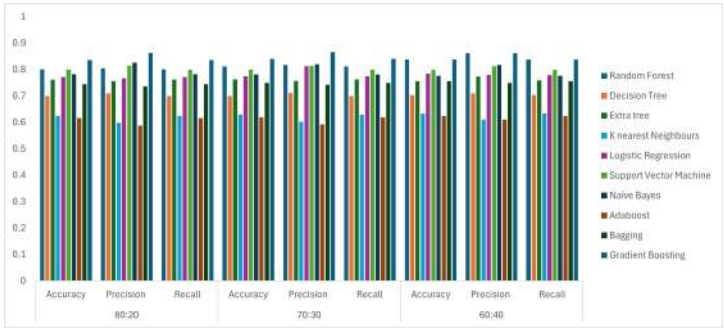

Figure 6. The bar graph of the split 60:40,70:30 and 80:20 of the ML for Early Adulthood

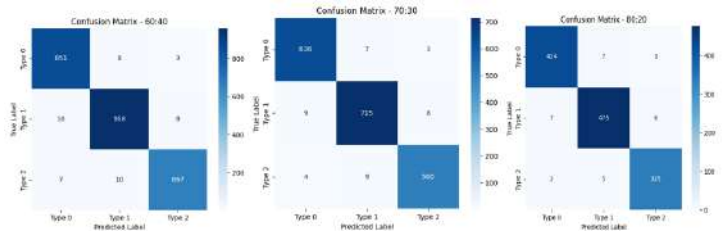

Figure 7. Confusion Matrix of the split 60:40,70:30 and 80:20 for the best model of Gradient Boosting for Early Adulthood

While, Extra Trees, Logistic Regression, and Naïve Bayes follow closely, all maintaining performance above

0.993 across splits. Decision Tree and Adaboost remain slightly lower in comparison, with Decision Tree peaking at 0.98239 and Adaboost at 0.98630 in 80:20 split. In the Figures 6 and 7 present the confusion matrices and feature importance results for the Gradient Boosting model applied to the middle-age diabetes dataset using 60:40, 70:30, and 80:20 train-test splits. The confusion matrices indicate that the model achieves very high classification accuracy across all splits, with most predictions correctly classified along the diagonal and only a small number of misclassifications. Moreover, the feature importance analysis reveals that HbA1c, insulin levels, fasting blood glucose, and BMI are the most influential predictors in the model.

Table 7. Results of Accuracy, Precision and Recall for Geriatric with Different Splitting

	80:20			70:30			60:40		
Model	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**
RF	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
DT	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
ET	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
KNN	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
LR	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
SVM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
NB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
AB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
BAG	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
GB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

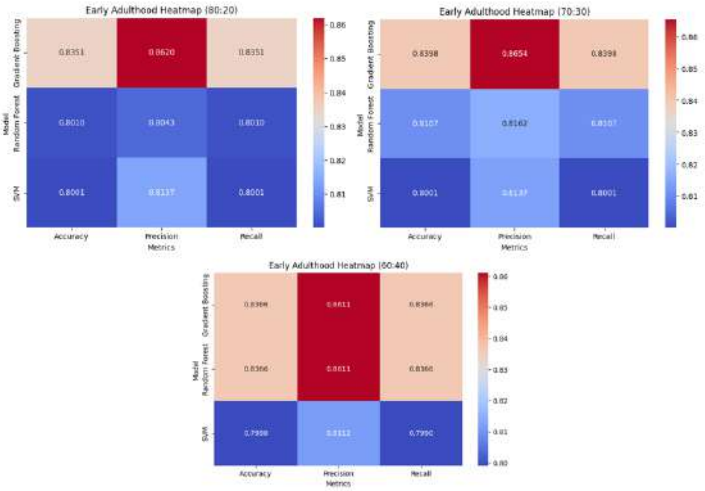

Figure 8. Top 3 Model performance Heatmap of the split 60:40,70:30 and 80:20 for Early Adulthood Group

Table 7 showed the results of the Geriatric age, indicated that Gradient Boosting achieves the highest and

