information

Article

A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh

Md. Jamal Uddin [1], Md. Martuza Ahamad [1], Md. Nesarul Hoque [1], Md. Abul Ala Walid [2], Sakifa Aktar [1], Naif Alotaibi [3], Salem A. Alyami [3], Muhammad Ashad Kabir [4] and Mohammad Ali Moni [5,*]

1 Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj 8100, Bangladesh; jamal.bsmrstu@gmail.com (M.J.U.); martuza.cse@bsmrstu.edu.bd (M.M.A.); mnhshisir@gmail.com (M.N.H.); sakifa.cse@bsmrstu.edu.bd (S.A.)
2 Department of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology (BAUET), Natore 6431, Bangladesh; abulalawalid@gmail.com
3 Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 13318, Saudi Arabia; nmaalotaibi@imamu.edu.sa (N.A.); saalyami@imamu.edu.sa (S.A.A.)
4 School of Computing, Mathematics, and Engineering, Charles Sturt University, Bathurst, NSW 2795, Australia; akabir@csu.edu.au
5 Artificial Intelligence & Data Science, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St Lucia, QLD 4072, Australia
* Correspondence: m.moni@uq.edu.au; Tel.: +61-414-701-759

Abstract: Diabetes is a chronic disease caused by a persistently high blood sugar level, causing other chronic diseases, including cardiovascular, kidney, eye, and nerve damage. Prompt detection plays a vital role in reducing the risk and severity associated with diabetes, and identifying key risk factors can help individuals become more mindful of their lifestyles. In this study, we conducted a questionnaire-based survey utilizing standard diabetes risk variables to examine the prevalence of diabetes in Bangladesh. To enable prompt detection of diabetes, we compared different machine learning techniques and proposed an ensemble-based machine learning framework that incorporated algorithms such as decision tree, random forest, and extreme gradient boost algorithms. In order to address class imbalance within the dataset, we initially applied the synthetic minority oversampling technique (SMOTE) and random oversampling (ROS) techniques. We evaluated the performance of various classifiers, including decision tree (DT), logistic regression (LR), support vector machine (SVM), gradient boost (GB), extreme gradient boost (XGBoost), random forest (RF), and ensemble technique (ET), on our diabetes datasets. Our experimental results showed that the ET outperformed other classifiers; to further enhance its effectiveness, we fine-tuned and evaluated the hyperparameters of the ET. Using statistical and machine learning techniques, we also ranked features and identified that age, extreme thirst, and diabetes in the family are significant features that prove instrumental in the detection of diabetes patients. This method has great potential for clinicians to effectively identify individuals at risk of diabetes, facilitating timely intervention and care.

Keywords: diabetes mellitus; machine learning; survey; feature selection; feature importance

Citation: Uddin, M.J.; Ahamad, M.M.; Hoque, M.N.; Walid, M.A.A.; Aktar, S.; Alotaibi, N.; Alyami, S.A.; Kabir, M.A.; Moni, M.A. A Comparison of Machine Learning Techniques for the Detection of Type-2 Diabetes Mellitus: Experiences from Bangladesh. *Information* **2023**, *14*, 376. https://doi.org/10.3390/info14070376

Academic Editors: Sidong Liu, Cristián Castillo Olea and Shlomo Berkovsky

Received: 13 May 2023
Revised: 25 June 2023
Accepted: 26 June 2023
Published: 2 July 2023

1. Introduction

Diabetes is a lifelong disease that prevents the body from obtaining energy from food sources due to a deficiency insulin, an influential factor in enhancing the cells' ability to absorb glucose and produce energy [1]. There are three primary types of diabetes: type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM), and gestational diabetes mellitus (GDM). In this study, we focused on T2DM, as it accounts for approximately 90% of all occurrences of diabetes [2]: insulin resistance, in which the body does not respond adequately to insulin, is a defining characteristic. T2DM is diagnosed most frequently in

elderly adults; however, it is becoming increasingly prevalent in children, teenagers, and young people due to rising rates of poor diet, physical inactivity, and obesity. Several risk factors, including diabetes in the family, excess weight, an unhealthy diet, a lack of exercise, increasing age, and higher blood pressure, are associated with T2DM.

Diabetes is the underlying cause of several associated diseases, including kidney disease, tuberculosis, cardiovascular disease, and eye management [3]. Patients with diabetes are susceptible to amputation, blindness, stroke, heart disease, kidney failure, and premature death [4]. According to the International Diabetes Federation (IDF), there were approximately 537 million diabetic patients worldwide—1 in 10 people—of whom 81% resided in low- and middle-income nations in 2021 [5]. In 2021, the IDF counted 6.7 million diabetes-related fatalities worldwide. In addition, the global cost of diabetes-related medical expenses in 2019 was USD 760 billion; this amount is set to grow to USD 825 billion by 2030 and to USD 845 billion by 2045 [6]. Chronic diseases induced by diabetes have imposed a financial burden on every nation [7].

The IDF estimates that in Bangladesh there are 7.1 million diabetics and approximately the same number of undiagnosed cases; this number is anticipated to double by 2025. In addition, the cost of diabetes imposes a significant burden on natural expenditures in low- and middle-income countries [8].

Our research focused on T2DM-type diabetes. Although this type of diabetes cannot be cured, most people can still avoid developing it. Early identification and lifestyle changes can minimize the chance of developing diabetes. T2DM risk can be accurately identified by doctors when treating an individual patient; however, clinicians encounter significant obstacles when screening thousands of patients with high-risk illnesses. In this situation, analytical methods are required for T2DM screening in the population.

Machine learning (ML) methods have been used to solve several problems recently, such as diagnosing cancer [9], COVID-19 [10], autism [11,12], meningitis, diabetes, and heart disease. Recent research suggests that ML can summarize patient characteristics and predict T2DM risk [13–17]. The authors of Haque and Alharbi [18] investigated 18 features of T2DM in Bangladesh. The principal contribution of this study was that the authors considered demographic and clinical data together and achieved the best output of 83.8% accuracy and 70% F1-score, using the LR model; however, the performance could be enhanced by including more features related to people's eating habits, lifestyle, and clinical diagnosis. In another work, Tasin et al. [19] specifically investigated females in Bangladesh. They developed a diabetes detection system with 81% accuracy using the XGBoost model with the ADASYN oversampling technique. They used an explainable AI approach with the LIME and Shapley additive explanations (SHAP) frameworks to provide feature weights relating to diabetes. They also deployed a website and an Android mobile application to detect diabetes in females. The authors only considered eight features in their detection system; therefore, the question of reliability arose. Nipa et al. [8] examined three datasets: the Sylhet Diabetes Hospital dataset (SDHD) of 520 samples; the pre-diagnosis diabetes dataset (PDD) of 558 samples; and the combined SDHD and PDD dataset (MDD). They tested 32 classifiers, where extra tree (ET), light gradient boosting machine (LGBM), stacking, multi-layer perceptron (MLP), histogram gradient boosting classifier (HGBC), RF, bagging, and gradient boosting classifier (GBC) presented more stable outputs. ET provided the best result of 97.11% accuracy for the SDHD dataset; MLP yielded the highest accuracy of 96.42%; and LGBM and HGBC showed the maximum output of 94.9% accuracy, separately. They applied the SHAP framework to find the features more responsible for diabetes. The authors of Kaur and Kumari [20] employed five ML models to identify a patient as diabetic or non-diabetic: linear SVM, radial bias SVM, k-NN, artificial neural network (ANN), and multi-factor dimensionality reduction (MDR). They obtained superior results with linear SVM and k-NN, where linear SVM displayed an accuracy of 89% and an F1-score of 87%, and k-NN exhibited an accuracy of 88% and an F1-score of 88%. Before applying the ML models, they filtered significant features using the Boruta wrapper feature selection method. In this research, the authors investigated the error analysis of the

detection system. Zheng et al. [21] analysed the T2DM-related features and experimented with various ML models to identify diabetic patients. The authors observed that SVM, J48, and RF presented a more stable output than the other ML models such as LR, NB, and k-NN.

We found many research works on the Pima Indians Diabetes Database (PIDD) dataset [22–31]. The dataset comprises 768 female patients, of whom 268 have diabetes, and 500 do not. The dataset has eight input features and one target variable. Here, Yahyaoui et al. [27], Abdulhadi and Al-Mousa [28], Tigga and Garg [30], Pranto et al. [31], and Ali et al. [23] proposed RF classifiers, while Saha et al. [22] and Wei et al. [25] preferred neural network-based models to obtain the best performance. On the other hand, Birjais et al. [26], Battineni et al. [29], and Howlader et al. [24] achieved the maximum output by utilizing gradient boosting (GB), LR, and the generalized additive model using LOESS (GAMLOESS) models, respectively.

The authors of Sneha and Gangil [32] attempted to employ a predictive analysis in the early detection of diabetes mellitus, ensuring that all significant features were utilized. With an accuracy of 82.03%, naive Bayes (NB) was the most accurate of the five ML algorithms utilized in the study. The highest specificities of RF and DT were 98% and 98.2%, respectively. The authors of [33] utilized LR, classification [34], and regression trees (CART), ANN, SVM, RF, and gradient boosting machine (GBM) classifiers to determine the likelihood an individual would develop T2DM. The GBM model achieved the best among those considered. In addition, they identified the significance of factors based on each classifier and the Shapley additive explanations approach and demonstrated the relevant features such as sweet affinity, urine glucose, age, heart rate, creatinine, waist circumference, uric acid, pulse pressure, insulin, and hypertension.

