2020 23rd International Conference on Computer and Information Technology (ICCIT), 19-21 December, 2020.

Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques Resolving Class Imbalance

Kazi Amit Hasan
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: kaziamithasan89@gmail.com

Md. Al Mehedi Hasan
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
Email: mehedi_ru@yahoo.com

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392694

***Abstract*—Being the most common and rapidly growing disease, Diabetes affecting a huge number of people from all span of ages each year that reduces the lifespan. Having a high affecting rate, it increases the significance of initial diagnosis. Diabetes brings other complicated complications like cardiovascular disease, kidney failure, stroke, damaging the vital organs etc. Early diagnosis of diabetes reduces the likelihood of transiting it into a chronic and severe state. The identification and analysis of risk factors of different spinal attributes help to identify the prevalence of diabetes in medical diagnosis. The prevalence measure and identification of diabetes in the early stages reduce the chances of future complications. In this research, the collective NHANES dataset of 1999-2000 to 2015-2016 was used and the purposes of this research were to analyze and ascertain the potential risk factors correlated with diabetes by using Logistic Regression, ANOVA and also to identify the abnormalities by using multiple supervised machine learning algorithms. Class imbalance, outlier problems were handled and experimental results show that age, blood-related diabetes, cholesterol and BMI are the most significant risk factors that associated with diabetes. Along with this, the highest accuracy score .90 was achieved with the random forest classification method.**

***Index Terms*—Risk Factors Analysis, Classification, Outliers, Class Imbalance, Logistic Regression, Machine Learning**

I. Introduction

Diabetes is recognized as one of the most frequent and rapidly growing diseases worldwide. Diabetes is a metabolic complication that is also referred as diabetes mellitus. It increases the blood sugar level significantly [1]. Diabetes enhances the chances of other long term complications like heart attack, kidney failure, cardiovascular disease etc. In late ages, diabetes affected patients also suffer from different severe damage to nerves, vital organs and blood vessels [2]. People with ages in the range from 20 to 80 years are at high risk of being impacted by diabetes. About 465 million grown-ups are affected by this disease which increases to 700 million by 2045 [3]. The mortality rates of diabetes are also extremely high due to its association with other complicated disorders. About 4.2 million deaths were happened because of diabetes [1]. The volume of affecting people with diabetes grew doubled in the last 20 years globally [3]. The mortality rate of diabetes cases is rapidly growing day by day as well as the affecting rates.

The consequences of diabetes become more frightening because of delayed diagnosis. Initial disclosure of diabetes can decrease the severe future impacts caused by this disease. The correlation with other chronic, long-standing diseases made diabetes one of the most investigated topics among the researchers. Nonlinear, complex, outliers etc medical data types made the investigation of diabetes more challenging and complicated [1]. During recent years, health informatics systems perform an essential part in recognizing and monitoring various diseases and disorders. Different machine learning-based systems can enhance the diagnosis of affected cases significantly and analyze the risk factors of the patients in the initial stages to estimate the severity of the disease. In recent studies, researchers proposed several machine learning strategies to classify diabetes patients. Several statistical analysis were also performed to determine the potential risk factors of diabetes [1]. In those investigations, various machine learning strategies were implemented. But some researches have lacking in data preprocessing [1]. Class imbalance, outliers etc were not handled optimally which can affect the outcome significantly.

In this research, the machine learning approaches and statistical methods were combinedly used to identify the potential clinical risk factors with their effects on diabetes and to propose a system that can identify the diabetes patients effectively. The analysis of risk factors of different independent variables were performed to identify the most impactful risk factors correlated with diabetes. Along with risk analysis, different machine learning strategies are also used to classify diabetes patients. The overall workflow of our proposed system is shown in Fig. 2 in Methodology section. The proposed ANN architecture is shown in Fig. 4. The outliers were handled by replacing the extreme values with median values. The class imbalance problem was handled by assigning class weights. The risk factors analysis were conducted by using statistical logistic regression and analysis of variance (ANOVA). After handling the outliers and class imbalance issue, the features

978-1-6654-2244-4/20/$31.00 ©2020 IEEE

were fed to different classification algorithms and the classification performances were measured and compared with other state of art methods are shown in Table I and Table II.