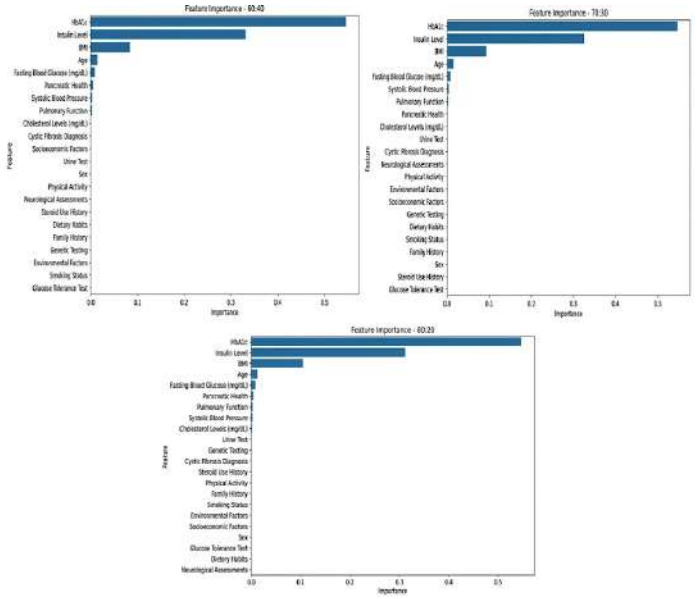

Figure 9. Feature importance of attributes of the split 60:40,70:30 and 80:20 for the best model of Gradient Boosting for Early Adulthood

most consistent performance, with accuracy, precision, and recall values of 0.9960 across all three splits, indicating its strong ability to correctly classify the instances in the dataset. Support Vector Machine also demonstrates very strong performance, achieving 0.9952 for all three evaluation metrics across the splits, followed closely by Logistic Regression with values of 0.9944 and Extra Trees with 0.9936. In addition, Random Forest and AdaBoost show similarly strong and stable performance, achieving 0.9928 and 0.9920, respectively, across all splits, which indicates the effectiveness of ensemble learning techniques for this classification problem.

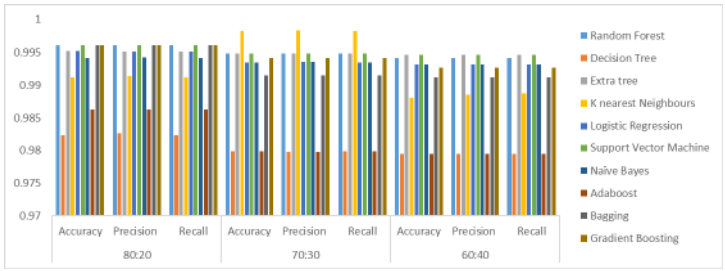

Figure 10. The bar graph of the splits 60:40, 70:30, and 80:20 of the ML for Middle Age

In contrast, Naïve Bayes shows slight variation across the different split ratios, recording 0.9905 accuracy in the

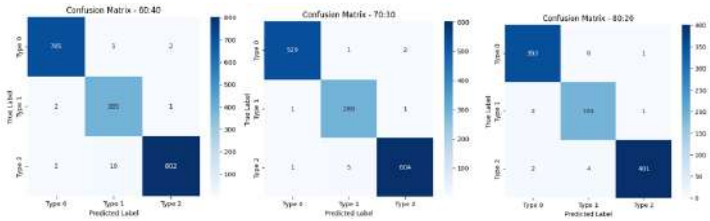

Figure 11. Confusion Matrix of the split 60:40,70:30 and 80:20 for the best model of Gradient Boosting for Middle Age

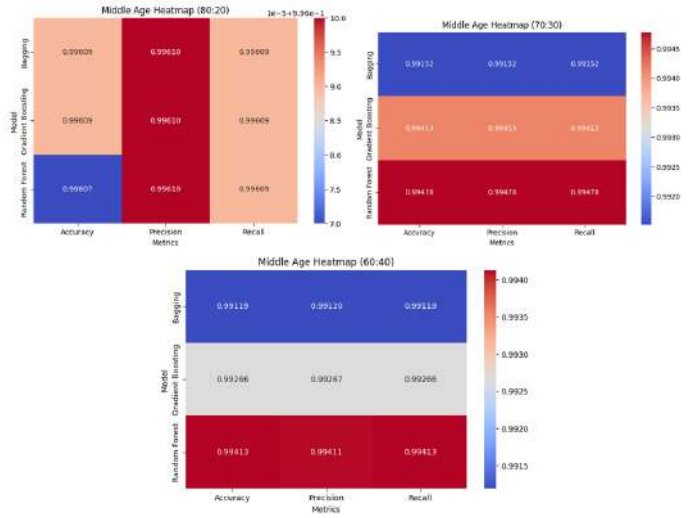

Figure 12. Top 3 Model performance Heatmap of the split 60:40,70:30 and 80:20 for Middle Age Group