Le et al. [35] presented a ML model to predict early onset diabetes in patients and employed grey wolf optimization (GWO) and adaptive particle swam optimization (APSO) to optimize the number of significant input attributes. Using their proposed strategy, their computational results indicated a higher degree of accuracy (96% for GWO-MLP and 97% for APGWO-MLP). Next, Islam et al. [36] utilized two statistical analyses to determine the diabetes risk factors. They used diabetes information from the 2011 Bangladesh Demographic and Health Survey. Six ML-based classifiers were used to predict and categorize diabetes. Eleven of the fifteen examined factors were found to be associated with diabetes, and the bagged CART model had the highest accuracy and area under the curve at 94.3% and 0.6, respectively. Using ML techniques, Haq et al. [37] developed a diagnostic system for diabetes. Furthermore, a filtering approach based on the DT algorithm was used to select the most critical features. Experiments showed that the proposed method for choosing features improved the accuracy of the classifying predictive models.

Shuja et al. [38] created a model for diabetes prognosis based on data mining categorization approaches using a dataset from the Kashmir Valley Clinical Institute. After performing SMOTE for data oversampling, the balanced dataset was fed to five ML algorithms: bagging, SVM, MLP, simple logistic, and DT. DT had the best accuracy of 94.7%, precision of 0.947, and sensitivity of 0.947. Then, Chatrati et al. [39] suggested developing a smart domestic system using five different supervised ML approaches to monitor a patient's glucose and blood pressure. The accuracy of all algorithms from SVM was 75%, from k-NN was 74%, from DT was 66.1%, LR was 74.5%, and DA was 74.7%. In another work, Islam et al. [40] identified risk variables for T2DM and offered an ML method to predict it. Next, five ML methods predicted T2DM. For 2009–2010, these researchers indicated six potential risks: age, learning, marital status, SBP, smoking, and BMI. For 2011–2012, they identified nine threat issues: age, race, martial status, SBP, DBP, direct cholesterol, bodily activity, smoking, and BMI. The RF-based classifier achieved a correctness of 95.9%, a sensitivity of 95.7%, an F-score of 95.3%, and an AUC of 0.946.

A significant number of diabetic patients in Bangladesh have gone undetected. Furthermore, inadequate healthcare equipment is incapable of accommodating for or treating an extensive number of diabetic patients, and the expense of diabetes carries an immense

strain on the citizens of the country. Unfortunately, we have discovered very little research on the inhabitants of Bangladesh. Therefore, we must establish an automated system for the early diagnosis of T2DM. We have made the following contributions in this regard:

- We created a dataset of 508 study populations for diabetes.
- We applied and compared state-of-the-art clinically applicable ML models to conduct a benchmark analysis, aiming to contribute to further research in the field.
- We proposed a framework that utilizes ML techniques to detect diabetic patients.
- We ranked and identified significant features associated with diabetes mellitus.

2. Materials and Methods

The workflow of this study is depicted in Figure 1.

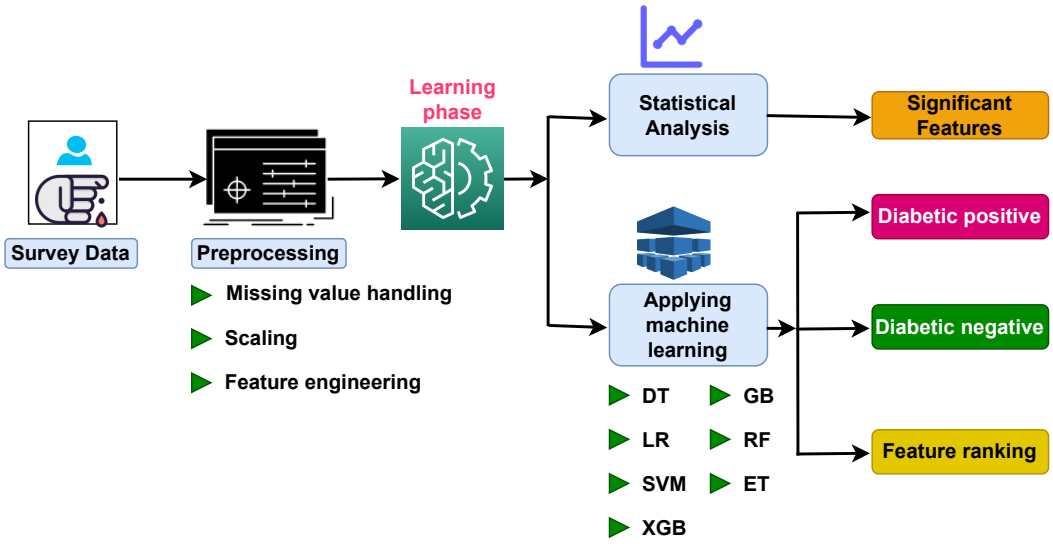

Figure 1. Workflow of this study.

The main dataset was gathered through surveys with closed-ended questions and direct interviews, also called in-person interviews. All respondents were from Bangladesh. In our study, we used the most commonly used variables (for preparing questionnaires) from recent articles on diabetes prediction models [33,35,41]. The primary dataset was converted into a secondary dataset in order to perform several mathematical operations. We performed an exploratory data analysis on our proposed Bangladeshi T2DM dataset to find out previously unknown information. Moreover, the dataset was processed through various stages, including missing value handling, data encoding, feature scaling, feature engineering, etc. Symmetrical analysis checked the amount of skewness in the target class. Next, the dataset was segmented into training and testing sets based on the target class ratio, and the training set exhibited the same level of skewness as before. Initially, seven ML models were trained and evaluated with this preliminary training set. SMOTE and ROS were applied to the training set to resolve the skewness property in the target class and improve the results of the ML model. Furthermore, seven ML models were applied to the datasets, balanced using SMOTE and ROS techniques. A statistical method known as the chi-squared test was also applied in order to identify any associative features, and several ML models were employed to detect any significant feature sets.

2.1. Data Collection and Description

In order to perform this research, we constructed a dataset from the perspective of the Bangladeshi population. Based on an analysis of the relevant studies, 33 questionnaires were made to obtain first-hand information. There were 508 respondents. The primary dataset was acquired by means of both online surveys and in-person interviews. Furthermore, we also collected the respective information about the participants' locations.

Therefore, it is seen that the study covered all 64 districts of Bangladesh with no selection bias. This dataset was turned into secondary data so that it could be used to build a model for ML. Each property is organized and briefly explained in Table 1, displaying the possible values for each attribute.

Table 1. Dataset description.

ID	Attribute Name	Feature Description	Count (N = 508), n (%) or Avg., Min, Max, Median
1	Age	Identifying age groups	1–18: 244 (48.03%) 19–40: 95 (18.7%) 40–65: 99 (19.49%) 65 or more: 70 (13.78%)
2	Gender	Gender of the respondent	Male: 249 (49.01%) Female: 259 (50.99 %)
3	Education Level	Level of education can be no education, primary, secondary or higher	Primary: 52 (10.36%) Secondary: 95 (18.7%) Higher: 305 (60.03%) No education: 51 (10.03%)
4	Diabetes in the family	Respondent's ancestors or parents suffered from the disease or not	Yes: 298 (58.66%) No: 210 (41.34%)
5	Occupation	Respondent can be unemployed, employed or find job	Looking work: 56 (11.02%) Not working: 189 (37.20%) Working: 260 (51.18%)
6	Household monthly income	Household monthly income of the respondent	Avg.: 30,477 Max: 300,000 Min: 1000 Median: 30,000
7	Wealth index	Wealth index can be low, middle or upper class	Poor: 46 (0.09%) Middle: 406 (79.91%) Rich: 58 (11%)
8	Place of residence	Place of residence indicates in which area the respondent usually lives	Urban: 233 (45.86%) Rural: 276 (54.14%)
9	Walk/ Run/ Physical exercise	How much jogging or walking the respondent did	None: 43 (8.46%) Less half: 90 (17.71%) More half: 151 (29.72%) One hour or More: 224 (44.09%)
10	BMI	Calculated from Height and Weight	Avg.: 23.2 Max: 58.44 Min: 12 Median: 23.12
11	Smoking	Whether the respondent smokes or not	Yes: 53 (10.44%) No: 455 (89.56%)
12	Alcohol consumption	Whether or not the respondent drinks alcohol	Yes: 22 (4.33%) No: 486 (95.67%)
13	Hours of sleep	Total hour of sleep each day	Avg.: 7.77 Max: 12 Min: 1 Median: 8
14	Regular intake of medicine (Except insulin)	Regular medication use, excluding insulin	Yes: 233 (45.86%) No: 275 (54.14%)
15	Junk food consumption	Prevalence of junk food consumption	Yes: 194 (38.18%) No: 314 (61.82%)