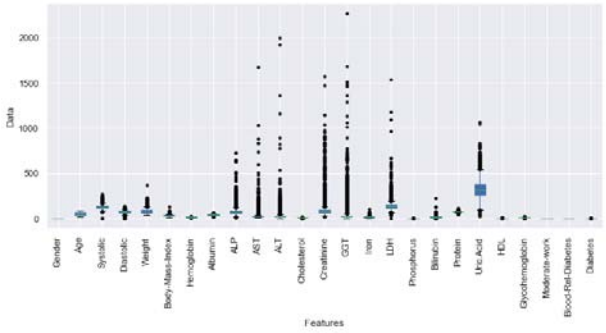
II. Dataset

A. Dataset Description

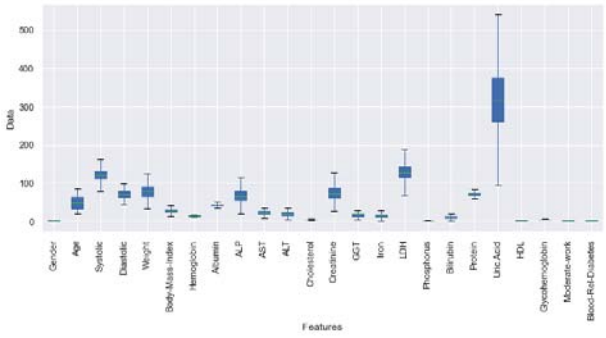
The dataset used in this research was obtained from the National Health and Nutrition Examination Survey (NHANES). The assembled version NHANES data from 1999-2000 to 2015-2016 was utilized for this study. NHANES is a health and nutrition assessment curriculum based on the United States population which manages diverse types of data. The dataset used in this research is a collective version of demographic, questionnaire, examination and laboratory data from 1999-2000 to 2015-2016. The dataset is also used in other studies [4]. The dataset has responses of 37079 people with 51 types of different attributes. A total of 25 independent variables were selected for the identification of diabetes diseases based on the effects. For this research, diabetes is considered as a dependent variable. There is a total of 37079 respondents in this dataset containing 4852 diabetes patients and 32227 normal patients. According to the class distribution of diabetes patients, the dataset has a class imbalance and this problem is handled by assigning class weights while training with different machine learning techniques.

B. Exploratory Data Analysis (EDA)

The dataset contains information of 37079 respondents with 51 different variables. The variables are *gender, ratio family income poverty, pulse rate (60s), weight, white blood cells, systolic, diastolic, lymphocyte, monocyte, eosinophils, basophils, red blood cells, hemoglobin, mean cell volume, mean cell hemoglobin concentrated, mean cell hemoglobin, platelet count, the mean volume of platelet, height, body mass index, segmented neutrophils, age, annual family income, hematocrit, red cell distribution width, albumin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), cholesterol, creatinine, glucose, gamma-glutamyl transferase (GGT), iron, lactate dehydrogenase (LDH), phosphorus, bilirubin, total cholesterol, high-density lipoprotein (HDL), glycohemoglobin, vigorous work, moderate work, health insurance, coronary heart disease, protein, uric acid, triglycerides, diabetes, blood-related diabetes, blood-related stroke, diabetics* [4].
Diabetes is considered as a dependent variable for this research. 25 independent variables were selected for this research based on their feature importance and effects on diabetes patients [1]. Some variables that are not correlated with the outcome variable (*ratio family income poverty, pulse rate (60s), annual family income, lymphocyte, monocyte, eosinophils, basophils, mean cell volume etc.*) are not considered for further analysis. Python, the programming language was used for all kinds of analysis and experiments done in this research [5].