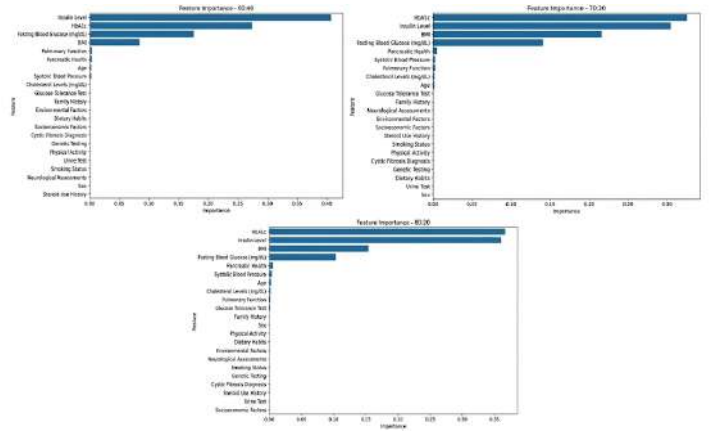

Figure 13. Feature importance of attributes of the split 60:40,70:30 and 80:20 for the best model of Gradient Boosting for Middle Age

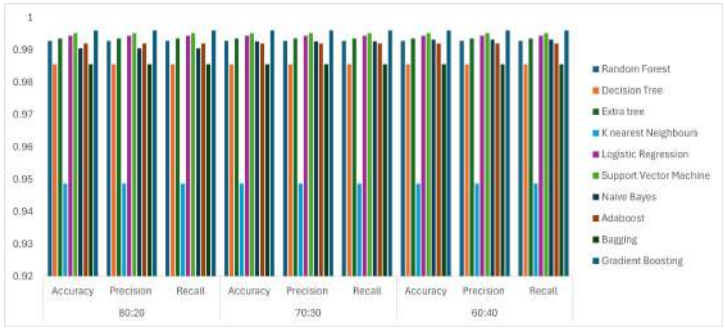

Figure 14. The bar graph of the split 60:40,70:30 and 80:20 of the ML for Geriatrics

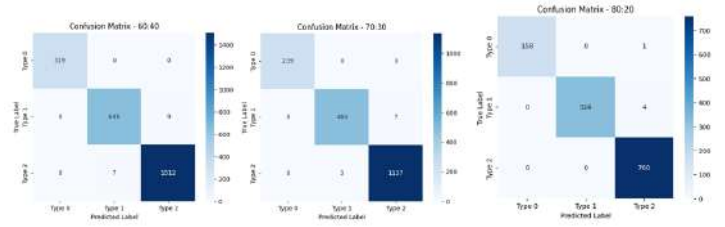

Figure 15. Confusion Matrix of attributes of the split 60:40,70:30 and 80:20 for the best model of Gradient Boosting for Geriatrics Age

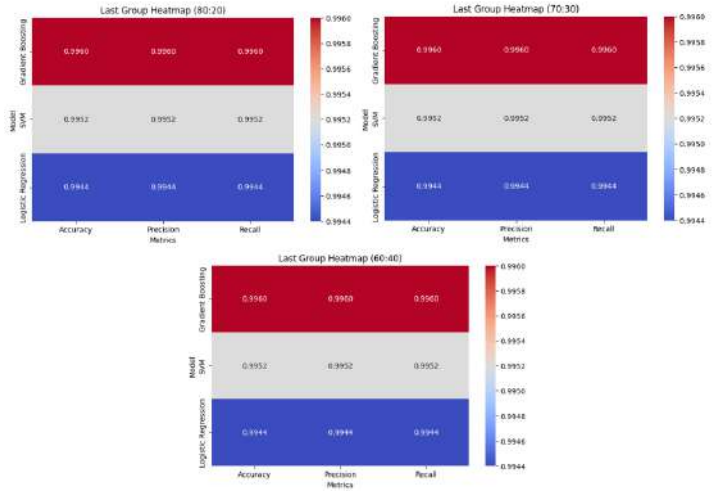

Figure 16. Top 3 Model performance Heatmap of the split 60:40,70:30 and 80:20 for Geriatric Group

