sustainability

MDPI

Article

Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique

Israt Jahan Kakoly [1], Md. Rakibul Hoque [2] and Najmul Hasan [3,*]

[1] Institute of Health Economics, University of Dhaka, Dhaka 1000, Bangladesh
[2] Department of Management Information Systems, Faculty of Business Studies, University of Dhaka, Dhaka 1000, Bangladesh
[3] BRAC Business School, BRAC University, Dhaka 1212, Bangladesh
* Correspondence: najmul.hasan@bracu.ac.bd; Tel.: +88-01711-016-143

Abstract: As type 2 diabetes becomes more prevalent across the globe, predicting its sources becomes more important. However, there is a big void in predicting the risk factors of this disease. Thus, the purpose of this study is to predict diabetes risk factors by applying machine learning (ML) algorithms. Two-fold feature selection techniques (i.e., principal component analysis, PCA, and information gain, IG) have been applied to boost the prediction accuracy. Then, the optimal features are fed into five ML algorithms, namely decision tree, random forest, support vector machine, logistic regression, and KNN. The primary data used to train the ML model were collected based on the safety procedure described in the Helsinki Declaration, 2013, and 738 records were included in the final analysis. The result has shown an accuracy level of over 82.2%, with an AUC (area under the ROC curve) value of 87.2%. This research not only identified the most important clinical and nonclinical factors in diabetes prediction, but it also found that the clinical risk factor (glucose) is the most relevant for diabetes prediction, followed by dietary factors. The noteworthy contribution of this research is the identification of previously unclassified factors left over from the previous study that considered both clinical and non-clinical aspects.

Keywords: diabetes; feature selection; risk factors; machine learning

Citation: Kakoly, I.J.; Hoque, M.R.; Hasan, N. Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique. *Sustainability* **2023**, *15*, 4930. https://doi.org/10.3390/su15064930

Academic Editor: Marc A. Rosen

Received: 13 November 2022
Revised: 4 February 2023
Accepted: 28 February 2023
Published: 10 March 2023

1. Introduction

The global burden of type 2 diabetes has been increasing. Diabetes is attributed to many other diseases, such as those of the eye, kidney, heart, and lower limb [1]; it decreases the quality of life and increases premature mortality. Diabetes affects 537 million people (20–79 years old), and it is anticipated that this number will increase to 578 million by the year 2030 and 700 million by the year 2045 [2]. Moreover, around three-quarters of the people with diabetes live in low- and middle-income countries. The high rate of diabetes among the world's population significantly raises the cost of per capita healthcare expenses. Diabetes caused 6.7 million deaths in 2021 [3] and is currently the ninth major cause of mortality [4]. It was estimated that, diabetes will be responsible for 747,000 deaths and will cost $10 billion to treat in the year 2021 [3]. Since the prevalence of diabetic patients has expanded exponentially, early identification of diabetic patients is an essential job. In addition, it is necessary to formulate forecasting models based on the most influential features.

Given the prevalent nature and serious consequences of diabetes, it is critical to explore solutions for the prevention and prediction of this disease. Many studies to date [5–9] have inquired about different factors, such as anthropometric factors (e.g., BMI), socio-demographic factors (e.g., occupation), lifestyle factors (e.g., smoking), and family history (e.g., heredity), as probable causes for diabetes. Findings from studies that looked for a link between diabetes and numerous risk factors have varied due to the use of different methodological approaches. For instance, a study based on participants from eight European countries, [6] has shown that lifestyle, anthropometric, and sociodemographic

factors are only marginally associated with diabetes. Another study in Australia reported that demographic, life quality (e.g., stress), and anthropometric factors were significant for predicting diabetes. Hence, because of the discrepancy in the results of various studies on different populations, contradictory evidence gaps have arisen in the prediction model, and there is a scope for further studies to learn how the above-mentioned factors are associated with diabetes. Diabetes affects the lives of 90 million individuals in Southeast Asia, which accounts for 11% of the population. To be more exact, the prevalence of diabetes among adults is 12.5% in Bangladesh [3]. Due to dietary habits, there might be different influential factors responsible for diabetes in this region than in developed countries. In light of this, more research into the prediction of diabetes risk factors, including dietary habit factors, is required in this area.

Prior studies considered diverse types of factors that are scrutinized when investigating the associations between various factors and the onset of diabetes, especially in predictive studies of diabetes using machine learning (ML). ML is a subset of artificial intelligence (AI) that enables software packages to grow increasingly accurate at predicting events without being precisely designed to do so. One of the major strengths of the ML technique is its ability to identify the uncovered factors that might cause diabetes from a large dataset with enhanced learning capability during training the model. At the same time, while traditional statistical models are unable to handle non-linear data, ML can handle non-linear data and provide robust prediction accuracy. It is used widely in various healthcare research areas, including diabetes risk factor prediction. For example, Zheng et al. [10] used ML algorithms to extract and analyze independent variables from patients' electronic health record (EHR) systems to predict type 2 diabetes. Perveen et al. [11] applied ML to predict diabetes based on the clinical risk factors listed by the NHLBI (National Heart, Lung, and Blood Institute) and AHA (American Heart Association). Narwane and Sawarkar [12] investigated how to improve system performance using ML that can handle unbalanced data for diabetic prediction.

Moreover, a systematic review of ML [13] reported that many studies have looked at different types of factors and several ML approaches to predict diabetes. Most of those studies mostly used different types of factors together to predict diabetes instead of considering dietary factors. Another systematic literature review by Bekele et al. [14] reveals that, among the 28 articles, 39% assessed and listed insufficient physical activity as a risk factor for diabetes, and 42.86% found that insufficient physical activity was a cause of the disease. This fact has prompted the authors to set the objective of this study as predicting risk factors for type 2 diabetes using the ML approach, taking into account dietary factors as well as non-clinical and clinical factors. To bridge this gap, we explored the current literature and proposed 18 initially selected features (Table 1) that include non-clinical factors (i.e., demographic, life quality, and dietary habits) and clinical factors. Therefore, the main objective of this study is to determine the most effective ML model for diabetes risk factor prediction by comparing the performance of several alternative ML models using a variety of assessment measures. Concurrently, several studies on diabetes prediction used either a single, outdated feature selection method (i.e., PCA) [2,15] or a cutting-edge IG method [16]. This study attempted to use both traditional and advanced feature selection techniques on the same dataset to obtain an understanding of which feature selection approach is the most effective. To the best of the author's knowledge, few studies, if any, have applied dual feature selection techniques, which include traditional PCA and advanced IG techniques, in the context of least-developed countries, such as Bangladesh. This allows us to address the following research questions:

RQ1: Does the prediction accuracy of diabetes risk factors improve by using the advanced feature selection technique?
RQ2: Which ML algorithm provides the most accurate and reliable predictions using which features for determining diabetes risk factors?

Table 1. Details of the dataset.