(a) Boxplot visualization before handling outliers



(b) Boxplot visualization after handling outliers

Fig. 1: Identifying and handling the outliers

III. Data Preprocessing

A. Outlier Handling

Outliers are referred as the data points which are different from other data points in dataset distribution [6]. Outliers represent the abnormality, unusual values of particular data distribution. Outliers can influence the statistical decision making and machine learning algorithms' performance significantly. The boxplot visualizations of the selected independent variables of the dataset are generated for outlier analysis and shown in Fig. 1. According to Fig. 1, it is clear that the dataset is affected by outliers. Without handling them, it may impact the performance of algorithms and the analysis of risk factors. In this research, Interquartile Range (IQR) method was applied for outliers identification purposes [7]. The Interquartile values lie between the third quartile and the first quartile. The mathematical representation of IQR is :

$$IQR = Q3 - Q1 \tag{1}$$

where Q3 and Q1 represent the third and first quartile respectively.

The lower and upper fence were calculated to denote the maximum and minimum value ranges and to determine the outliers. Values fall outside of the minimum and maximum acceptable range considered as outliers shown in Fig. 1. The mathematical representation of the lower and upper fence is:

$$Upper fence = Q3 + (1.5 * IQR) \tag{2}$$

$$Lower fence = Q1–(1.5 * IQR) \tag{3}$$

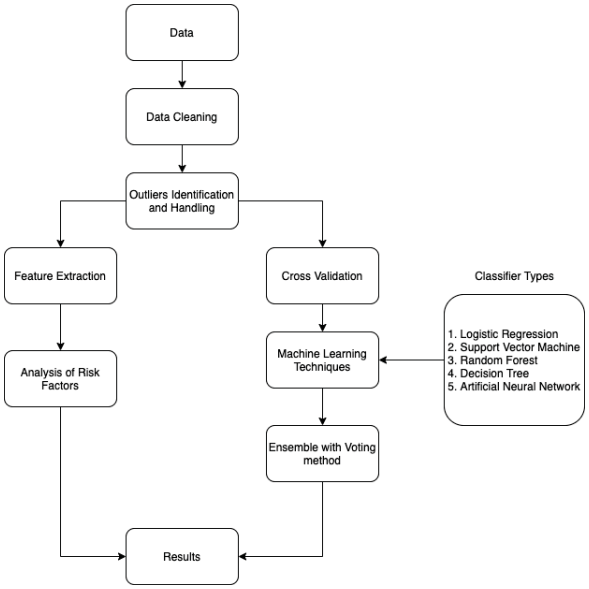


Fig. 2: Graphical representation of proposed system

A total number of 15 variables are affected with outliers: *systolic, diastolic, weight, body mass index, hemoglobin, albumin, ALP, AST, ALT, cholesterol, GGT, iron, LDH, HDL, glycohemoglobin.*

The outliers were handled by replacing the extreme values with their median values than removing them completely. So, the outliers are replaced with the median values of their respective variables shown in Fig.1.

B. Assigning Class Weights

As there is a total of 37079 respondents in the dataset containing 4852 diabetes and 32227 normal patients which indicate the class imbalance problem. The class imbalance problem is handled by assigning class weights *(Class 0: 0.52, Class 1: 0.47)* to the classes using scikit-learn tool [8].

C. Data Standardization

Standardization rescales the distribution of values in a way that it holds the mean 0 and standard deviation (std) 1. It is also less affected by outliers [9]. In this research, after handling the outliers of the dataset, data standardization was performed by using the scikit-learn tool [8].

IV. Methodology

A. Feature Importance

The dataset used in this research has 51 independent variables. But for this research, 25 independent variables were taken into consideration by using a univariate feature selection method and statistical logistic regression method which represents their effects on diabetes patients [1]. The ANOVA F-scores of individual variables validated the outcomes of risk factors identified by using statistical logistic regression. The resultant risk factors are also compared and validated with other research findings [1].

	Features	F_Scores
1	Age	2686.409705
23	Blood-Rel-Diabetes	1536.785191
5	Body-Mass-Index	1000.670115
2	Systolic	766.368625
4	Weight	526.900940
7	Albumin	525.811066
21	Glycohemoglobin	477.417483
20	HDL	378.750379
11	Cholesterol	334.923024
13	GGT	305.472574
6	Hemoglobin	270.428079
14	Iron	211.216877
22	Moderate-work	188.266414
8	ALP	156.172300
19	Uric.Acid	144.722845
12	Creatinine	102.880413
17	Bilirubin	80.463331
3	Diastolic	63.316044
15	LDH	49.826302
10	ALT	27.645892
9	AST	22.682305
0	Gender	12.219556
18	Protein	8.581358
16	Phosphorus	0.000433