Table 1. *Cont.*

ID	Attribute Name	Feature Description	Count (N = 508), n (%) or Avg., Min, Max, Median
16	Stress	Stress level of respondent	Not at all: 66 (13%) Sometimes: 329 (64.76%) Often: 60 (11.8%) Always: 55 (10.82%)
17	Blood pressure level	Average level of blood pressure	High: 83 (16.33%) Normal: 383 (75.4%) Low: 43 (8.26%)
18	Hypertension	Whether or not the respondent has hypertension	Yes: 269 (52.95%) No: 239 (47.05%)
19	Frequency of urination	Frequency of urination each day	Not much: 360 (70.86%) Quite much: 148 (29.14%)
20	Extreme thirst	Whether or not the respondent is extremely thirsty	Yes: 253 (49.8%) No: 255 (50.2%)
21	Sudden weight loss	Whether the respondent ever noticed sudden weight loss	Yes: 110 (21.65%) No: 334 (65.74%) May be: 64 (12.6%)
22	Weakness	Whether the respondent feels more vulnerable	Yes: 225 (44.29%) No: 187 (36.81%) May be: 96 (18.9%)
23	More appetite	Whether the respondent feels more hungry	Yes: 151 (29.72%) No: 250 (49.21%) May be: 106 (20.86%)
24	Irritability	Whether the respondent feels irritability	Yes: 250 (49.2%) No: 258 (50.8%)
25	Delayed healing	Whether the respondent notices delayed healing of the wound	Yes: 213 (41.92%) No: 295 (58.08%)
26	Muscle stiffness	Whether the respondent notices muscle stiffness of their body	Yes: 234 (46.06%) No: 274 (53.94%)
27	Partial paralysis.	Partial paralysis of any part of the body is noticed or not	Yes: 69 (13.58%) No: 396 (77.95%) May be: 33 (8.46%)
28	Hair loss	Whether the respondent notices gradually losing hair or not	Yes: 231 (45.47%) No: 203 (39.76%) May be: 70 (14.76%)
29	Other diseases	Whether the respondent has other serious diseases	Yes: 260 (51.18%) No: 248 (48.82%)
30	Number of dependent family members	Total number of dependent family members	Avg.: 3.43 Max: 15 Min: 0 Median: 3
31	Living house type	Respondent home can be rental or owned	Owned: 409 (80.51%) Rented: 99 (19.49%)
32	Anxiety	Whether the respondent has extreme anxiety or not	Yes: 293 (57.67%) No: 215 (42.33%)
33	Diabetes (output factor)	Diabetes, non-diabetes	Yes: 275 (54.14%) No: 233 (45.86%)

2.2. Explanatory Data Analysis

Exploratory data analysis was conducted on our dataset in an attempt to reveal buried information. This dataset comprised 508 samples, and the output variable contained two distinct categories: Diabetes (affected with diabetes) and non-diabetes (not affected with diabetes). A property of imbalance was observed by analysing the distribution of output classes. There were 275 (54.14%) observations belonging to the majority class named diabetes, and the rest belonged to the minority class, non-diabetes. The maximum age of the individuals was 78 years, with a minimum of 18 years, and a median of 41 years. A total of 49.01% were men and 50.99% were women in our dataset. Two hundred and fifty-three individuals suffered from extreme thirst, while 148 had a high frequency of urination. A total of 576.7% of the population suffered from anxiety, whereas 51.8% suffered from other diseases.

The number of people with diabetes who had a family history of the disease was significantly higher than the number of those without a family history of diabetes. Diabetes affected 223 individuals whose ancestors or parents had the condition, but only 52 individuals whose ancestors or parents did not have the disease. Figure 2 represents the correlation between the features. We found no significant correlation between the features. Therefore, we did not remove any characteristics from the dataset. Figure 3 depicts the association between the category of the output variable non-diabetes and the walk/run/physical exercise variable, indicating that a person has a higher likelihood of falling into the category of non-diabetes if they are physically active for at least one hour. We also saw that most diabetic observations involved higher blood pressure levels compared to non-diabetic observations.

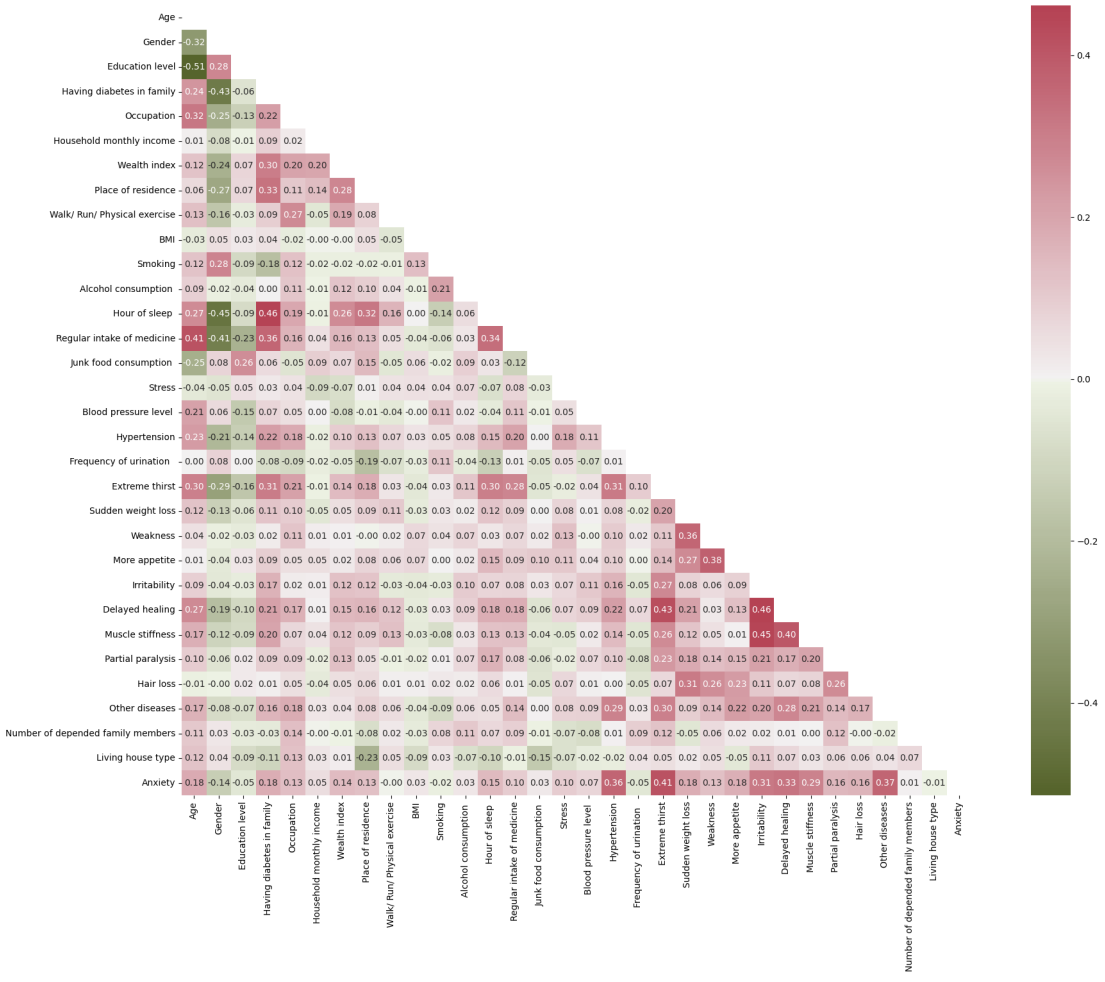

Figure 2. Heatmap between each of the features, representing the correlations between them.

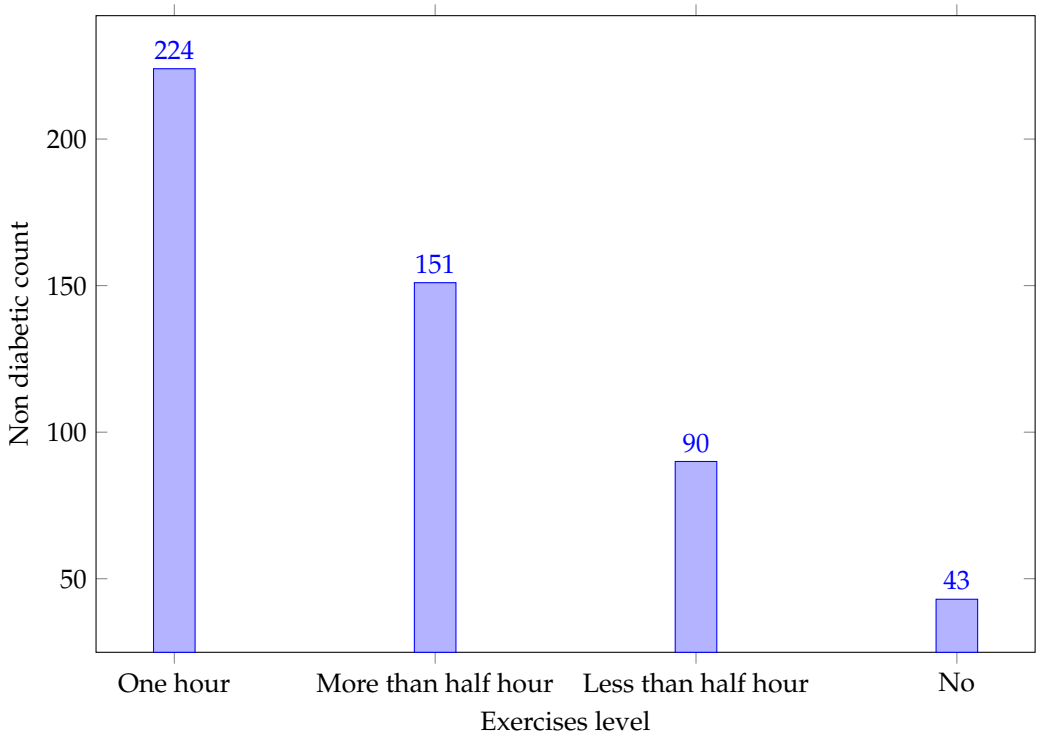

Figure 3. Relation between the output variable (non-diabetes) and the walk/run/physically active variable.