Demographic Features	Value Range	Life Quality and Food Habit Features (Weekly)	Value Range	Clinical Features	Value Range
Area	1–2 (N)	Work stress	1–3 (A)	Systolic blood pressure	96–230 (N)
Age	19–78 (N)	Physical activity	1–2 (N)	Diastolic blood pressure	51–142 (N)
Gender	1–2 (A)	Eating meat	1–3 (A)	Glucose (random)	3–30.5 (N)
Marital status	1–2 (A)	Eating fish	1–3 (A)	Family history of diabetes	1–2 (A)
Education level	1–8 (A)	Eating vegetables	1–3 (A)	Number of pregnancies	0–13 (N)
		Eating fruits	1–3 (A)	BMI	14.1–52.3 (N)
		Soft drinks	1–3 (A)		

Note: N = Numeric, A = Attribute.

This study made use of the data collected from participants in Bangladesh, which is categorized as a least-developed country [17]. While researchers are increasingly trying to predict diabetes and its complications using the ML approach, they all ultimately vie for achieving higher accuracy in prediction with most of the influential features. Therefore, we conducted the feature selection process (Figure 1) carefully for the identification of the most prominent features that can yield a high degree of accuracy in diabetes prediction. Subsequently, we compared all the results based on features selected in different ways. In the process, we also evaluated the performance of different ML algorithms we used. The authors assume that this study will help to make a preliminary judgment about the onset of diabetes based on the factors associated with clinical factors, as well as the demographic, life quality, and dietary habit factors that are changing in a country with fast economic growth.

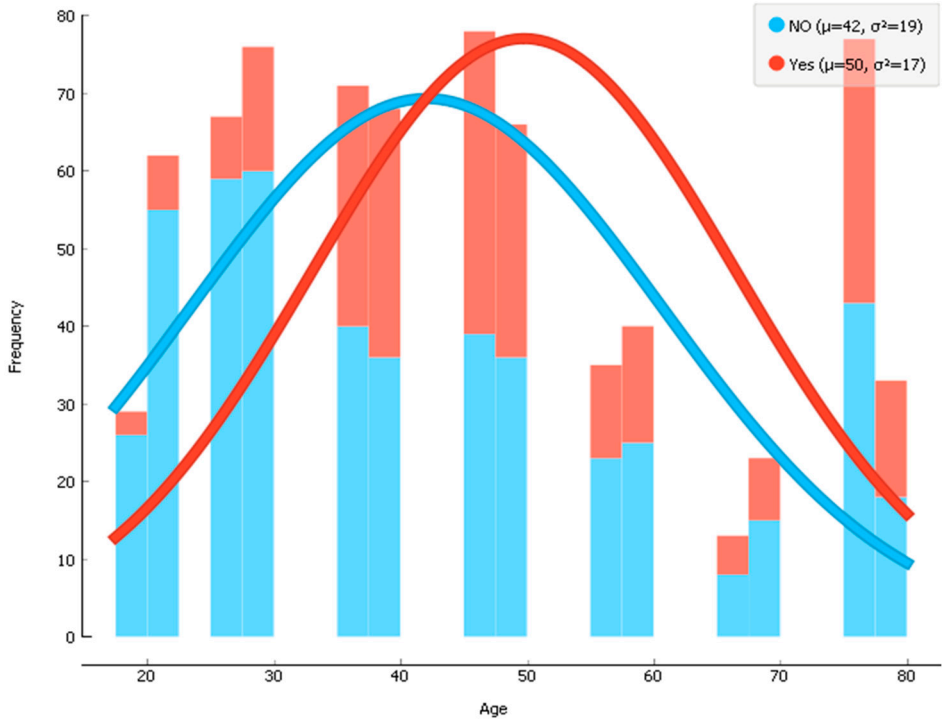

Figure 1. Histogram of the age of participants.

The major contributions of this study are as follows:

- The PCA and IG approaches were both put into practice to investigate the most important potential risk factors of type 2 diabetes, and a multivariable feature analysis was carried out, which lessened the dimensionality of the dataset.

- Since there is a significant correlation between one's dietary habits and their risk of developing diabetes, those with more varied eating patterns are at greater risk for the disease. This is particularly relevant in countries, such as Bangladesh. In this research, a more in-depth investigation was carried out to investigate the influence that dietary routine has on the clinical component in question (i.e., glucose) to better comprehend the clinical factor.
- This study adds an entirely new dataset to the diabetes risk factor classification, and it does so from the perspective of a least-developed country (Bangladesh), which means that it may be used in future research.

The remaining parts of the article are structured as described below. In Section 2, we discuss the works that are linked. In addition, a detailed description of the methods and materials used can be found in Section 3, and the findings can be found in Section 4. In Section 5, we go through the findings of the study that was conducted. Finally, this study concludes the results and presents future directions in Section 6.

2. Related Works

Machine learning has been used widely in the last decade to not only identify the factors contributing to diabetes onset but also predict the likelihood of a patient being diabetic. Apart from predicting diabetes onset, ML is also used to predict diabetes-related complications, such as nephropathy, neuropathy, and retinopathy [1,18]. In diabetes complication studies, researchers have also used psychological data for predicting depression among diabetic patients [19]. Numerous researchers retrieved data from secondary sources, such as electronic medical records (EMR), instead of performing data collection through surveys. Though EMRs contain a large amount of clinical and non-clinical data which is better for predictive analysis, EMRs often contain data on blood lipid profiles, such as triglyceride (TG) level and blood glucose level, that the researchers use for diabetes prediction [20]. On the other hand, a large number of researchers have evaluated the widely used Pima Indian Diabetes Database (PIDD), and those studies have not achieved a satisfactory level of prediction accuracy. Using PIDD as an example, Chatrati et al. [21] achieved 75% accuracy, Jashwanth Reddy et al. [22] attained 80% accuracy, and Goyal and Jain [23] accomplished 77% accuracy. Moreover, there is a dearth of trustworthy and relevant labeled data for diabetes prediction research in least-developed countries, such as Bangladesh [24]. This highlights the need for a trustworthy diabetes dataset from this region that might provide more robust prediction accuracy. Hence, this research is notably different from previous studies in terms of obtaining data from primary sources in a country, such as Bangladesh, which has a high prevalence of diabetes.

In addition, the most commonly used ML algorithms for predicting diabetes are known as decision tree (DT), naïve Bayes (NB), random forest (RF), support vector machine (SVM), logistic regression (LR), k-nearest neighbor (KNN) [25]. These algorithms have emerged from different theories found in the related fields of statistics and probability studies. When applying unsupervised and supervised ML techniques, a study on the application of ML for diabetes identification and its complication prediction shows that about 85% of the studies used a supervised ML approach to predict diabetes [13]. In the practice of diabetes risk factor prediction, many ML methods have been employed. However, there is conflicting evidence about the acquisition of accurate prediction models [26]. This is a need for further attempts in the academic community to study the improved accuracy of diabetes prediction. On the other hand, most of the studies used clinical and non-clinical factors to predict diabetes. however, in the more advanced studies, genetic factors are used to identify the genes contributing early onset of diabetes. For instance, in a more advanced study, Pedersen et al. [27] used ML to find the genomic factors to predict diabetes remission weight-loss surgeries. In addition, the dietary factors that have the greatest influence on diabetes were not included in almost any of the studies at all. This study incorporated dietary factors in order to determine how important they are in order to fill in the gaps that were left by previous studies.