Fig. 3: Feature importance based on their ANOVA F-score

*1) **Analysis of Variance (ANOVA):*** The dataset used in this research has 25 independent variables. To find the optimal features a feature selection algorithm was implemented. Feature selection methods also enhance prediction quality. In this study, as a feature selection technique, the analysis of variance (ANOVA) was used and it is also used in different studies [10]. ANOVA is a simple and strong statistical method that examines the means of a couple or more groups. It also determines how much the groups are significantly different from each other. In this research, the topmost features were chosen and sorted according to their ANOVA F-score shown in Fig. 3. The scikit-learn [8] tool was used to measure the F-score which represents the individual feature importance of the independent variables.

*2) **Risk Factors Analysis using Logistic Regression:*** In this research, the logistic regression model is used for statistical data exploration. The statsmodels tool was used to measure and identify the relative risk factors in our dataset [11]. To identify the risk and effects of individual attributes on the dependent variable, p-value, confidence interval (C.I.), odd ratio (O.R.) were generated of each variable to investigate the relative risks of independent variables towards the outcome variable and shown in Table I. The statistical significance can be represented by p-value. The range of p-value lies in the range of 0 and 1 [12]. In this research, the p-value is considered less than 0.05 for statistical significance. According to Table I, *age, diastolic, cholesterol, blood-related diabetics* are the most significant risk factors along with the risk factors associated with diabetics disease. Odd ratio (OR) is also used to investigate the relative risk associated with the outcome variable.

B. Cross Validation

Cross-validation (CV) is a data partitioning method that divides the dataset into two groups as train data and test/validation data. In this research, the *ShuffleSplit* function was selected as a cross-validation method that generates a user-defined number of independent train and validation splits. In

this method, the samples are shuffled and split into a set of the train and test/validation sets. The aim of selecting this method is to have control over the number of iterations and proportion data samples on both sides of the train/validation set. The dataset was divided into 80% training set and 20% set/validation set in the cross-validation procedure. With defined train and validation splits, the data was tested with cv/split values 3, 5, 10 respectively by using scikit-learn tool [8]. The results are shown in Table II.

C. Prediction Models

In this research, several machine learning predictive models were used for evaluating the performance of the proposed method.

1) ***Logistic Regression (LR):*** Logistic Regression is used in previous risk factors identification studies of diabetes patients [1]. LR is one of the various common supervised machine learning algorithms because of it's binary in nature. In this research, statistical LR is also used to investigate the potential risk factors associated with diabetes patients. In LR, the dependent variable is a logit and the equation is:

$$logit(x) = log(\frac{x}{1-x}) = \beta_0 + \beta_1 y_1 + ... + \beta_y y_y \quad (4)$$

where, *x* represents the probability of diabetes patients when *y = 1* and *1-x* represents the probability of non-diabetes patients when *y = 0.* The odd ratio, confidence intervals were also identified by using the statistical logistic regression method which identifies the possible risk factors associated with diabetes. The features were selected and sorted according to their respective p-values. P-value less 0.05 was considered for this research to identify the individual risk factors. The risk factors with their p-value, odd ratio and confidence interval are shown in Table. I.

2) ***Support Vector Machine (SVM):*** Support vector machine (SVM) is also used in previous risk factor identification studies of diabetes [13]. SVM is a popular supervised learning algorithm used for investigating the hidden patterns as well as solving classification problems. SVM is also known as a decision boundary function by finding a hyperplane. In this method, with the guidance of nonlinear mapping, the original train data is transformed into higher dimensional space. Later, SVM differentiates the patterns of different classes [14]. The mathematical equation of hyperplane is:

$$W.Y + p = 0 \quad (5)$$

where, *W* represents the weighted vector and *b* represents the scalar data.