2.3. Data Preprocessing

We used the mean instead of the missing values for age and monthly household income found in our dataset. The BMI attribute was derived from the height and weight attributes. To scale the dataset, we also employed the standard scalar technique. The most important characteristic was determined using the recursive feature elimination technique. We rectified the class disparity using the SMOTE and ROS techniques.

2.4. Data Balancing Technique

In particular, when working with medical datasets, class imbalance is a prominent issue. This issue arises when instances are not distributed uniformly throughout the classes. As a result, the classifications of the classifiers become skewed, and the minority class is ignored. This problem can be resolved through either oversampling or undersampling. We utilized two oversampling techniques: the SMOTE and the ROS technique.

SMOTE is a method of oversampling in which the data points of minority classes are oversampled in order to balance the dataset [42]. It generates synthetic samples from minority classes and prevents duplication of samples, unlike standard oversampling algorithms. It randomly selects examples of the minority class, identifies their k nearest minority class neighbours, and then selects one of the neighbours [43].

On the other hand, ROS is a technique for balancing an unbalanced dataset prior to feeding it to ML classifiers in order to improve the performance and eliminate bias towards the majority class of the classifiers. Typically, it substantially increases the size of the dataset. It is a non-heuristic strategy that achieves data balancing by duplicating or replicating minority class samples at random [44].

2.5. Feature Transformation

The standard scalar is a process for scaling characteristics that removes the mean of every feature and normalizes its variance to one. It offers many benefits, including

being smooth, bidirectional, swift, and highly scalable. The standard scalar's equation is presented below:

$$\hat{X}_i = \frac{X_i - \bar{X}}{\sigma} \tag{1}$$

where the standard scalar is represented as $\hat{X}_i$, each observation as X_i, the mean as $\bar{X}$, and the standard deviation as σ [45].

2.6. Feature Selection

Feature selection is the primary data dimension reduction procedure. It increases the accuracy of the classifier's predictions by identifying a collection of attributes that strongly contribute to the target class. It shortens the procedure and reduces the cost of computation. We used the recursive feature elimination (RFE) technique in our work. It is an iterative method for selecting features based on the model's accuracy. In each iteration, it calculates the ranking score metric and eliminates low-ranking characteristics. Until the required number of attributes has been reached, the recursive operation continues [46].

2.7. Statistical Methods Identifying the Most Significant and Associative Diabetes Features

The chi-square test employs the p value to determine the significance of a dependent variable-related characteristic. H_0 is a "null hypothesis", denoting that the target variable and categorical feature have nothing in common. H_1 is an alternative hypothesis that says there is a strong link between the categorical feature and target variable. If the p value is greater than 0.05, the null hypothesis cannot be rejected because there is no connection between the target variable and categorical features. If the p value is less than 0.05, the null hypothesis is rejected as there is evidence of a connection between the categorical characteristics and the target variable. All of these features are then used in the next step of the ML pipeline [47]. The equation for $\tilde{\chi}^2$ is given below:

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{2}$$

where the observed frequencies is denoted as O_k, the expected frequencies as E_k, and the number of samples as n [46].

2.8. Machine Learning Model

In our work, we applied multiple ML algorithms to predict diabetes, including DT, LR, SVM, GBs, XGBs, RF, and our custom ETs.

- DT is a white box concept that has an effective learning component. Numerous leaf nodes, multiple internal nodes, and a central root node constitute DT. Each leaf node is labelled according to its class and linked to the root of the tree via internal nodes. A DT's root node serves as its beginning point, and the route from this node to its leaf nodes produces the classification rules [48].
- LR is an excellent method for predicting the probability of a result in a variety of classification situations. Commonly, the LR model is used when people can make predictions about health or illness. The LR algorithm predicts the probability of the target category dependent variable by applying the training examples to a logistic sigmoid activation function. In LR, the target attribute's calculated probability ranges from 0 to 1. Additionally, a threshold is established to classify an event into a certain target class. The predicted probability is input into a certain target category based on the threshold value [49,50].
- The SVM [51,52] is a type of linear generalized classifier that sorts binary data using supervised learning. It is appropriate for small data collections with minimal outliers. The goal is to identify a hyperplane that can be used to connect data points. This hyperplane divides the space into separate domains, each of which holding different

kinds of data. There are numerous hyperplanes from which to choose to split the two groups of data. Our objective was to find the plane with the largest margin. The margin is the distance between the hyperplane and two data points that are closest to it that represent two subclasses. The SVM attempts to optimize the algorithm by increasing this margin value, thereby determining the optimal superplane to divide the dataset into two layers. The nearest data points to the hyperplane are referred to as support vectors.

- GB is a prominent supervised ML method for disease forecasting since it creates an ensemble forecasting model using weak classifiers based on a DT. It constructs DTs using a gradient decent iterative optimization technique to discover the best parameter values, unlike RF. Then, we use the weighted majority votes from each DT to forecast the predicted value [53,54].
- XGBoost [55,56] constructs multiple new algorithms and merges them into a single ensemble model. First, the inaccuracy the of residuals for every observation is determined based on an established model. Based on previous errors, a revised model is developed to predict the residuals. The predictions of this model are then incorporated into the ensemble models. XGBoost is superior to GB algorithms because it finds a balance between bias and variation.
- RF is an ML algorithm that uses a random subspace approach and bagging ensemble learning. In the training stage, RF builds several DTs for arbitrarily partitioning data. For each node in the root DT, a subset of K attributes is chosen at random from the node's attribute set. From this subset, an effective attribute is then chosen for partitioning. Each tree submits a classification as a vote for the other trees, and the RF selects the classification with the most votes [29,57].
- ET is a procedure for data mining that combines multiple methods into a single optimal predictive model to improve predictions. This technique provides superior predictive performance when compared to a single model. We combined DT, XGB, and RF to benefits from all the algorithms to improve the overall predictive performance. By uniting the strengths of multiple models, ETs can provide enhanced generalization, increased robustness, and enhanced precision. It can help reduce individual model biases and improve model performance overall [58].

2.9. Model Evaluation

Employing accuracy, precision, recall, ROC-AUC, F1-score, geometric mean (GM), and log-loss, we assessed the ML classifiers. The entire dataset was divided into 10 parts using a 10-fold cross-validation technique. One part was utilized for model testing, while the other parts were employed for training the model in each fold. The evaluation procedure was performed 10 times [59]. Accuracy provides an accurate rating of the categorization. It is calculated as the ratio of the summation of the true positive (TPS) and true negative (TNG) in the whole population [60]. Precision is the percentage of predicted positives that are real [61]. Recall measures the models ability to categorize samples inside a class [62]. The F1-score maintains a balance between the classifier's precision and recall [63]. The log-loss computes the ambiguity of the method's probability by evaluating its exact labels [64]. A lower log-loss number suggests a more accurate forecast. ROC-AUC demonstrates the link between sensitivity and specificity as well as reflecting the model's capacity for discrimination. TPSs are those in which the model correctly recognizes the positive class. A TNG is a result which the model forecasts the negative class properly. When the model wrongly predicts the positive class, a false positive (FPS) is generated. False negatives (FNG) occur when models incorrectly predict the negative class.

$$Accuracy = \frac{TPS + TNG}{TPS + TNG + FPS + FNG} \tag{3}$$

$$Precision = \frac{TPS}{TPS + FPS} \tag{4}$$

$$Recall = \frac{TPS}{FNG + TPS} \tag{5}$$

$$FS = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

$$Log - loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i, \log(p(y_i)) + (1 - y_i)\log(1 - \log(p(y_i))) \tag{7}$$

$$GM = \sqrt{\frac{TPS}{TPS + FNG} * \frac{TNG}{TNG + FPS}} \tag{8}$$

3. Results and Discussion

We implemented seven ML models: DT, LR, SVM, GB, XGBoost, RF, and custom-built ET on the diabetes dataset. First, we discuss the experimental setup and hardware configuration. After this, we present the implementation output of each model with a comparative analysis. Finally, we provide a detailed discussion of our detection system.

3.1. Experimental Setup

We used the Google Co-laboratory platform, which provides the Jupyter Notebook Python language editor (version 3.7.13) and offers many built-in Python modules and packages through which we applied every ML model. In addition, we generated every plot and figure using 'matplotlib' in Python and 'ggplot2' within the R language. To determine the significance of a variable for the DT, GB, XGB, RF, and ET methods, we used the 'feature_importance_' approach; for SVM and LR, we used the 'coef_' technique. Various evaluation indicators were employed to assess their performance. We used a 10-fold cross-validation method on the dataset to produce a more reliable detection system.

3.2. Result Analysis

The experiment outcome for the primary dataset is displayed in Table 2. All classifiers had an accuracy between 80 and 90%. However, ET provided the best results with an accuracy of 87.60%. Then, GB, XGB, SVM, LR, and DT successively delivered the best results.

Table 2. Performance analysis of the various classifiers using the main diabetes dataset.