Moreover, many studies have used different types of factors to predict diabetes. For example, age, gender, family history, body mass index (BMI), exercise, and hypertension are commonly used features in applying ML to predict diabetes [28]. While classifying high-risk individuals with diabetes, the World Health Organization (WHO) considers both the clinical (e.g., blood glucose) and non-clinical factors (e.g., demographic) [29]. Furthermore, the living standards or quality is intrinsically related to the increase in diabetic patients, which is increasing evidence in fast-growing countries. Thus, as a part of incorporating diverse types of features for predicting diabetes, both the clinical, non-clinical, and dietary factors were selected in this study. This research also makes a contribution to the existing body of knowledge by including dietary factors that have a significant impact on diabetes, particularly in the South Asian region.

3. Materials and Methods

3.1. Data

This study collected data with a survey questionnaire from 738 participants from different urban and rural areas in Bangladesh. The survey questions were basically three-fold, covering demographic, clinical, and non-clinical factors. To develop the questionnaire and select the question items, we reviewed the previous literature and then consulted with 2 diabetes experts and 1 public health researcher, and initially selected 18 features. As the questionnaire contains clinical data (i.e., blood pressure, glucose level), informed consent had been confirmed by all the participants of this study that explained the study objective, prospective outcome, and benefits to society, while the participation was completely voluntary. After manual inspection, we divided and obtained two classes of data: healthy and diabetic people.

There were 256 participants (~34.6% of 738 participants) with diabetes and 482 non-diabetic participants. The average age of the participants was 44 years, with the age range of 19–78 years old. Before training the data, the dataset was divided by an 80:20 ratio where 80% data were used for the train set and the remaining 20% data were used for the test set. Then, the training data were fed to our model for conducting supervised machine learning techniques. The histogram of the age of participants who had diabetes (mean age = 50) and who did not have diabetes (mean age = 42) is shown in Figure 1. Details of the data properties are shown in Table 1.

3.2. Data Preprocessing

The data are helped to transform via the process of preprocessing, which enables a more accurate ML model to be constructed. Several tasks, including the rejection of outliers, the removal of records that have missing values, and the standardization of the data, are carried out during the preprocessing stage. Prior to carrying out the fitting procedure on the dataset, the MinMaxScaler was used to standardize the data values. Within the dataset, 256 samples have been identified as having diabetes, whereas 482 of the samples did not have diabetes. After that, the data has separated into a training set and a test set before being fed into the model. The given dataset is divided for training and testing in a ratio of 80:20, with the former functioning as the primary focus.

3.3. Feature Construction

As raw data can be often noisy, we assume that not all of the features are informative for classification and, hence, we followed the feature selection process to remove the less significant features and reduce input dimensionality. Constructing a good feature model through feature selection and extraction is often essential to ensure better and more effective prediction. The feature selection will not only remove the insignificant and redundant features that do not contribute to the accuracy of a predictive model but also reduces the computational complexity of analysis by feeding significant features to the classification model. This study used the PCA technique for feature extraction and the IG technique for feature selection.

3.4. Principal Component Analysis (PCA)

Principal component analysis is a technique for eliminating dimensions or features from a dataset [15]. The PCA shows that most information falls in the directions along which the variations are the greatest. After reducing the dimensions in PCA, a comparatively smaller number of dimensions represent the most information, which could also be reflected by the original dimensions.

In PCA, standardization is carried out to maximize the variance in the projected space. Then, the covariance matrix of the data is calculated for deducing the eigenvectors. When setting the eigenvalue as 1, our PCA analysis extracted five principal components that explain 58.5% of the total variance. However, after conducting a parallel analysis that calculates eigenvalues from randomly generated correlation matrices, two principal components were finally selected that explain 35.4% of the total variance. The KMO value and *p*-value in Bartlett's test were 0.73 and 0.000, respectively, indicating a good fit.

Table 2 presents the rotated component matrix in the PCA, where we can see that the major features that show covariance together in principal component 1 are area (−87.7%), eating meat (83.5%), an education level (82.3%), and eating fruits (80.9%). For principal component 2, the major features are the number of pregnancies (65.5%), gender (60.0%), and age (58.8%). As some features, such as eating fish, eating vegetables, and work stress, did not show any significant (less than 30%) covariance either in the principal component analysis 1 or 2, these 3 features were removed from the initially selected 18 features. Figure 2 shows the scatter plot of principal component 1 (PC1) and principal component 2 (PC2).

Table 2. Rotated component matrix in PCA.

No.	Features	PC 1	PC 2
1.	Area	−87.7%	
2.	Eating meat	83.5%	
3.	Education level	82.3%	
4.	Eating fruits	80.9%	
5.	Physical activity	−56.8%	
6.	Soft drinks	53.4%	
7.	Family history of diabetes	−35.9%	
8.	Eating fish		
9.	Eating vegetables		
10.	No. of pregnancies	−41.4%	65.6%
11.	Gender	−41.8%	60.0%
12.	Age		58.8%
13.	Systolic BP		58.3%
14.	Diastolic BP		53.7%
15.	Marital status		−52.3%
16.	BMI		31.4%
17.	Glucose		30.4%
18.	Work stress		

3.5. Information Gain Technique

Information gain [16] is a technique for feature selection to reduce high-dimensional data by evaluating the worth of an attribute by measuring the IG of a class. During the process of filter-based feature selection, the IG is used to choose important features according to the degree to which an attribute is relevant to a class [30]. Important attributes, meanwhile, are chosen according to rank. In other words, it lets us know how much information a feature provides regarding a class. This approach assigns rankings to each of the features according to a user-specified threshold value. IG is an entropy-based evaluation technique that helps us to identify the best features that give the most information. Using the combined indications of the objective factor, the IG approach calculates the anticipated value [31]. The IG entails applying various statistical formulas to evaluate every feature

(Table 1) by measuring their relevance to the target class (i.e., diabetes). In our test, we selected 10 features according to their ranking, as shown in Figure 3.