In this research, Radial Basis Function (RBF) kernel of SVM are used as one of the classification methods. The mathematical representation of RBF kernel is:

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2) \quad (6)$$

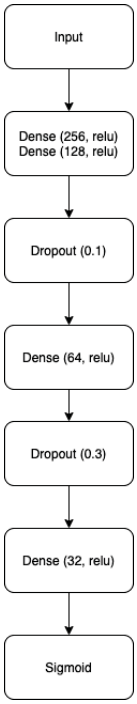


Fig. 4: Proposed architecture of Artificial Neural Network (ANN)

3) ***Random Forest (RF):*** Random Forest classification is also used in previous diabetes diagnosis studies [1]. The random forest can be implemented in both classification and regression problems. It is an extended version of the bagging algorithm, fast and robust to over-fitting that takes its root from the decision tree's characteristics [15].

4) ***Decision Tree (DT):*** Decision tree-based classification is also used in the studies [1]. DT is a popular supervised learning method used as a classification and regression model both. It builds classification or regression models in the type of tree structure. In this method, a tree is constructed containing the input features as nodes. It selects the feature by using information gain [1].

5) ***Artificial Neural Network (ANN):*** Artificial neural networks (ANN) are also used in other diabetes prediction studies [2]. In this research, along with different dense layers, dropout layers were also used to prevent overfitting. The dataset was trained with class weights to handle the class imbalance problem. The proposed ANN architecture is shown in Fig. 4.

D. Ensemble with Voting Method

Ensemble is an extremely efficient process of evaluating the model's performance [16]. Hard voting method was applied for ensembling in this research [17]. In this research, all the previous methods were ensembled with hard voting method. In terms of the hard voting method, the output equation is:

$$X_{pred} = mode(P_1(y), P_2(y), P_3(y), ..., P_n(y)) \quad (7)$$

In Eq. (7), X_{pred} represents the predicted value, $P_1, P_2, P_3, ...P_n$ represents the classifiers and y represents the input to the classifier. The $mode()$ generates the mode value of the predicted labels by the classifiers $P_1, P_2, P_3, .., P_n$.

V. Experiment and Results

A. Evaluation Metrics

In this research, several statistical measures odd ratio (OR), P-value, confidence interval (CI) were used as evaluation metrics to estimate the risk factors and effects of individual variables on the independent variable. Along with these, accuracy and AUC score were also used as evaluation metrics to evaluate the performance of our proposed model.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

where, TP, TN, FP, FN represents the True Positives, True Negative, False Positives and False Negatives values respectively [18].

The AUC score is the resultant score of (TP) True Positive and (FP) False Positive rate which outlines the sensitivity versus 1 - specificity that implies the overall effectiveness of the model.

B. Statistical Analysis of Risk Factors

In this research, the variables which have a p-value less than 0.05 were selected for analysis for their statistical significance. The *age, diastolic, cholesterol, blood-related diabetes, BMI etc.* independent variables are the most significant risk factors associated with diabetes class. Feature importance after handling the outliers by using Analysis of Variance (ANOVA) was also shown in the Fig. 3. The Fig. 3 validates the selected risk factors according to F-score.

After implementing statistical logistic regression and ANOVA methods, the intersection of two methods indicates that *age, blood-related diabetes, cholesterol, BMI* are the most significant risk factors of diabetes. These findings of risk factors were confirmed and validated by comparing with other studies [1]. The higher odd ratio is associated with higher chances of belonging to the diabetes class. Similarly, the lower odd ratio is related to lower chances of belonging to the diabetes class. According to Table I, *age, diastolic, cholesterol, BMI, gender, blood-related diabetes* variables can be acknowledged as high-risk factors of diabetes disease.

C. Results

The Table. II represents the comparison of accuracy and AUC scores with different CV values. The dataset was tested with different split/cv values 3, 5, 10 respectively. Several classification methods support vector machine, logistic regression, random forest, decision tree, artificial neural network and ensemble approach were used in this research. There are significant changes in the evaluation metrics shown in Table II. Among all of them, the random forest classification method (RF) achieved the highest accuracy (90%) and AUC score (0.98) than other classification methods.