Classifier	DT	LR	SVM	XGB	GB	RF	ET
Accuracy	0.813	0.8386	0.8602	0.8661	0.8681	0.874	0.876
Precision	0.8409	0.8459	0.8864	0.8791	0.8741	0.8809	0.8926
Recall	0.8073	0.8582	0.8509	0.8727	0.8836	0.8873	0.8764
ROC-AUC	0.8135	0.8368	0.8611	0.8655	0.8667	0.8728	0.8760
F1-Score	0.8237	0.852	0.8683	0.8759	0.8788	0.8841	0.8844
Geometric Mean	0.8135	0.8365	0.861	0.8655	0.8665	0.8727	0.8759
Log-Loss	6.7404	5.8181	5.0376	4.8247	4.7538	4.5409	4.47

The classification results for the balanced dataset using SMOTE are subsequently displayed in Table 3. ET had the highest accuracy of 87.45%, precision of 87.05%, recall of 88%, ROC-AUC of 0.8745, F1-score of 87.52%, GM of 87.45%, and the lowest log-loss of 4.5218 in this scenario. DT, on the other hand, yielded the lowest results across all evaluation metrics. Therefore, the results demonstrate that GB, XGB, and SVM produce are clearly better than DT and LR.

Table 3. Performance analysis of the different classifiers using SMOTE.

Classifier	DT	LR	SVM	XGB	GB	RF	ET
Accuracy	0.8018	0.8509	0.8655	0.8691	0.8673	0.8636	0.8745
Precision	0.8029	0.8561	0.8708	0.8664	0.8686	0.8676	0.8705
Recall	0.8	0.8436	0.8582	0.8727	0.8655	0.8582	0.88
ROC-AUC	0.8018	0.8509	0.8655	0.8691	0.8673	0.8636	0.8745
F1-Score	0.8015	0.8498	0.8645	0.8696	0.867	0.8629	0.8752
Geometric Mean	0.8018	0.8509	0.8654	0.8691	0.8673	0.8636	0.8745
Log-Loss	7.1432	5.3738	4.8495	4.7184	4.784	4.915	4.5218

Table 4 shows the results of different classifiers in the balanced dataset using the ROS technique. In this scenario, ET had the best accuracy of 89.27%, precision of 89.71%, recall of 88.73%, ROC-AUC of 0.8927, F1-score of 89.21%, GM of 89.27%, and the lowest log-loss of 3.8665. DT, on the other hand, produced the worst results across all the evaluation metrics. XGB and RF all produced results that were very close for all evaluation metrics.

Table 4. Performance analysis of the different classifiers using the ROS technique.

Classifier	DT	LR	SVM	XGB	GB	RF	ET
Accuracy	0.8473	0.8527	0.8691	0.8873	0.8764	0.8891	0.8927
Precision	0.8577	0.8514	0.883	0.8959	0.8906	0.8934	0.8971
Recall	0.8327	0.8545	0.8509	0.8764	0.8582	0.8836	0.8873
ROC-AUC	0.8473	0.8527	0.8691	0.8873	0.8764	0.8891	0.8927
F1-Score	0.845	0.853	0.8667	0.886	0.8741	0.8885	0.8921
Geometric Mean	0.8471	0.8527	0.8689	0.8872	0.8762	0.8891	0.8927
Log Loss	5.5048	5.3082	4.7184	4.0631	4.4563	3.9976	3.8665

Table 5 depicts the outcomes of several classifiers using 20 significant features. All of them had an accuracy greater than 80%. ET displayed a maximum accuracy of 88.18%, precision of 89.77%, ROC-AUC of 0.8818, F1-score of 87.94%, GM of 88.16%, and a minimum log-loss of 3.8665. On the other hand, SVM generated the highest recall of 87.27% in comparison with other classifiers. However, XGB produced the second-best outcome. Furthermore, the other classifiers, such as SVM, GB, and RF, also produced excellent outcomes.

Table 5. Classification results (evaluation metrics) with 20 important features.

Classifier	DT	LR	SVM	XGB	GB	RF	ET	ET (Tuning)
Accuracy	0.8364	0.8545	0.8673	0.8782	0.8655	0.8764	0.8818	0.9927
Precision	0.8627	0.847	0.8633	0.891	0.8764	0.8848	0.8977	1
Recall	0.8	0.8655	0.8727	0.8618	0.8509	0.8655	0.8618	0.9855
ROC-AUC	0.8364	0.8545	0.8673	0.8782	0.8655	0.8764	0.8818	0.9927
F1-Score	0.8302	0.8561	0.868	0.8762	0.8635	0.875	0.8794	0.9927
Geometric Mean	0.8356	0.8545	0.8673	0.878	0.8653	0.8763	0.8816	0.9927
Log-Loss	5.8981	5.2427	4.784	4.3908	4.8495	4.4563	4.2597	0.2621

In addition, we optimized the hypeparameters of the ET algorithm for the highest performance. ET exhibited 99.27% accuracy, 100% precision, 98.55% recall, 0.9927 ROC-AUC, 99.27% F1-score, 99.27% GM, and 0.2621 log-loss.

Figure 4 illustrates the associative features with T2DM using $-log_{10}(P)$. The $-log_{10}(P)$ function changes the p value into a range of positive numbers that can be used to make good decisions for each feature. If the value of $-log_{10}(P)$ is greater than 1.301, then the feature is considered significant. Furthermore, a high $-log_{10}(P)$ value indicates a highly significant feature. The results show that age is the most significant element, whereas smoking is the least significant component. Other significant factors include extreme thirst, gender, regular intake of medicine, and having diabetes in the family; while stress and living house type are less significant.

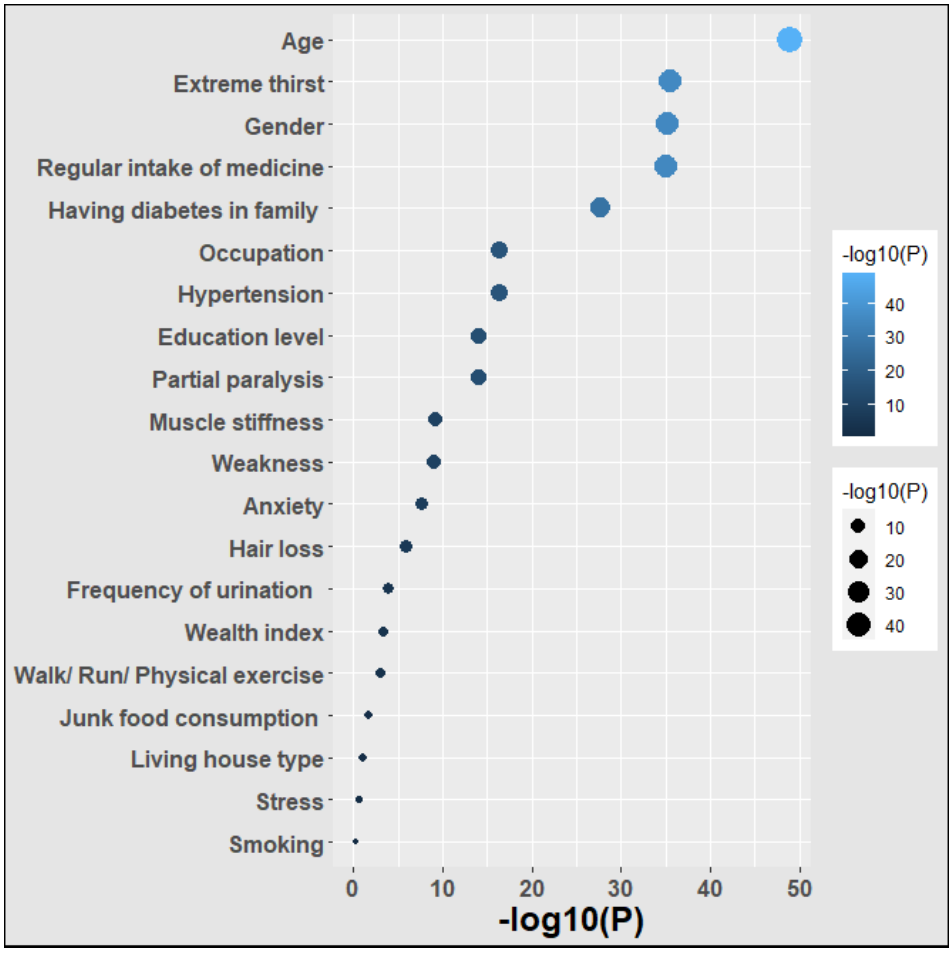

Figure 4. Significance of the features, *p* values with negative 10 base logarithm. The lighter and larger bubbles represent more significance.

Our research also revealed the significance of features, estimated based on the mean coefficient value of every classifier employed. We calculated feature significance values for every method and then normalized them by applying the min–max technique to ensure that they ranged between 0 and 1. Next, we determined the mean scores for every characteristic. In Table 6, we examined the significance of the diabetic features and found that age was the most significant attribute (average coefficient value 0.99). Other crucial elements included having diabetes in the family, regular intake of medicine, extreme thirst, etc. The least important characteristics included muscle stiffness, living house type, stress, wealth index, etc. The ROC curves [65] for each classification method are shown in Figure 5. We observed that the ET classifier performed better than the rest of the classifiers in the ROC curves.

Table 6. Feature ranking using ML techniques based on coefficient values.