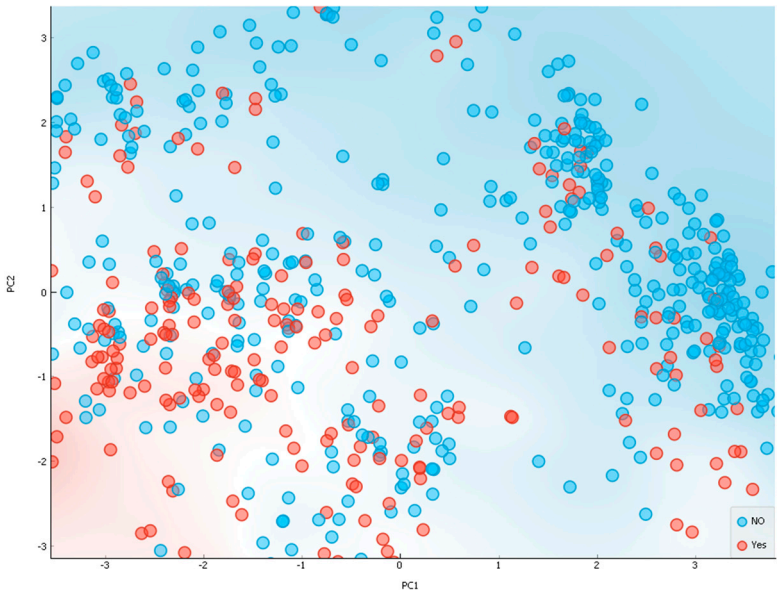

Figure 2. Scatter plot of PC 1 and PC 2 (red = diabetic and blue = non-diabetic patients).

Serial	Features	Value
1	Glucose	22.6%
2	Eating Meat	7.4%
3	Eating Fruits	6.6%
4	Area	6.6%
5	Age	5.9%
6	Education level	5.7%
7	Family History of Diabetes	4.1%
8	Physical Activity	2.5%
9	Eating Fish	1.7%
10	BMI	1.3%

Figure 3. The top 10 features found in information gain.

3.6. Predicting Flowchart

Figure 4 represents the workflow that was conducted in the following four steps: (1) data preprocessing, (2) dual-stage hybrid feature selection, (3) classification model training with hyperparameter optimization, and (4) performance evaluation after testing the ML model (Figure 4). The Python-based Orange toolkit (version 3.23.0) was used for both the feature construction step and the model training step. The feature construction step followed two methods, namely feature selection by the IG technique and feature extraction by the PCA technique.

Initially, the data were used which had 18 features selected based on expert opinions. These 18 features were classified as demographic features, life quality features, food habits, and clinical features as shown in Table 1. The model building started with the feed data with all 18 features. Then, the data were subjected to PCA and IG processes for feature selection and extraction to attain a higher accuracy rate in prediction. In the second step, we used multiple ML algorithms to predict diabetes, which was optimized by grid search. In this step, we created groups of features (Figure 5) based on the 18 initial features to compare the accuracy results following the clinical and non-clinical factors (i.e., demographics, life

quality, and dietary habits). Then, in the third step, the performance of the classification models was evaluated by applying the confusion matrix [32] and ROC curve (receiver operating characteristic curve) [33] to evaluate the model accuracy and performance score based on performance metrics.

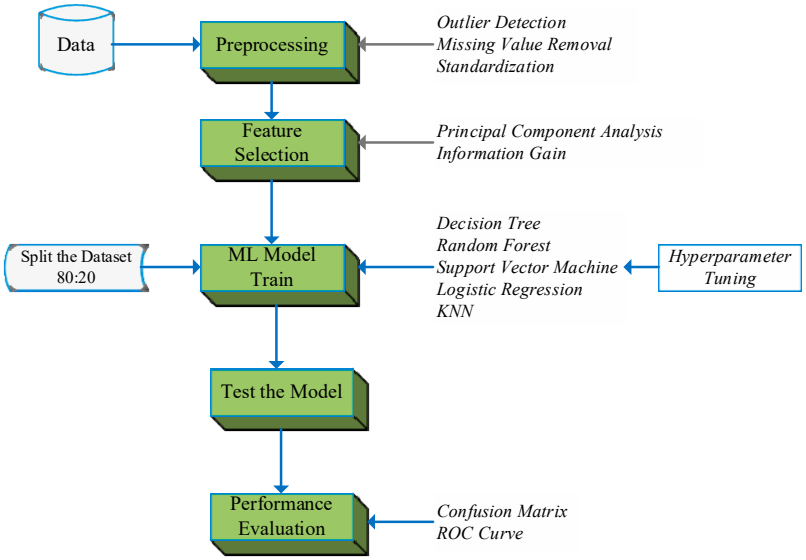

Figure 4. Predicting flowchart.

			Change from Group 1: All Features	
	AUC	CA	AUC	CA
Group 1:	87.14%	81.44%	0.00%	0.00%
Group 2:	87.19%	81.98%	0.06%	0.67%
Group 3:	87.07%	80.76%	-0.07%	-0.83%
Group 4:	87.16%	82.25%	0.03%	1.00%
Group 5:	86.59%	80.35%	-0.63%	-1.33%
Group 6:	87.55%	81.71%	0.47%	0.33%
Group 7:	72.17%	70.05%	-17.17%	-13.98%
Group 8:	73.70%	70.19%	-15.43%	-13.81%

Figure 5. The change in performance of all groups in comparison to Group 1 for logistic regression.

3.7. Model Validation

Model validation is a process of evaluating the capability of a model design. For model validation, two validation methods, namely the hold-out method and the k-fold cross-validation method, are used. In the hold-out method, we divide the dataset into two parts called training data and test data. The purpose of training data is to train the model, and the purpose of test data is to evaluate the model for its preference in diabetes prediction. On the other hand, in the k-fold (ideally the k = 5 to 10), cross-validation method, the complete dataset is used to both train and test the model. This study used a 10-fold cross-validation method. The benefits of using k-fold cross-validation are that it minimizes bias and variance by using all available data, as every observation in the original dataset has an equal chance to be both in the training and test dataset.

3.8. Classification Models

Many ML algorithms are used for achieving improved accuracy in predicting diabetes. For example, the decision tree is widely used ML algorithm for the problems of both regression and classification. On the other hand random forest used to combine many

decision trees and then merges them to generate a more accurate prediction that is also known as ensemble learning. The support vector machine algorithm generates a hyperplane that defines decision boundaries for classification. Moreover, the logistic regression algorithm uses a sigmoid function to predict the probability of a categorical variable, and the k-nearest neighbor algorithm is non-parametric supervised learning technique and does not need to depend on numbers, but rather on a ranking, and can be used for both classification and regression problem. Hence, for this study, we used decision tree, random forest, support vector machine, logistic regression, and k-nearest Neighbor algorithms as a classification model (Figure 4). The purpose of using multiple algorithms is to identify the comparative performance of ML algorithms for our dataset. The tuning parameters used in the classification models used in this study are shown in Table 3.

Table 3. Tuning parameters for the algorithms.