80:20 split, improving to 0.9926 in the 70:30 split, and further increasing to 0.9932 in the 60:40 split, suggesting that its performance improves with a larger training dataset. Decision Tree and Bagging produce identical results across the splits, achieving 0.9856 for accuracy, precision, and recall, which indicates moderate predictive performance compared to the leading models. With the lowest performance at the lower end, K-Nearest Neighbors (KNN) has the lowest results of 0.9487 on all met-

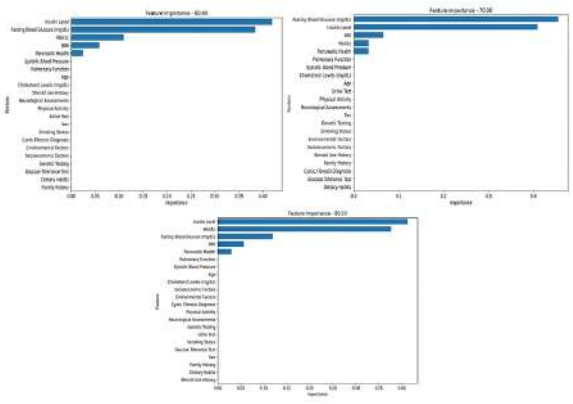

Figure 17. Feature importance of attributes of the split 60:40,70:30 and 80:20 for the best model of Gradient Boosting for Geriatrics Age

rics and splits, indicating that distance-based algorithms are ineffective on this data. Lastly, Figures 8 and 9 below show the confusion matrices and feature importance outputs of the Gradient Boosting model on the geriatrics dataset using 60:40, 70:30 and 80:20 train-test splits. The confusion matrices show that the classification accuracy is very high with the majority being correctly classified on the diagonal and few instances of misclassification between Type 1 and Type 2 diabetes. Moreover, the feature importance analysis shows that insulin level, fasting blood glucose, BMI, and HbA1c are the most significant predictors of diabetes classification among elderly individuals.

4 DISCUSSIONS

The confusion matrices in Figure 3, Figure 7, Figure 11 and Figure 15 offer a comprehensive foundation to investigate classification behaviour in age groups. The confusion matrix of the paediatric group demonstrates that there is a strong tendency to concentrate correct cases on the diagonal, with a little bit more off-diagonal cases than middle age and geriatric groups.

This shows that the models are reliable, but there is average uncertainty in the classification of classes, presumably because of natural physiological variability and not so stable clinical patterns in young cohorts. Conversely, the middle age and geriatric groups are almost perfectly diagonally dominant and these groups are said to have highly separable classes and stronger decision boundaries.

Misclassification is relatively larger in the early adulthood group, however, supporting the existence of overlapping feature distributions. It is interesting to note that the structure of the confusion matrices of the various data splits are similar, indicating that the behaviour of models is not very dependent on the sample size of the training data, but rather by the inherent properties of the data themselves.

Also, the performance heatmaps that are presented in Figure 4, Figure 8, Figure 12, Figure 16 and Figure 14 further confirm these findings as they allow comparing cross-models and cross-metrics.. The heatmaps in the paediatric group indicate a higher uniformity of values of accuracy, precision, and recall but with slight variations relative to that of the near uniformity of the values of the middle age and geriatric groups. This is an indication that, although the performance is good, the models still encounter slight inconsistencies in regard to balancing false positives and false negatives. Conversely, the early adulthood heatmaps are more variable and less intense, which suggests its susceptibility to complexity of data and lower predictors.

In all the categories, ensemble models are invariably in the highest performing places proving their ability to capture nonlinear relationships and reduce overfitting. The fact that the patterns of the heatmaps are almost identical between 80:20, 70:30 and 60:40 splits supports the thesis that the models are characterized by consistent generalization, and that the difference in performance is mostly conditioned by the structure of the datasets, as opposed to the configuration of the experiment.