Feature Name	DT	SVM	LR	RF	XGB	GB	ET	Avg.	Rank
Age	1	0.95	1	1	1	1	1	0.99	1
Having diabetes in family	0.26	1	0.94	0.45	0.34	0.23	0.23	0.49	2
Regular intake of medicine	0.08	0.89	0.79	0.48	0.13	0.2	0.19	0.39	3
Extreme thirst	0.08	0.82	0.75	0.39	0.15	0.13	0.19	0.36	4
Occupation	0.06	0.77	0.67	0.17	0.03	0.04	0.06	0.26	5
Frequency of urination	0.04	0.66	0.57	0.06	0.05	0.04	0.04	0.21	6
Walk/ Run/ Physically exercise	0.06	0.62	0.56	0.13	0.01	0.02	0.03	0.21	6

Table 6. *Cont.*

Feature Name	DT	SVM	LR	RF	XGB	GB	ET	Avg.	Rank
Weakness	0.06	0.57	0.58	0.11	0.01	0.02	0.05	0.2	7
Smoking	0.05	0.67	0.66	0	0.01	0	0.01	0.2	7
Junk food consumption	0.05	0.64	0.6	0.02	0.03	0	0.01	0.19	8
Partial paralysis	0	0.6	0.58	0.06	0.05	0.02	0.03	0.19	8
Education level	0.08	0.36	0.37	0.22	0.02	0.04	0.09	0.17	9
Hypertension	0.01	0.51	0.53	0.09	0	0.01	0.02	0.17	9
Gender	0.18	0	0	0.48	0.21	0.14	0.21	0.17	9
Hair loss	0.05	0.43	0.38	0.12	0.01	0.01	0.04	0.15	10
Anxiety	0.02	0.43	0.37	0.05	0.01	0	0.02	0.13	11
Wealth index	0.01	0.34	0.33	0.04	0.02	0	0.02	0.11	12
Stress	0.04	0.25	0.25	0.1	0.03	0.02	0.04	0.11	12
Living house type	0.02	0.38	0.34	0	0	0	0	0.11	12
Muscle stiffness	0.04	0.26	0.27	0.06	0.01	0.01	0.05	0.1	13

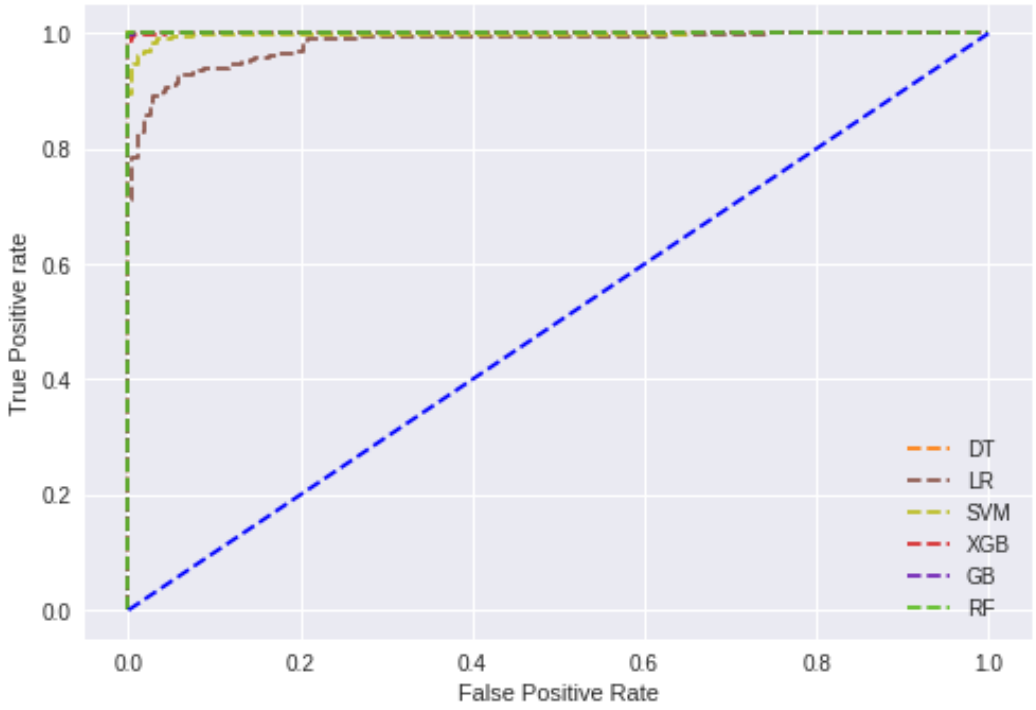

Figure 5. Comparison of the ROC curves obtained by using the seven ML classifiers.

3.3. Discussion

The detection of diabetes may play a crucial role in the management of this disease. Initially, we preprocessed the dataset and then normalized the entire dataset using scaling techniques. We independently applied statistical and ML algorithms to the dataset. In the statistical study, the most relevant characteristics were found, but in the ML classification approaches, the patients were divided into diabetic and non-diabetic. We ordered the characteristics of diabetes according to their significance.

ML techniques are widely accepted as a way to display disease-related characteristics as distinguishing indicators in predicting disease diagnoses such as diabetes [66–68]. The potential of ML methods to uncover hidden trends in data by analysing a set of attributes

might result in a deeper comprehension. The classification results with a high degree of precision imply a reliable prognosis and ensure practical applicability. The majority of models studied here are capable of making correct predictions since their accuracy, precision, recall, ROC-AUC, F1-score, and GM were all greater than 80%. Excellent model performance was shown by a small log-loss value in binary classification. In comparison to other models, the ET model reached the highest degree of accuracy.

Our findings imply a number of crucial and relevant characteristics. Depending on the log-based association, the most significant characteristics include age, extreme thirst, and gender. In the ML models, the most essential features are age, having diabetes in the family, regular intake of medicine, and extreme thirst. Our analysis reveals that significant features are adequate for detecting diabetes, and this will aid in the implementation of diagnosing diabetes.

We also compared our findings to prior research based on ML techniques, represented in Table 7. Pranto et al. [31] utilized female diabetic patients from Bangladesh to predict an accuracy of 81.2%, precision of 80%, and an F1-score of 88%. In another study, Syed and Khan [69] developed a data-driven predictive model to screen for T2DM in the western region of Saudi Arabia, achieving 82.1% accuracy, 77.6% precision, 89% recall, 0.867 ROC-AUC, and 82.9% F1-score. Next, Chou et al. [70] predicted the onset of diabetes using ML methods, achieving 95.3% accuracy, 92.7% precision, 93.1% recall, 0.991 ROC-AUC, and 92.9% F1-score. Then, Laila et al. [71] used an ensemble-based ML model to predict diabetes with an accuracy of 97.11%, precision of 97.1%, recall of 97.1%, and an F1-score of 97.1%. In contrast, our proposed framework outperformed the previous studies, achieving an accuracy of 99.27%, precision of 100%, ROC-AUC of 0.9927, and an F1-score of 99.27%. Additionally, we computed a GM of 99.27% and a log-loss of 0.2621. We also determined the most significant characteristic using both log-based correlations and ML models. It is important to note that this study was conducted with a limited sample size, restricting the generalizability of the findings. Nevertheless, this approach shows promise as a predictive method for real-world applications and could greatly assist practitioners in prompt diabetes diagnosis.

Table 7. Comparing the performance of the proposed model with existing studies.

Reference	Dataset	Accuracy	Precision	Recall	ROC-AUC	F1-Score	Geometric Mean	Log-Loss
Pranto et al. [31]	PIMA, Kurmitola Hospital, Dhaka	0.8120	0.8	1	0.84	0.88	–	–
Syed and Khan [69]	Western Region of Saudi Arabia	0.821	0.776	0.89	0.867	0.829	–	–
Chou et al. [70]	Taipei Municipal Medical Center	0.953	0.927	0.931	0.991	0.929	–	–
Laila et al. [71]	UCI Repository	0.9711	0.971	0.971	–	0.971	–	–
This study	Bangladesh, 2022	0.9927	1	0.9855	0.9927	0.9927	0.9927	0.2621

The proposed framework for diabetes detection, utilizing an ensemble-based ML approach, offers several advantages compared to the existing gold standard methods. This is particularly crucial in Bangladesh, where there is a shortage of diabetologists relative to the country's population. While the gold standard relies on costly and time-consuming lab-based diagnostics and manual interpretation of clinical data by diabetologists, our automated approach can help alleviate the burden on healthcare providers, allowing them to efficiently screen a larger number of individuals for diabetes risk. Through the application of statistical and ML techniques, this study identified age, extreme thirst, and family history of diabetes as key features that play instrumental roles in diabetes detection. By implementing the proposed framework in clinical settings, healthcare providers can proactively identify individuals at risk of diabetes based on their risk factors. This prompt identification enables timely intervention and personalized care, helping to mitigate the risks and severity associated with diabetes.

4. Conclusions

Our proposed ET with hyperparameter tuning outperformed the other ML models to identify T2DM patients in this region. We obtained an outstanding accuracy of 99.27% and an F1-score of 99.27%. In addition, using various statistical and ML models we determined four key factors: age, having diabetes in the family, regular intake of medicine, and extreme thirst, highly associated with diabetes. Since our proposed diabetes detection system has a high degree of precision, physicians and clinicians may use our proposed framework to assess diabetes risk. In Bangladesh, the ratio of diabetic patients to physicians is insufficient, and there are less detection instruments available across the country. Therefore, a fast diagnosis system based on ML is effective for patients and diabetologists. In this study, we found that a number of lifestyle factors are associated with the development of diabetes; therefore, sustaining a lifestyle that includes the observance of these factors could reduce the rate of diabetes progression. The quantity of data was a primary limitation in our study. In the future, we will use additional data to investigate diabetes as well as other medical conditions such as kidney disease, heart disease, and breast cancer. We will also incorporate our proposed system into intuitive web and mobile application platforms. Finally, we will concentrate on diabetes mellitus disease prevention and recovery strategies.