SL No.	Algorithms/Model Name	Tuning Parameters
1	Decision tree	Depth = 10
2	Random forest	Estimators = 100
3	Support vector machine	C = 1.0, kernel = rbf, degree = 3, γ = auto
4	Logistic regression	C = 10, random_state = 0
5	KNN	Neighbor = 5

4. Results

Performance Measurement

For performance measurement, we used the confusion matrix for measuring the performance of each ML algorithm or model used based on several parameters, known as accuracy (the ratio of correct predictions to total predictions made), precision (the ability of a model of correctly make positive predictions), recall (the ratio of the total number of correct positive predictions to the total number of positives in an actual class.), specificity, and F1 score (simply the value of the weighted average of precision and recall) [34]. The formulas of these parameters are given in the last row of Table 4. The confusion matrix helps us to count the true positive (TP), false negative (FN), false positive (FP), and true negative (TN) (Table 4) to ultimately know when a classification model is confused while making predictions. Once the confusion matrix is applied for performance evaluation, the receiver operating characteristic (ROC) curve [33] was plotted to present the performance of the binary classification models/algorithm by calculating the performance of all the classification models at various threshold levels (ranging 0 to 1), and then we calculated the AUC (area under the ROC curve).

Table 4. Confusion matrix.

		Predicted Class	
Actual Class		Positive	Negative
	Positive	True positives (TP)	False negatives (FN)
	Negative	False positives (FP)	True negative (TN)
	Formulas: Accuracy = (TP + TN)/(TN + TP + FP + FN) Precision = TP/(TP + FP) Recall = TP/(TP + FN) F1 Score = 2∗(Recall∗Precision)/(Recall + Precision) Specificity = TN/TN + FP		

The results section has discussed the findings of the parameters (i.e., accuracy, F1 score, precision, specificity, and recall) measured from the confusion matrix and the AUC from the ROC curve. As Table 5 shows, to compare the results, the results are presented in eight groups containing various numbers of features, which are as follows:

Group 1: all features;
Group 2: features after principal component analysis (PCA);
Group 3: features after information gain (top 10 features);
Group 4: features after information gain (top 5 features);
Group 5: features without demographic factors;
Group 6: features without life quality and food factors;
Group 7: features without clinical factors;
Group 8: features without glucose.

Table 5. Performance measurement.

Group 1: All Features							**Group 5: Features without Demographic Factors**						
Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**	**Model**	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**
LR	87.1%	81.4%	70.0%	79.2%	62.7%	90.5%	LR	86.6%	80.4%	68.3%	77.2%	61.2%	91.7%
RF	86.0%	81.0%	69.8%	77.5%	63.5%	93.2%	RF	82.5%	78.9%	66.5%	73.5%	60.8%	92.7%
DT	73.6%	78.0%	68.4%	68.1%	68.6%	89.2%	DT	72.6%	76.8%	65.0%	67.9%	62.4%	96.9%
KNN	74.8%	73.3%	54.9%	65.9%	47.1%	77.0%	KNN	72.9%	72.5%	52.2%	65.3%	43.5%	74.0%
SVM	78.2%	70.2%	62.2%	55.4%	71.0%	96.3%	SVM	75.9%	72.0%	58.2%	60.0%	56.5%	92.7%
Group 2: Features after PCA							**Group 6: Features without Food Factors**						
Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**	**Model**	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**
LR	87.2%	82.0%	70.9%	80.2%	63.5%	83.2%	LR	87.5%	81.7%	70.6%	79.4%	63.5%	90.5%
RF	84.3%	78.3%	65.8%	72.3%	60.4%	83.4%	RF	87.2%	80.6%	70.0%	75.2%	65.5%	89.9%
DT	73.3%	78.0%	67.7%	68.8%	66.7%	80.7%	DT	70.7%	76.2%	64.5%	66.4%	62.7%	90.3%
KNN	74.8%	73.2%	55.0%	65.4%	47.5%	83.9%	KNN	74.2%	72.9%	54.3%	65.0%	46.7%	76.2%
SVM	77.9%	71.0%	59.3%	57.6%	61.2%	82.8%	SVM	79.5%	70.9%	63.1%	56.1%	72.2%	97.9%
Group 3: Features after IG (Top 10 Features)							**Group 7: Features without Clinical Factors**						
Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**	**Model**	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**
LR	87.1%	80.8%	67.9%	80.2%	58.8%	92.1%	RF	70.8%	70.7%	52.2%	59.9%	46.3%	89.6%
RF	86.1%	79.9%	68.5%	74.9%	63.1%	89.0%	LR	72.2%	70.1%	49.9%	59.1%	43.1%	90.3%
DT	71.6%	78.0%	66.9%	69.8%	64.3%	90.9%	KNN	71.0%	69.2%	51.6%	56.5%	47.5%	77.8%
KNN	82.0%	77.8%	63.9%	72.9%	56.9%	79.9%	DT	63.0%	66.3%	45.5%	51.5%	40.8%	90.7%
SVM	75.2%	69.6%	58.5%	55.4%	62.0%	96.5%	SVM	61.3%	59.2%	40.2%	40.7%	39.6%	95.7%
Group 4: Features after IG (Top 5 Features)							**Group 8: Features without Glucose**						
Model	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**	**Model**	**AUC**	**CA**	**F1**	**Precision**	**Recall**	**Speci**
LR	88.5%	82.2%	70.6%	82.6%	61.6%	92.5%	LR	73.7%	70.2%	49.5%	59.7%	42.4%	85.1%
RF	79.4%	80.2%	68.7%	75.8%	62.7%	85.5%	RF	72.4%	68.8%	48.7%	56.5%	42.7%	86.3%
DT	72.9%	77.5%	65.8%	69.3%	62.7%	91.1%	KNN	68.1%	67.8%	48.9%	54.0%	44.7%	72.3%
KNN	75.4%	77.4%	63.3%	72.0%	56.5%	82.2%	DT	62.6%	66.3%	49.7%	51.3%	48.2%	86.3%
SVM	80.0%	71.7%	63.4%	57.3%	71.0%	98.6%	SVM	62.7%	60.6%	47.2%	43.9%	51.0%	84.1%

Note: AUC, area under curve; CA, accuracy; Speci, specificity.

The accuracy of correct prediction is presented as ACC in Table 5. The result for Group 1 contains all 18 features which were initially selected. The comparison of the result of Group 1 (all features before feature selection) with the other groups in Table 5 has given us a better understating of the performance of the model before and after the feather selection step. The highest accuracy rates for Group 1 to Group 8 were 81.4%, 82.0%, 80.0%, 82.2%, 80.4%, 81.7%, 70.7%, and 70.2% respectively.

When comparing the performance of the ML algorithms used, logistic regression yielded the best result (accuracy of 81.4%, 82.0%, 80.8%, 82.2%, 80.4%, 81.7%, and 70.2%, respectively, for Group 1 to Group 6 and Group 8, except Group 7) in comparison to the other algorithms. The second-best result for all the groups was yielded by the random forest (accuracy of 81.0%, 78.3%, 79.9%, 80.2%, 78.9%, 80.6%, and 68.8%, respectively, for Group 1 to Group 6 and Group 8).

