HYBRID PREDICTION MODEL FOR TYPE-2 DIABETES WITH CLASS IMBALANCE

Balasubramanian S
Electronics and Communication Engineering Department
PES University
Bengaluru, India
balavish580@gmail.com

Rishi Kashyap
Electronics and Communication Engineering Department
PES University
Bengaluru, India
rk.rishikashyap@gmail.com

Surya Teja CVN
Electronics and Communication Engineering Department
PES University
Bengaluru, India
cvnsuryateja@gmail.com

Anuradha M
Electronics and Communication Engineering Department
PES University
Bengaluru, India
anuradha@pes.edu

2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT) | 978-1-7281-8885-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLANT50963.2020.9355975

***Abstract*— Diabetes mellitus is a fast-growing disease affecting millions of people around the globe. According to the statistics obtained from the International Diabetes Federation, the growth of diabetic cases around the world is estimated to be around 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. Machine learning techniques can be used for early diagnosis of diabetes so that individuals can adapt to better lifestyle habits and diet plan to prevent further complications of symptoms. In this paper we have implemented an efficient model to classify type-2 diabetes using a hybrid approach. The dataset which is relatively imbalanced with respect to the number of diabetic or non-diabetic individuals was obtained from the Biostatistics program of Vanderbilt University, USA which contains parameters like Glucose, Age, Cholesterol, Weight, Waist/Hip ratio of 390 African-American individuals. The features from the dataset were obtained by filter and wrapper methods. The dataset was balanced by applying oversampling and undersampling methods. A voting ensemble model consisting of 5 algorithms was then used as a classifier with model validation done using stratified K-fold technique. The model's performance was assessed using recall, precision, F1 score and accuracy. The results obtained by comparison of the models with and without sampling have shown significant improvements in recall. The main motive is to improve recall and have reasonable precision at the same time.**

Keywords*— *Diabetes mellitus, Relative Imbalance, Machine learning, Feature Extraction, Ensemble learning, Oversampling, Undersampling

I. INTRODUCTION

Type-2 Diabetes is a long-standing condition that influences the body's mechanism in processing glucose. It leads to a situation where the body is not able to effectively utilize insulin. Insulin is a hormone produced by the pancreas which plays a central role in controlling the blood glucose levels in the body. Type-2 diabetes is usually prevalent among individuals who have excess body weight and lack of physical activities along with other reasons like genes, environmental factors, etc. Uncontrolled diabetes over a prolonged period of time can lead to damage to the heart, nerves, kidneys and eyes. According to WHO, in 2014, 8.5% of adults aged 18 years and older had diabetes. In 2012 high blood glucose was the cause of 2.2 million deaths and in 2016, diabetes was the direct cause of 1.6 million deaths [1]. Currently, there is no cure for diabetes but it can be controlled if early detection is accurately possible. The importance of machine learning in health care is its ability to process raw data outside the scope of human capabilities, and then reliably transform the analysis of the information into clinical insights that help doctors plan and dispense care, which ultimately results in better outcomes. Machine learning techniques can help physicians analyze patient symptoms to suggest better lifestyle changes. The hybrid model in this paper focuses on building a prediction model which can classify a person as diabetic or non-diabetic. A hybrid model combines of data processing techniques along with classification algorithms to form a combined model to solve the imbalance in a dataset and predict the target variable efficiently [2], [3], [4]. Pre-processing is an essential step in any machine learning model, which deals with parameters such as handling the missing values, extraction of important features, and data normalization. Repeated Stratified K-fold cross-validation is performed to preserve class distribution in each fold for better learning of the entire dataset. Feature selection was done using point bi-serial correlation, Boruta algorithm which is a wrapper-based method [5], and information gain. Different ML classifiers (Logistic Regression, Random Forest, Naïve Bayes, K-nearest Neighbor, and Support Vector Machines) were used in the Ensemble Voting Classifier. The grid search technique was performed for selecting the appropriate parameters for each model. To solve the class imbalance problem, we used sampling algorithms [6], [7], [8]. Ensemble Learning is a process through which multiple classifiers are used to find a solution to a problem. Sampling algorithms like SMOTE [9], ADASYN [10] and validation techniques were pipelined with various combinations of ML classifiers for obtaining the highest recall. Since recall is one of the most important performance metrics in medical applications, it was very important for us to build a robust model with a higher recall as we wanted to decrease the number of false negatives. The most challenging part of our project was to deal with the imbalanced nature of the dataset which was handled well by using the hybrid model whose components are as shown in Fig.1.

The rest of the paper is organized as follows: Section II discusses all the components involved in the hybrid model. In section III, the results obtained from the model are analysed. In section IV conclusion and future work are discussed.