Lastly on the feature importance results in Figure 5, Figure 9, Figure 13 and Figure 17 give some additional information on these performance differences. The importance of features in the paediatric group is moderately distributed with some of the features playing a significant role without a single most important predictor. This distribution implies that the classification decisions are based on a mix of factors, which can cause moderate uncertainty and justify the little performance variability seen in the confusion matrices and heatmaps.

By comparison, middle age and geriatric populations show very high levels of concentrated feature importance, and a very few variables are highly predictive,

leading to increased accuracy and consistency. The early adulthood group once again exhibits a more dispersed pattern of importance, with less feature discriminating power, and with this being the cause of poor model performance. The stability of the feature rankings in the various splits suggests that these trends are not data partition-dependent.

Thus, the analysis of the paediatric group alone and in combination with older age groups indicates that the paediatric group performs strongly, but it is the middle ground between highly separable age groups and the more complicated early adulthood group and that feature structure is crucial in determining the effectiveness of the model.

This paper provides a longer benchmarking story of machine learning algorithms on the Egyptian diabetes dataset, and makes a number of significant contributions to the existing literature. This study is more extensive than the previous work carried out by the same authors on the same topic, as it tests six algorithms on three age groups, and adds a Pediatric subgroup never studied before. The findings show a better predictive performance as the accuracy values of 0.98–0.99 in Pediatric group and a higher value of 0.99 in the Middle Age and Older groups, are found. Additionally, the use of multiple data split ratios (80:20, 70:30, and 60:40) enhances validation and confirms model robustness. The findings highlight that ensemble methods, particularly Random Forest and Gradient Boosting, consistently outperform other models and that classification performance improves across age groups, indicating increased feature separability.

5 CONCLUSIONS

Machine learning prediction of diabetes has a different degree of accuracy in different age groups because physiologic, behavioral, and clinical differences are linked with each stage of life. This research compared different machine learning algorithms in four age groups that included, pediatric, early adulthood, middle age and geriatric groups. The same dataset features, validation strategies and evaluation metrics were used consistently, which allowed performing a reliable comparison of the model performance between these groups.

The findings indicated that the ensemble learning algorithms including the Random Forest, Gradient Boosting and Extra Trees along with the Support Vector machine had high predictive performance in the pediatric and early adulthood group. In such populations, indicators associated with diabetes are more likely to exhibit more definite patterns, which can be successfully modeled by machine learning models to identify the connections between clinical and lifestyle characteristics.

On the same note, the middle age data set resulted in close to perfect prediction modeling in various models. This effect can probably be explained by the fact that the well defined clinical risk indicators are more prevalent in this age group and they form classes that are more separable in the dataset and allow models to learn predictive patterns that are stable.

Lastly, geriatric group showed better predictive performance in all the algorithms. This could be due to the clinical indicators of diabetes becoming clearer and more predictable with age. In elderly populations, physiological alterations associated with age, comorbidity, medications, and coinciding symptoms with normal aging produce unique health patterns that may result in diabetes being more prominent to machine learning models. These clearer patterns may improve the ability of algorithms to distinguish Type 1 and Type 2 diabetic and non-diabetic cases. The findings emphasize that diabetes prediction is age-dependent, highlighting the importance of developing age-specific modeling strategies to further enhance predictive accuracy.

Author Contributions

Saida O. Said : Conceptualization, Methodology, Software, Validation, Writing- Original draft preparation **Nurul Liyana Binti Mohamad Zulkufli** : Data curation, Visualization, Investigation and Supervision.**Asmarani Binti Ahmad Puzi** : Software, Validation and Supervision. **Asadullah Shah** : Writing- Reviewing and Editing and Supervision.

Compliance with Ethical Standards

It is declared that all authors don't have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

Data Availability

The dataset used in this study is not publicly available due to data access restrictions and ethical considera-

tions. However, the implementation code is publicly available at:

https://github.com/brown-panther/diabetes-prediction-ml

The experimental setup and additional details can be provided by the authors upon reasonable request

AI Usage Disclosure:

The authors used AI-assisted tools only for language and grammar improvement. No AI was used to generate research content, results, or interpretations. The authors confirm that all work is original and take full responsibility for the manuscript.