After feature extraction using PCA, following the total variance for the top two principal components and rotated component matrix (Table 2) we removed three features (i.e., eating fish, eating vegetables, and work stress) from the initially selected 18 features. As a result of conducting PCA, the results for Group 2 (i.e., features after principal component

analysis) slightly improved compared to Group 1 (i.e., all features), as the accuracy and AUC for logistic regression in Group 2 were 82.0% and 87.2%, respectively (Table 5). Furthermore, after conducting the information gain technique for feature selection, we found the top 10 features, as shown in Table 5.

When using the top 10 features in Group 3, the accuracy and AUC for logistic regression in Group 3 were 80.8% and 87.1%, respectively, which were not better than the respective parameters' values in Group 1 and Group 2. However, when using the top five features (i.e., glucose, eating meat, eating fruits, area, and age) in Group 4, we obtained the best performance among all eight groups. The accuracy and AUC for logistic regression in Group 4 were 82.2% and 87.2%, respectively. Again, when comparing the changes in results of CA between Group 2 (after selecting the features after PCA) and Group 1 (all features), the tiny increase in performance (from 81.4% to 82.0%) indicates that PCA does not increase the accuracy for this dataset significantly and, hence, is not very useful for this particular dataset.

Figure 5 shows the group-wise AUC and accuracy (CA) for logistic regression and the respective changes in comparison to Group 1. For logistic regression, there were 0.06% and 0.67% increases in AUC and CA, respectively, in Group 2 in comparison to Group 1 after PCA, and 0.03% and 1.00% increases in AUC and CA, respectively, in Group 4 in comparison to Group 1 after selecting the top five features (Table 5). Furthermore, to test the effect of blood glucose level, which is a significant clinical feature, on the machine learning model performance, we excluded this clinical feature from the dataset with the initial 18 features. Then, it was found that there were 15.43% and 13.81% decreases for AUC and CA, respectively, in Group 8 when we removed glucose. Furthermore, there were 17.17% and 13.98% decreases, respectively, in AUC and CA in Group 7 when we removed all the clinical features from Group 7. Overall, the performance of the logistic regression for the features in Group 4 was the best after selecting the top four features, and the features in Group 8 were the worst after removing the clinical factors. Moreover, when removing the clinical factors in Group 7, it was found that random forest performed the best (AUC = 70.8% and CA = 70.7%) instead of logistic regression (AUC = 72.2% and CA = 70.1%) for Group 7 only.

5. Discussion

The major objectives of this study were to find how much differently the clinical factors and diverse non-clinical factors are related in predicting diabetes in Bangladesh, where fast-changing demographic, dietary habits, and life quality are prevalent. To the best of our knowledge, no previous studies, if any, have investigated or incorporated similar factors or features, more specifically in a least-developed country, such as Bangladesh. Hence, this is a significant study from the perspective of Bangladesh, where the prevalence of diabetes is progressively greater than before. Before conducting this experimental study, we collected clinical and non-clinical data from the urban and rural populace in Bangladesh with informed consent for applying the ML model which consisted of three major steps, namely 'feature selection' for selecting the most significant features, 'classification model building' of diabetic cases, and the performance evaluation using the confusion matrix and area under the ROC curve. Figure 6 shows the ROC curve of all classifiers for Group 4, which has an AUC of 82.2%, which is a good result as per the general reference.

From Table 5 and Figure 5, we can find that both PCA and IG (in the case of the top five features) increase the model performance to a very small extent. Although this study has shown a good level of accuracy in classifying diabetes compared to [21–23], the accuracy was not satisfactory for all groups. The large decrease in the performance of Group 7 and Group 8 implies that the inclusion of clinical features/factors is significant, and that glucose level is especially significant as a clinical factor for predicting diabetes. Although logistic regression performed the best in our study, its performance did not differ greatly from the random forest. The classification model KNN showed a change in performance consistency

for Group 7 and Group 8, which could be because of the reduction in noise and data dimensionality.

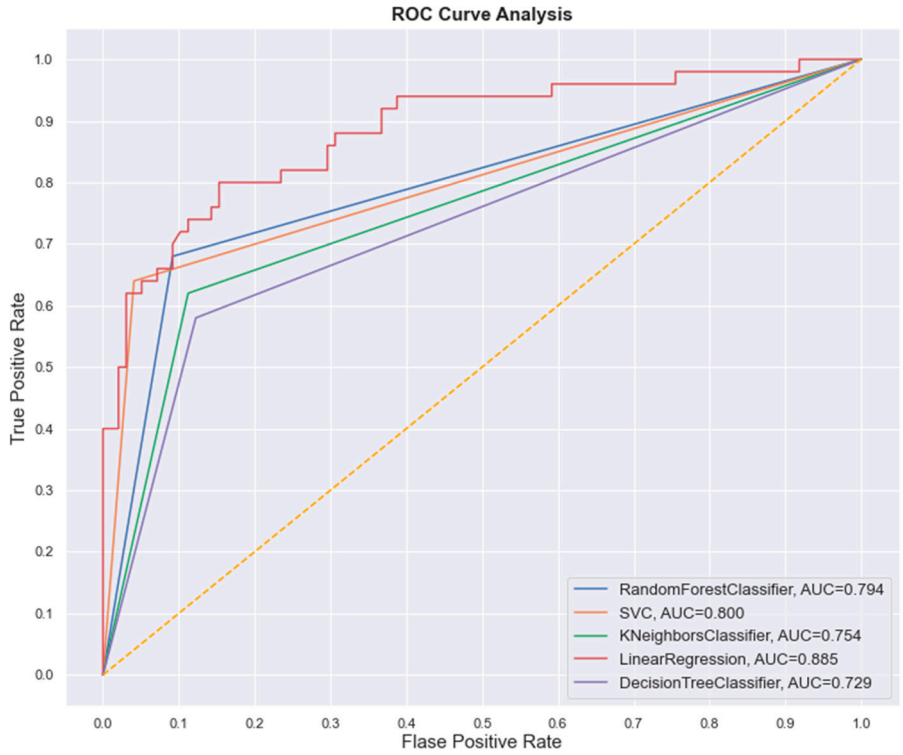

Figure 6. AUC of the ROC curve of logistic regression for Group 4 with the top five features.

The best performance by the features in Group 4 indicates that glucose and age as clinical factors, eating meat and eating fruits (food habit), and area (demographic) as non-clinical factors are the most significant in predicting diabetes. In a previous study, it was also found that meat consumption is linked to the onset of diabetes, which is in line with the finding of Feskens et al. [35]. On the other hand, following the top five features in Group 4, we can say that living areas, such as an urban or rural residence (the demographic factor which might be related correlated to life quality), can have a big influence on the prevalence of diabetes. In our dataset, it was found that 46.5% of the participants from the urban areas (425 urban participants) had diabetes in comparison to 18.2% of the rural participants (313 rural participants). In general, urban areas have a higher prevalence of diabetes (close to double) than rural areas [36] because of lifestyle differences. Hence, the information gain techniques have listed or ranked the area as one of the top five features in predicting diabetes. Figure 7 graphically shows the diabetic and non-diabetic patients against the top five features.