Author Contributions: Conceptualization, M.J.U. and M.M.A.; methodology, M.J.U.; software, M.J.U. and M.M.A.; validation, M.J.U., M.A.K. and M.M.A.; formal analysis, M.J.U., S.A. and M.M.A.; investigation, M.J.U., N.A, S.A.A. and M.M.A.; resources, M.J.U., N.A, and S.A.A.; data curation, M.J.U.; writing—original draft preparation, M.J.U., S.A., M.N.H., M.A.A.W., N.A. and M.M.A.; writing—review and editing, S.A.A., N.A, M.A.K. and M.A.M.; visualization, M.J.U., M.A.A.W., M.A.K. and M.M.A.; supervision, M.A.M.; project administration, S.A.A., N.A and M.A.M.; funding acquisition, S.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding and supporting this work through Research Partnership Program no RP-21-09-09.

Informed Consent Statement: Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: Data will be available only for research purposes. Please email to the corresponding author for further information.

Acknowledgments: The authors would like to thank all the participants who participated in this survey. Furthermore, we also give thanks to Bangabandhu Sheikh Mujibur Rahman Science and Technology University Research Cell for supporting the data collection.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine learning
DT	Decision tree
RF	Random forest
LR	Logistic regression
SVM	Support vector machine
GB	Gradient boosting

References

1. Association, A.D. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **2014**, *37*, S81–S90. [CrossRef] [PubMed]
2. IDF. Type 2 Diabetes. Available online: https://www.idf.org/aboutdiabetes/type-2-diabetes.html (accessed on 7 May 2023).
3. John, J.E.; John, N.A. Imminent risk of COVID-19 in diabetes mellitus and undiagnosed diabetes mellitus patients. *Pan Afr. Med. J.* **2020**, *32874422*. [CrossRef] [PubMed]
4. Gahlan, D.; Rajput, R.; Singh, V. Metabolic syndrome in north indian type 2 diabetes mellitus patients: A comparison of four different diagnostic criteria of metabolic syndrome. *Diabetes Metab. Syndr.* **2019**, *13*, 356–362. [CrossRef] [PubMed]

5. Atlas, I.D. Diabetes around the World in 2021. Available online: https://diabetesatlas.org/ (accessed on 7 May 2023).
6. Williams, R.; Karuranga, S.; Malanda, B.; Saeedi, P.; Basit, A.; Besançon, S.; Bommer, C.; Esteghamati, A.; Ogurtsova, K.; Zhang, P.; et al. Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res. Clin. Pract.* **2020**, *162*, 108072. [CrossRef] [PubMed]
7. Htay, T.; Soe, K.; Lopez-Perez, A.; Doan, A.H.; Romagosa, M.A.; Aung, K. Mortality and cardiovascular disease in type 1 and type 2 diabetes. *Curr. Cardiol. Rep.* **2019**, *21*, 45. [CrossRef]
8. Nipa, N.; Riyad, M.M.H.; Satu, M.S.; Walliullah, M.; Howlader, K.C.; Moni, M.A. Clinically Adaptable Machine Learning Model To Identify Early Appreciable Features of Diabetes In Bangladesh. *Intell. Med.* **2023**. [CrossRef]
9. Huang, Y.; Roy, N.; Dhar, E.; Upadhyay, U.; Kabir, M.A.; Uddin, M.; Tseng, C.L.; Syed-Abdul, S. Deep Learning Prediction Model for Patient Survival Outcomes in Palliative Care Using Actigraphy Data and Clinical Information. *Cancers* **2023**, *15*, 2232. [CrossRef]
10. Panday, A.; Kabir, M.A.; Chowdhury, N.K. A survey of machine learning techniques for detecting and diagnosing COVID-19 from imaging. *Quant. Biol.* **2022**, *10*. [CrossRef]
11. Uddin, M.J.; Ahamad, M.M.; Sarker, P.K.; Aktar, S.; Alotaibi, N.; Alyami, S.A.; Kabir, M.A.; Moni, M.A. An Integrated Statistical and Clinically Applicable Machine Learning Framework for the Detection of Autism Spectrum Disorder. *Computers* **2023**, *12*, 92. [CrossRef]
12. Hossain, M.D.; Kabir, M.A.; Anwar, A.; Islam, M.Z. Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets. *Health Inf. Sci. Syst.* **2021**, *9*, 17. [CrossRef]
13. Aguilera-Venegas, G.; López-Molina, A.; Rojo-Martínez, G.; Galán-García, J.L. Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *J. Comput. Appl. Math.* **2023**, *427*, 115115. [CrossRef]
14. Zhao, M.; Wan, J.; Qin, W.; Huang, X.; Chen, G.; Zhao, X. A machine learning-based diagnosis modelling of type 2 diabetes mellitus with environmental metal exposure. *Comput. Methods Programs Biomed.* **2023**, *235*, 107537. [CrossRef] [PubMed]
15. Xia, S.; Zhang, Y.; Peng, B.; Hu, X.; Zhou, L.; Chen, C.; Lu, C.; Chen, M.; Pang, C.; Dai, Y.; et al. Detection of mild cognitive impairment in type 2 diabetes mellitus based on machine learning using privileged information. *Neurosci. Lett.* **2022**, *791*, 136908. [CrossRef] [PubMed]
16. Ejiyi, C.J.; Qin, Z.; Amos, J.; Ejiyi, M.B.; Nnani, A.; Ejiyi, T.U.; Agbesi, V.K.; Diokpo, C.; Okpara, C. A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. *Healthc. Anal.* **2023**, *3*, 100166. [CrossRef]
17. Hennebelle, A.; Materwala, H.; Ismail, L. HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System. *Procedia Comput. Sci.* **2023**, *220*, 331–338. [CrossRef]
18. Haque, M.; Alharbi, I. A Dataset-Specific Machine Learning Study for Predicting Diabetes (Type-2) in a Developing Country Context. *Indian J. Sci. Technol.* **2022**, *15*, 1932–1940. [CrossRef]
19. Tasin, I.; Nabil, T.U.; Islam, S.; Khan, R. Diabetes prediction using machine learning and explainable AI techniques. *Healthc. Technol. Lett.* **2022**, 1684017. [CrossRef]
20. Kaur, H.; Kumari, V. Predictive modelling and analytics for diabetes using a machine learning approach. *Appl. Comput. Inform.* **2022**, *18*, 90–100. [CrossRef]
21. Zheng, T.; Xie, W.; Xu, L.; He, X.; Zhang, Y.; You, M.; Yang, G.; Chen, Y. A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int. J. Med. Inform.* **2017**, *97*, 120–127. [CrossRef]
22. Saha, P.K.; Patwary, N.S.; Ahmed, I. A widespread study of diabetes prediction using several machine learning techniques. In Proceedings of the 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 18–20 December 2019; IEEE: Piscataway NJ, USA, 2019; pp. 1–5.
23. Ali, M.S.; Islam, M.K.; Das, A.A.; Duranta, D.; Haque, M.; Rahman, M.H. A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights. *Biomed Res. Int.* **2023**, 8583210. [CrossRef]
24. Howlader, K.C.; Satu, M.S.; Awal, M.A.; Islam, M.R.; Islam, S.M.S.; Quinn, J.M.; Moni, M.A. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. *Health Inf. Sci. Syst.* **2022**, *10*, 2. [CrossRef] [PubMed]
25. Wei, S.; Zhao, X.; Miao, C. A comprehensive exploration to the machine learning techniques for diabetes identification. In Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, 5–8 February 2018; IEEE: Piscataway NJ, USA, 2018; pp. 291–295.
26. Birjais, R.; Mourya, A.K.; Chauhan, R.; Kaur, H. Prediction and diagnosis of future diabetes risk: A machine learning approach. *SN Appl. Sci.* **2019**, *1*, 1112. [CrossRef]
27. Yahyaoui, A.; Jamil, A.; Rasheed, J.; Yesiltepe, M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In Proceedings of the 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 6–7 November 2019; IEEE: Piscataway NJ, USA, 2019; pp. 1–4.
28. Abdulhadi, N.; Al-Mousa, A. Diabetes detection using machine learning classification methods. In Proceedings of the 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 14–15 July 2021; IEEE: Piscataway NJ, USA, 2021; pp. 350–354.
29. Battineni, G.; Sagaro, G.G.; Nalini, C.; Amenta, F.; Tayebati, S.K. Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods. *Machines* **2019**, *7*, 74. [CrossRef]