Funding Information

This research is funded by International Institute of Islamic Thought (IIIT) and International Islamic University Malaysia (IIUM) under IIIT-IIUM Scholarship and IIIT-RMC Sponsorship Grant.

References

[1] A. D. Association, "Standards of medical care in diabetes—2019 abridged for primary care providers," *Clinical Diabetes*, vol. 37, p. 11, 2019.

[2] W. H. Organization, *Global Report on Diabetes*. Geneva, Switzerland: World Health Organization, 2024.

[3] T. A. A. Aaty, M. M. Rezk, M. H. Megallaa, M. E. Yousseif, and H. S. Kassab, "Serum leptin level and microvascular complications in type 2 diabetes," *Clinical Diabetology*, vol. 9, pp. 239–244, 2020.

[4] A. Sapra, P. Bhandari, and A. W. Hughes, *Diabetes Mellitus (Nursing)*. Treasure Island, FL, USA: StatPearls Publishing, 2021.

[5] W. H. Organization, *Classification of Diabetes Mellitus*. Geneva, Switzerland: World Health Organization, 2019.

[6] V. Khalate and P. B. Sukeshkumar, "Diabetes prediction using machine learning algorithm," *International Journal of Research in Applied Science Engineering Technology*, vol. 12, pp. 1963–1969, 2024.

[7] H. A. et al., "The application of unsupervised clustering methods to alzheimer's disease," *Frontiers in Computational Neuroscience*, vol. 13, p. 31, 2019.

[8] A. Spector, W. Zhu, J. Hossain, and N. Roy, "Simulated forest environment and robot control framework for integration with cover detection algorithms," in *Proc. IEEE/ACM Int. Conf. Big Data Comput., Appl. Technol. (BDCAT)*. New York, NY, USA: IEEE, 2022, pp. 277–283.

[9] M. A. Khan, "Real-world applications and research directions for machine learning: Challenges and defies," *Cloud Computing and Data Science*, pp. 2949–2954, 2023.

[10] H. Lee, M.-B. Park, and Y.-J. Won, "Ai machine learning–based diabetes prediction in older adults in south korea: Cross-sectional analysis," *JMIR Formative Research*, vol. 9, p. e57874, 2025.

[11] S. Afolabi, N. Ajadi, A. Jimoh, and I. Adenekan, "Predicting diabetes using supervised machine learning algorithms on e-health records," *Informatics in Health*, vol. 2, pp. 9–16, 2025.

[12] S. K. Bigdeli, M. Ghazisaedi, S. M. Ayyoubzadeh, S. Hantoushzadeh, and M. Ahmadi, "Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: A cross-sectional study," *BMC Medical Informatics and Decision Making*, vol. 25, p. 3, 2025.

[13] H. K. et al., "Machine-learning algorithms in screening for type 2 diabetes mellitus: Data from the fasa adults cohort study," *Endocrinology, Diabetes & Metabolism*, vol. 7, p. e00472, 2024.

[14] G. K. Teimoory and M. R. Keyvanpour, "An effective feature selection for type ii diabetes prediction," in *Proc. 10th Int. Conf. Web Research (ICWR)*. New York, NY, USA: IEEE, 2024, pp. 64–69.

[15] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *Journal of Personalized Medicine*, vol. 13, p. 406, 2023.

[16] I. J. Kakoly, M. R. Hoque, and N. Hasan, "Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique," *Sustainability*, vol. 15, p. 4930, 2023.

[17] S. O. Said, N. L. B. M. Zulkufli, A. B. A. Puzi, and A. Shah, "Performance evaluation metrics comparison between weka and google colab for predicting type 1 and type 2 diabetes using machine learning algorithms on an egyptian dataset," in *Proc. 10th Int. Conf. Inf. Commun. Technol. Muslim World (ICT4M)*, Kuala Lumpur, Malaysia, 2025, pp. 1–6.

[18] S. O. Said, N. L. Zulkufli, A. B. Puzi, and A. Shah, "Performance metrics analysis for diabetes prediction using machine learning algorithm," in *Proc. 9th IEEE Int. Conf. Eng. Technol. Appl. Sci. (ICETAS)*, Kuala Lumpur, Malaysia, 2024, pp. 1–5.