In comparison to the studies by others that show accuracies exceeding 95%, for example, the studies with models by [29,37], our model showed a maximum level of accuracy (for Group 4) of slightly more than 82%. A major reason for it could be the inclusion of fewer clinical factors. For instance, the study by Sneha and Gangil [37] used other clinical factors, such as serum sodium and serum potassium levels, to predict diabetes. Therefore, from the comparison of the results of this study, the authors stress that instead of the inclusion of more non-clinical factors, the clinical factors can significantly improve the prediction accuracy. Furthermore, it indicates that there is a necessity for more optimization by using various advanced machine learning techniques, such as ensemble methods that used a combination of multiple classification models, so that the local doctors and other stakeholders can consider using the result of our model more seriously.

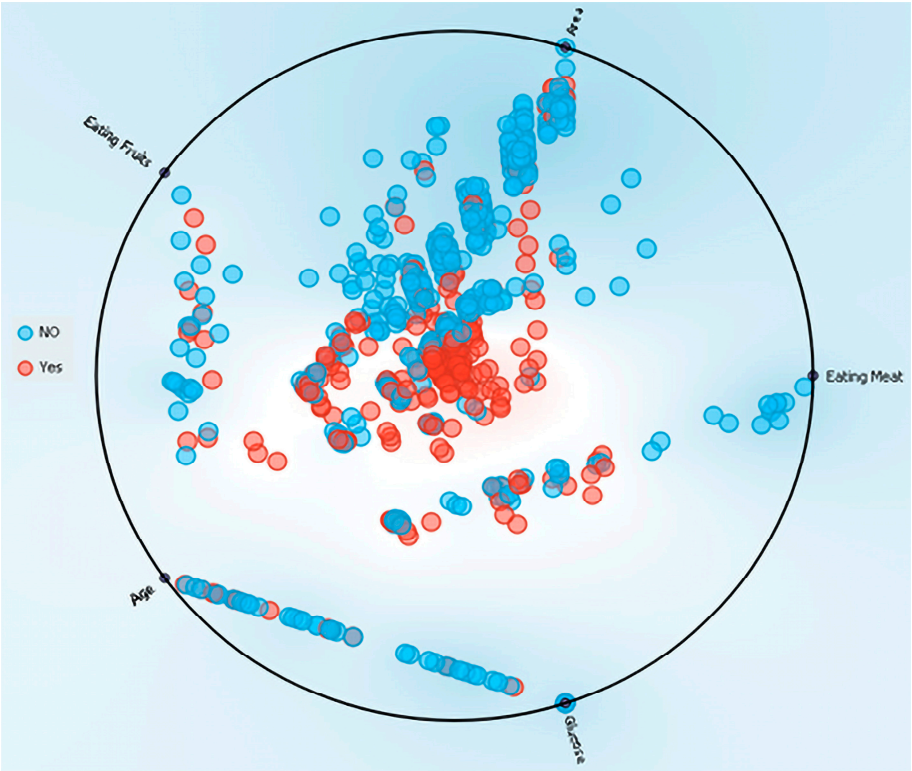

Figure 7. The graphical representation of diabetic and non-diabetic patients against the top five features.

6. Conclusions

Our study has yielded an accuracy level of over 82.2% (for the top five features selected) with an AUC (area under the ROC curve) value of 87.2%, which is a good result. To yield good prediction results, we used the feature selection process carefully by constructing representative features for diabetes classification or prediction. Subsequently, we trained our model based on constructed features to predict diabetes. In addition to the identification of the significant clinical and non-clinical factors in predicting diabetes, from the comparison of the results of this study, the authors conclude that instead of the inclusion of more non-clinical factors, the clinical factors can significantly improve the prediction accuracy.

This study is significant in the context of Bangladesh, as it is a fast-growing economic country where the fast change in lifestyle as a result of an increase in per capita income has led to a big change in the prevalence of diabetes and the consequent higher cost of regular healthcare [38]. Furthermore, as per the best knowledge of the authors, few previous studies, if any, have used the above-mentioned dietary habit features to compare their influence on the prediction of diabetes in Bangladesh that can help to make a preliminary judgment about the onset of diabetes based on lifestyle, food habits, and the other factors mentioned above.

Finally, we can conclude that the result of this study is good but not yet satisfactory enough to assist doctors to make any final decision without reluctance. However, the authors assume that more advanced feature selection techniques (i.e., ensemble techniques), and advanced optimization techniques (i.e., particle swarm optimization, orthogonal array tuning methods) might provide satisfactory performance with the proposed model. Therefore, the authors will conduct future studies considering more advanced feature selection and optimization techniques using the same dataset. Moreover, in future research, the authors will also consider the computational complexity calculation due to the application of advanced optimization techniques.

Author Contributions: Conceptualization, I.J.K.; methodology, I.J.K. and M.R.H.; formal analysis, I.J.K.; investigation, M.R.H.; data curation, I.J.K.; writing—original draft, I.J.K.; writing—review and editing, N.H. (initial, review one and review two); supervision, M.R.H. and N.H.; funding acquisition, I.J.K. All authors have read and agreed to the published version of the manuscript.

Funding: The authors gratefully acknowledge the technical and financial support by University of Dhaka, Bangladesh.

Institutional Review Board Statement: All participants involved were treated based on the safety procedure described in the Helsinki Declaration, 2013. The authors obtained ethical approval from the National Research Ethics Committee (NREC) of the Bangladesh Medical Research Council (BMRC) with approval No. 18325022019.

Informed Consent Statement: Informed consent had been confirmed by all the participants of this study that explained the study objective, prospective outcome, and benefits to society, while the participation was completely voluntary.

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: The authors declare no potential conflict of interest with respect to the research, authorship, and/or publication of this article.