TABLE I: Individual risk factors with their p-value, odd ratio, confidence intervals using logistic regression

Features	P value	Odd Ratio (OR)	5% CI for OR	95% CI for OR
Age	0.00	1.055761	1.047528	1.064160
Diastolic	0.00	1.025792	1.013955	1.037768
Cholesterol	0.00	1.032076	1.020663	1.043616
Blood-Rel-Diabetes	0.00	1.063944	1.052669	1.075340
Hemoglobin	0.0004	1.026249	1.011516	1.041195
Body-Mass-Index	0.0050	1.173713	1.55784	1.291980
AST	0.0013	1.021564	1.008342	1.034961
HDL	0.0022	1.018803	1.006727	1.031023
Gender	0.0048	1.023693	1.007169	1.040487
Phosphorus	0.0092	0.985706	0.975092	0.996436
Systolic	0.0299	0.986081	0.973684	0.998637
Protein	0.0306	0.986909	0.975190	0.998769
ALT	0.0309	0.985280	0.972104	0.998635

TABLE II: Comparison of accuracy and AUC score with different CV values

Methods	Evalution Metrics	CV		
		3	5	10
SVM	**Acc.**	.88	0.88	.89
	AUC	.89	.90	.90
LR	**Acc.**	.87	.88	.88
	AUC	.88	.83	.84
RF	**Acc.**	.90	.90	**.90**
	AUC	.89	.97	**.98**
DT	**Acc.**	.85	.86	.86
	AUC	.87	.95	.96
ANN	**Acc.**	.84	.84	.85
	AUC	.80	.80	.82
Ensemble	**Acc.**	.87	.88	.90
	AUC	.86	.89	.90

VI. Conclusion

The main purposes of this research are to identify and analyze the potential clinical risk factors of diabetes disease and to propose a system that classifies diabetes patients according to their different attributes efficiently. In this research, the total number of 37079 observations of patients with different attributes were analyzed. The outliers were handled by substituting the extreme values with median values rather than completely removing them. The class imbalance problem was handled by assigning class weights while training. The potential clinical risk factors were identified and investigated using Logistic Regression and ANOVA methods. As a result, age, blood related diabetes, cholesterol and BMI attributes are found as the most significant risk factors associated with diabetes. Multiple classification algorithms are implemented in the processed dataset and high accuracy, AUC scores are achieved and compared with other state of art methods. The random forest algorithm achieved the highest accuracy (.90) and AUC score (.98) in our proposed method. Though these scores are higher than other existing works, it still could be improved by applying other enhanced techniques.

References

[1] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, "Classification and prediction of diabetes disease using machine learning paradigm," *Health Information Science and Systems*, vol. 8, no. 1, p. 7, 2020.

[2] N. S. El_Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," 2018.
[3] P. Zimmet, K. G. Alberti, D. J. Magliano, and P. H. Bennett, "Diabetes mellitus statistics on prevalence and mortality: facts and fallacies," *Nature Reviews Endocrinology*, vol. 12, no. 10, p. 616, 2016.
[4] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, "An efficient convolutional neural network for coronary heart disease prediction," *Expert Systems with Applications*, p. 113408, 2020.
[5] M. Pilgrim and S. Willison, *Dive Into Python 3*. Springer, 2009, vol. 2.
[6] D. M. Hawkins, *Identification of outliers*. Springer, 1980, vol. 11.
[7] X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC medical research methodology*, vol. 14, no. 1, p. 135, 2014.
[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
[9] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.
[10] A. Smiley, D. King, J. Harezlak, P. Dinh, and A. Bidulescu, "The association between sleep duration and lipid profiles: the nhanes 2013–2014," *Journal of Diabetes & Metabolic Disorders*, vol. 18, no. 2, pp. 315–322, 2019.
[11] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, vol. 57. Austin, TX, 2010, p. 61.
[12] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.
[13] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, p. 51, 2011.
[14] B. Scholkopf, K.-K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2758–2765, 1997.
[15] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, vol. 31, no. 14, pp. 2225–2236, 2010.
[16] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
[17] K. Tumer and A. K. Agogino, "Ensemble clustering with voting active clusters," *Pattern Recognition Letters*, vol. 29, no. 14, pp. 1947–1953, 2008.
[18] K. A. Hasan and M. Al Mehedi Hasan, "Classification of parkinson's disease by analyzing multiple vocal features sets," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 758–761.