30. Tigga, N.P.; Garg, S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Comput. Sci.* **2020**, *167*, 706–716. [CrossRef]
31. Pranto, B.; Mehnaz, S.M.; Mahid, E.B.; Sadman, I.M.; Rahman, A.; Momen, S. Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information* **2020**, *11*, 374. [CrossRef]
32. Sneha, N.; Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data* **2019**, *121*, 54–64. [CrossRef]
33. Zhang, L.; Wang, Y.; Niu, M.; Wang, C.; Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. *Sci. Rep.* **2020**, *10*, 4406. [CrossRef]
34. Bonifazi, G.; Enrico Corradini, D.U.; Virgili, L. Defining user spectra to classify Ethereum users based on their behavior. *J. Big Data* **2022**, *9*, 37. [CrossRef]
35. Le, T.M.; Vo, T.M.; Pham, T.N.; Dao, S.V.T. A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic. *IEEE Access* **2020**, *9*, 7869–7884. [CrossRef]
36. Islam, M.M.; Rahman, M.J.; Roy, D.C.; Maniruzzaman, M. Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**, *14*, 217–219. [CrossRef]
37. Haq, A.U.; Li, J.P.; Khan, J.; Memon, M.H.; Nazir, S.; Ahmad, S.; Khan, G.A.; Ali, A. Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. *Sensors* **2020**, *20*, 2649. [CrossRef] [PubMed]
38. Shuja, M.; Mittal, S.; Zaman, M. Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. In *Proceedings of the Advances in Computing and Intelligent Systems*; Sharma, H., Govindan, K., Poonia, R.C., Kumar, S., El-Medany, W.M., Eds.; Springer: Singapore, 2020; pp. 195–211.
39. Chatrati, S.P.; Hossain, G.; Goyal, A.; Bhan, A.; Bhattacharya, S.; Gaurav, D.; Tiwari, S.M. Smart home health monitoring system for predicting type 2 diabetes and hypertension. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 862–870. [CrossRef]
40. Islam, M.M.; Rahman, M.J.; Menhazul Abedin, M.; Ahammed, B.; Ali, M.; Ahmed, N.F.; Maniruzzaman, M. Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques. *Health Syst.* **2022**, *12*, 243–254. [CrossRef] [PubMed]
41. Islam, S.M.S.; Islam, M.T.; Uddin, R.; Tansi, T.; Talukder, S.; Sarker, F.; Mamun, K.A.A.; Adibi, S.; Rawal, L.B. Factors associated with low medication adherence in patients with Type 2 diabetes mellitus attending a tertiary hospital in Bangladesh. *Lifestyle Med.* **2021**, *2*, e47. [CrossRef]
42. Nnamoko, N.; Korkontzelos, I. Efficient treatment of outliers and class imbalance for diabetes prediction. *Artif. Intell. Med.* **2020**, *104*, 101815. [CrossRef] [PubMed]
43. Ganie, S.M.; Malik, M.B. An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators. *Healthc. Anal.* **2022**, *2*, 100092. [CrossRef]
44. Petmezas, G.; Haris, K.; Stefanopoulos, L.; Kilintzis, V.; Tzavelis, A.; Rogers, J.A.; Katsaggelos, A.K.; Maglaveras, N. Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. *Biomed. Signal Process. Control* **2021**, *63*, 102194. [CrossRef]
45. Mehedi Hassan, M.; Mollick, S.; Yasmin, F. An unsupervised cluster-based feature grouping model for early diabetes detection. *Healthc. Anal.* **2022**, *2*, 100112. [CrossRef]
46. Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3317. [CrossRef]
47. Aktar, S.; Ahamad, M.M.; Rashed-Al-Mahfuz, M.; Azad, A.; Uddin, S.; Kamal, A.; Alyami, S.A.; Lin, P.I.; Islam, S.M.S.; Quinn, J.M.; et al. Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: Statistical analysis and model development. *JMIR Med. Inform.* **2021**, *9*, e25884. [CrossRef]
48. Azad, C.; Bhushan, B.; Sharma, R.; Shankar, A.; Singh, K.K.; Khamparia, A. Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. *Multimed. Syst.* **2022**, *28*, 1289–1307. [CrossRef]
49. Maniruzzaman, M.; Rahman, M.; Al-MehediHasan, M.; Suri, H.S.; Abedin, M.; El-Baz, A.; Suri, J.S. Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. *J. Med. Syst.* **2018**, *42*, 92. [CrossRef]
50. Ahlqvist, E.; Storm, P.; Käräjämäki, A.; Martinell, M.; Dorkhan, M.; Carlsson, A.; Vikman, P.; Prasad, R.B.; Aly, D.M.; Almgren, P.; et al. Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* **2018**, *6*, 361–369. [CrossRef]
51. Boubin, M.; Shrestha, S. Microcontroller implementation of support vector machine for detecting blood glucose levels using breath volatile organic compounds. *Sensors* **2019**, *19*, 2283. [CrossRef] [PubMed]
52. Muhammad, L.; Algehyne, E.A.; Usman, S.S. Predictive supervised machine learning models for diabetes mellitus. *SN Comput. Sci.* **2020**, *1*, 240. [CrossRef] [PubMed]
53. Islam, M.M.; Rahman, M.J.; Roy, D.C.; Tawabunnahar, M.; Jahan, R.; Ahmed, N.F.; Maniruzzaman, M. Machine learning algorithm for characterizing risks of hypertension, at an early stage in Bangladesh. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2021**, *15*, 877–884. [CrossRef] [PubMed]
54. Ahamad, M.M.; Aktar, S.; Rashed-Al-Mahfuz, M.; Uddin, S.; Liò, P.; Xu, H.; Summers, M.A.; Quinn, J.M.; Moni, M.A. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. *Expert Syst. Appl.* **2020**, *160*, 113661. [CrossRef] [PubMed]

55. Dutta, A.; Hasan, M.K.; Ahmad, M.; Awal, M.A.; Islam, M.A.; Masud, M.; Meshref, H. Early prediction of diabetes using an ensemble of machine learning models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12378. [CrossRef]
56. Kibria, H.B.; Nahiduzzaman, M.; Goni, M.O.F.; Ahsan, M.; Haider, J. An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI. *Sensors* **2022**, *22*, 7268. [CrossRef] [PubMed]
57. Ijaz, M.F.; Alfian, G.; Syafrudin, M.; Rhee, J. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. *Appl. Sci.* **2018**, *8*, 1325. [CrossRef]
58. Amelio, A.; Bonifazi, G.; Corradini, E.; Di Saverio, S.; Marchetti, M.; Ursino, D.; Virgili, L. Defining a deep neural network ensemble for identifying fabric colors. *Appl. Soft Comput.* **2022**, *130*, 109687. [CrossRef]
59. Islam, S.M.S.; Talukder, A.; Awal, M.A.; Siddiqui, M.M.U.; Ahamad, M.M.; Ahammed, B.; Rawal, L.B.; Alizadehsani, R.; Abawajy, J.; Laranjo, L.; et al. Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. *Front. Cardiovasc. Med.* **2022**, *9*, 839379. [CrossRef] [PubMed]
60. Akter, T.; Ali, M.H.; Khan, M.I.; Satu, M.S.; Uddin, M.J.; Alyami, S.A.; Ali, S.; Azad, A.; Moni, M.A. Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. *Brain Sci.* **2021**, *11*, 734. [CrossRef] [PubMed]
61. Ahamad, M.M.; Aktar, S.; Uddin, M.J.; Rahman, T.; Alyami, S.A.; Al-Ashhab, S.; Akhdar, H.F.; Azad, A.; Moni, M.A. Early-Stage Detection of Ovarian Cancer Based on Clinical Data Using Machine Learning Approaches. *J. Pers. Med.* **2022**, *12*, 1211. [CrossRef]
62. Ahamad, M.M.; Aktar, S.; Uddin, M.J.; Rashed-Al-Mahfuz, M.; Azad, A.; Uddin, S.; Alyami, S.A.; Sarker, I.H.; Khan, A.; Liò, P.; et al. Adverse effects of COVID-19 vaccination: Machine learning and statistical approach to identify and classify incidences of morbidity and postvaccination reactogenicity. *Healthcare* **2022**, *11*, 31. [CrossRef]
63. Akter, T.; Khan, M.I.; Ali, M.H.; Satu, M.S.; Uddin, M.J.; Moni, M.A. Improved machine learning based classification model for early autism detection. In Proceedings of the 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 5–7 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 742–747.
64. Akter, T.; Satu, M.S.; Khan, M.I.; Ali, M.H.; Uddin, S.; Lio, P.; Quinn, J.M.; Moni, M.A. Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access* **2019**, *7*, 166509–166527. [CrossRef]
65. Xiong, Y.; Lin, L.; Chen, Y.; Salerno, S.; Li, Y.; Zeng, X.; Li, H. Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques. *J. Matern. Fetal Neonatal Med.* **2022**, *35*, 2457–2463. [CrossRef]
66. Olisah, C.C.; Smith, L.; Smith, M. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Comput. Methods Programs Biomed.* **2022**, *220*, 106773. [CrossRef]
67. Wei, H.; Sun, J.; Shan, W.; Xiao, W.; Wang, B.; Ma, X.; Hu, W.; Wang, X.; Xia, Y. Environmental chemical exposure dynamics and machine learning-based prediction of diabetes mellitus. *Sci. Total Environ.* **2022**, *806*, 150674. [CrossRef]
68. Rawat, V.; Joshi, S.; Gupta, S.; Singh, D.P.; Singh, N. Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study. *Mater. Today Proc.* **2022**, *56*, 502–506.
69. Syed, A.H.; Khan, T. Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: A retrospective cross-sectional study. *IEEE Access* **2020**, *8*, 199539–199561. [CrossRef]
70. Chou, C.Y.; Hsu, D.Y.; Chou, C.H. Predicting the Onset of Diabetes with Machine Learning Methods. *J. Pers. Med.* **2023**, *13*, 406. [CrossRef] [PubMed]
71. Laila, U.E.; Mahboob, K.; Khan, A.W.; Khan, F.; Taekeun, W. An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. *Sensors* **2022**, *22*, 5247. [CrossRef] [PubMed]