References

1. Maniruzzaman, M.; Islam, M.M.; Rahman, M.J.; Hasan, M.A.M.; Shin, J. Risk prediction of diabetic nephropathy using machine learning techniques: A pilot study with secondary data. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2021**, *15*, 102263. [CrossRef]
2. Febrian, M.E.; Ferdinan, F.X.; Sendani, G.P.; Suryanigrum, K.M.; Yunanda, R. Diabetes prediction using supervised machine learning. *Procedia Comput. Sci.* **2023**, *216*, 21–30. [CrossRef]
3. IDF Diabetes around the World in 2021. Available online: https://diabetesatlas.org/ (accessed on 12 September 2021).
4. Pradeepa, R.; Mohan, V. Epidemiology of type 2 diabetes in India. *Indian J. Ophthalmol.* **2021**, *69*, 2932–2938. [CrossRef]
5. Chen, L.; Magliano, D.J.; Balkau, B.; Colagiuri, S.; Zimmet, P.Z.; Tonkin, A.M.; Mitchell, P.; Phillips, P.J.; Shaw, J.E. AUSDRISK: An Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. *Med. J. Aust.* **2010**, *192*, 197–202. [CrossRef] [PubMed]
6. The InterAct, C. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: The EPIC-InterAct study. *Diabetologia* **2013**, *56*, 60–69. [CrossRef] [PubMed]
7. Lee, D.H.; Keum, N.; Hu, F.B.; Orav, E.J.; Rimm, E.B.; Willett, W.C.; Giovannucci, E.L. Comparison of the association of predicted fat mass, body mass index, and other obesity indicators with type 2 diabetes risk: Two large prospective studies in US men and women. *Eur. J. Epidemiol.* **2018**, *33*, 1113–1123. [CrossRef]
8. Sulaiman, N.; Mahmoud, I.; Hussein, A.; Elbadawi, S.; Abusnana, S.; Zimmet, P.; Shaw, J. Care, Diabetes risk score in the United Arab Emirates: A screening tool for the early detection of type 2 diabetes mellitus. *BMJ Open Diabetes Res.* **2018**, *6*, e000489. [CrossRef]
9. Wainberg, M.; Mahajan, A.; Kundaje, A.; McCarthy, M.I.; Ingelsson, E.; Sinnott-Armstrong, N.; Rivas, M.A. Homogeneity in the association of body mass index with type 2 diabetes across the UK Biobank: A Mendelian randomization study. *PLoS Med.* **2019**, *16*, e1002982. [CrossRef] [PubMed]
10. 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]
11. Perveen, S.; Shahbaz, M.; Keshavjee, K.; Guergachi, A. Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques. *IEEE Access* **2018**, *7*, 1365–1375. [CrossRef]
12. Narwane, S.V.; Sawarkar, S.D. Is handling unbalanced datasets for machine learning uplifts system performance?: A case of diabetic prediction. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2022**, *16*, 102609. [CrossRef] [PubMed]
13. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine Learning and Data Mining Methods in Diabetes Research. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 104–116. [CrossRef]
14. Bekele, B.B.; Manzar, M.D.; Alqahtani, M.; Pandi-Perumal, S.R. Diabetes mellitus, metabolic syndrome, and physical activity among Ethiopians: A systematic review. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2021**, *15*, 257–265. [CrossRef] [PubMed]
15. Kamadi, V.V.; Allam, A.R.; Thummala, S.M. A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. *Appl. Soft Comput.* **2016**, *49*, 137–145. [CrossRef]
16. Win, T.Z.; Kham, N.S.M. Information Gain Measured Feature Selection to Reduce High Dimensional Data. Ph.D. Thesis, University of Computer Studies, Yangon, Myanmar, 2019.
17. UNCTAD UN List of Least Developed Countries. Available online: https://unctad.org/topic/least-developed-countries/list (accessed on 21 October 2022).

18. Dagliati, A.; Marini, S.; Sacchi, L.; Cogni, G.; Teliti, M.; Tibollo, V.; De Cata, P.; Chiovato, L.; Bellazzi, R. Machine Learning Methods to Predict Diabetes Complications. *J. Diabetes Sci. Technol.* **2018**, *12*, 295–302. [CrossRef]
19. Khalil, R.M.; Al-Jumaily, A. Machine learning based prediction of depression among type 2 diabetic patients. In Proceedings of the 12th International Conference on Intelligent Systems Knowledge Engineering, Nanjing, China, 24–26 November 2017; pp. 1–5.
20. Lee, B.J.; Kim, J.Y. Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides based on Machine Learning. *IEEE J. Biomed. Health Inform.* **2015**, *20*, 39–46. [CrossRef] [PubMed]
21. 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]
22. Reddy, D.J.; Mounika, B.; Sindhu, S.; Reddy, T.P.; Reddy, N.S.; Sri, G.J.; Swaraja, K.; Meenakshi, K.; Kora, P. WITHDRAWN: Predictive machine learning model for early detection and analysis of diabetes. *Mater. Today Proc.* **2020**. [CrossRef]
23. Goyal, P.; Jain, S. Prediction of Type-2 Diabetes using Classification and Ensemble Method Approach. In Proceedings of the 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 10–11 March 2022; pp. 658–665.
24. 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]
25. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Front. Genet.* **2018**, *9*, 515. [CrossRef]
26. 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]
27. Pedersen, H.K.; Gudmundsdottir, V.; Pedersen, M.K.; Brorsson, C.; Brunak, S.; Gupta, R. Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. *NPJ Genom. Med.* **2016**, *1*, 16035. [CrossRef] [PubMed]
28. Tsao, H.-Y.; Chan, P.-Y.; Su, E.C.-Y. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinform.* **2018**, *19*, 111–121. [CrossRef] [PubMed]
29. Wu, H.; Yang, S.; Huang, Z.; He, J.; Wang, X. Type 2 diabetes mellitus prediction model based on data mining. *Inform. Med. Unlocked* **2018**, *10*, 100–107. [CrossRef]
30. Selvakumar, B.; Muneeswaran, K. Firefly algorithm based feature selection for network intrusion detection. *Comput. Secur.* **2019**, *81*, 148–155. [CrossRef]
31. Gokulnath, C.B.; Shantharajah, S.P. An optimized feature selection based on genetic approach and support vector machine for heart disease. *Clust. Comput.* **2018**, *22*, 14777–14787. [CrossRef]
32. Caelen, O. A Bayesian interpretation of the confusion matrix. *Ann. Math. Artif. Intell.* **2017**, *81*, 429–450. [CrossRef]
33. Narkhede, S. Understanding auc-roc curve. *Towards Data Sci.* **2018**, *26*, 220–227.
34. Sisodia, D.; Sisodia, D.S. Prediction of diabetes using classification algorithms. *Procedia Comput. Sci.* **2018**, *132*, 1578–1585. [CrossRef]
35. Feskens, E.J.; Sluik, D.; van Woudenbergh, G.J. Meat consumption, diabetes, and its complications. *Curr. Diabetes Rep.* **2013**, *13*, 298–306. [CrossRef] [PubMed]
36. Seck, S.M.; Dia, D.G.; Doupa, D.; Diop-Dia, A.; Thiam, I.; Ndong, M.; Gueye, L. Diabetes Burden in Urban and Rural Senegalese Populations: A Cross-Sectional Study in 2012. *Int. J. Endocrinol.* **2015**, *2015*, 163641. [CrossRef] [PubMed]
37. Sneha, N.; Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J. Big Data* **2019**, *6*, 13. [CrossRef]
38. Mohiuddin, A.K. Diabetes fact: Bangladesh perspective. *Int. J. Diabetes Res.* **2019**, *2*, 14–20.