978-1-7281-8885-0/20/$31.00 ©2020 IEEE

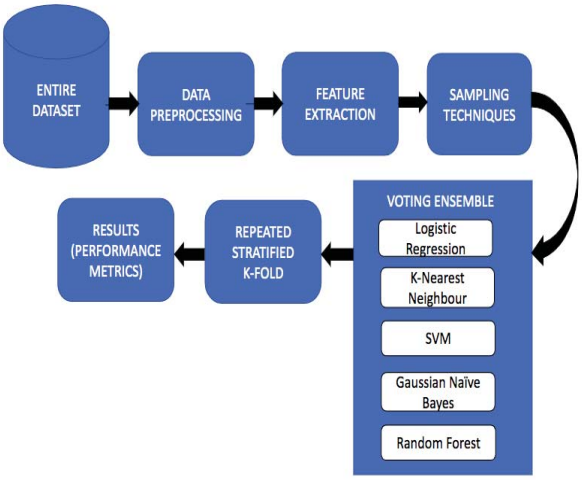

Fig. 1. Overview of the Hybrid Model

II. MATERIALS AND METHODS

A. Dataset

The original dataset is from Vanderbilt University, Tennessee, USA where the data samples were collected from rural African Americans in Virginia. This dataset was cleaned and modified by Robert Hoyt in data.world website [11]. Any patient without hemoglobin A1C was excluded. If their hemoglobin A1C was 6.5 or greater they were labelled with diabetes = yes. The threshold of 6.5 is the standard threshold used as per medical regulations. The dataset consists of 390 samples out of which only 15.38% (60 samples) belong to the minority class (type-2 diabetic) thus indicating the relative imbalanced nature of the dataset. The features in dataset were Glucose, Cholesterol, CHOL/HDL Ratio, Age, Height, Waist, Body Mass Index, Hip, Systolic pressure, Gender, Waist/Hip Ratio, and Diastolic pressure. The target column was Diabetes which had 'yes' and 'no' as its values. Gender and Diabetes columns were converted to binary values by one-hot encoding. T-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique used for envisioning high-dimensional data. This was applied in our dataset as shown in Fig.2 to get a better understanding of the outliers and the composition present.

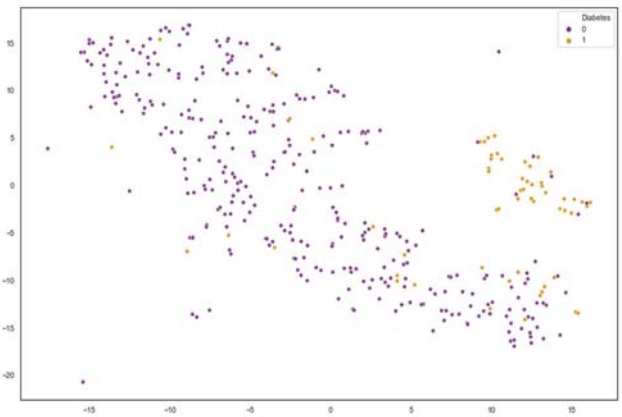

Fig. 2. 2-D T-SNE Components of the Data where '0' or purple dots represent non-diabetic samples and '1' or orange dots represent diabetic samples

B. Feature selection

The extraction of important features is essential to obtain better accuracy of the model, improve the interpretability of the model through graphs, and understand the contribution of each feature. Point Biserial correlation shown in Table I as a filter method that uses only the independent variables to measure relationships between each other. It is used when the target variable is a binary variable and the independent variables are continuous/discrete.

TABLE I: Bi-serial Correlation of featuers with Diabetes

Features	Correlation	pvalue
Cholesterol	0.204	4.8273e-05
Glucose	0.689	0
HDL Chol	-0.123	0.014
Chol/HDL Ratio	0.273	4.298e-08
Age	0.3019	1.146e-09
Height	0.0234	0.644
Weight	0.1629	0.00123
BMI	0.145	0.004
Systolic BP	0.1985	7.85e-05
Diastolic BP	0.049	0.331
Waist	0.2233	8.44e-06
Hip	0.1437	0.00443
Waist/Hip Ratio	0.1756	0.00049

The second feature selection method we have used is a wrapped method known as Boruta. Wrapper methods train a model using a subset of features and then using the results or predictions, the method tries to select the most important features. Boruta adds 'shadow' features which are shuffled (across different samples of a feature column) copies of the original features to the dataset and trains the model using Random Forest which gives a feature importance score known as Mean Decrease Accuracy/Impurity. The original features are then compared with the threshold which is the highest Mean Decrease Accuracy among the shadow features. The original features which are higher than the threshold are recorded as hits. This process is repeated with randomly shuffled shadow features created every time. Boruta then fits a binomial distribution and uses Z-score (number of standard deviations a data point is from the mean) to decide whether to accept or reject features based on statistical significance. The algorithm stops when all features are either selected or rejected or when the preset number of iterations is reached as shown in Fig. 3. Feature relevance is estimated by checking if a feature is doing better than a shuffled copy of itself which will imply that the particular feature is important to the model by statistical significance and not by chance. Information gain measures the reduction in entropy or increase in information about a dataset when the dataset is split according to a random variable or a feature. Features with higher information gain imply reduction in entropy or uncertainty about the class of the dataset. $IG(Y, x) = H(Y) - H(Y \mid x)$, where IG (Y, x) is the information for the dataset Y for the feature x. H(Y) is the entropy for the dataset and $H(Y \mid x)$ is the conditional entropy for the dataset given feature x. Features with higher information gain can partition the dataset better than those

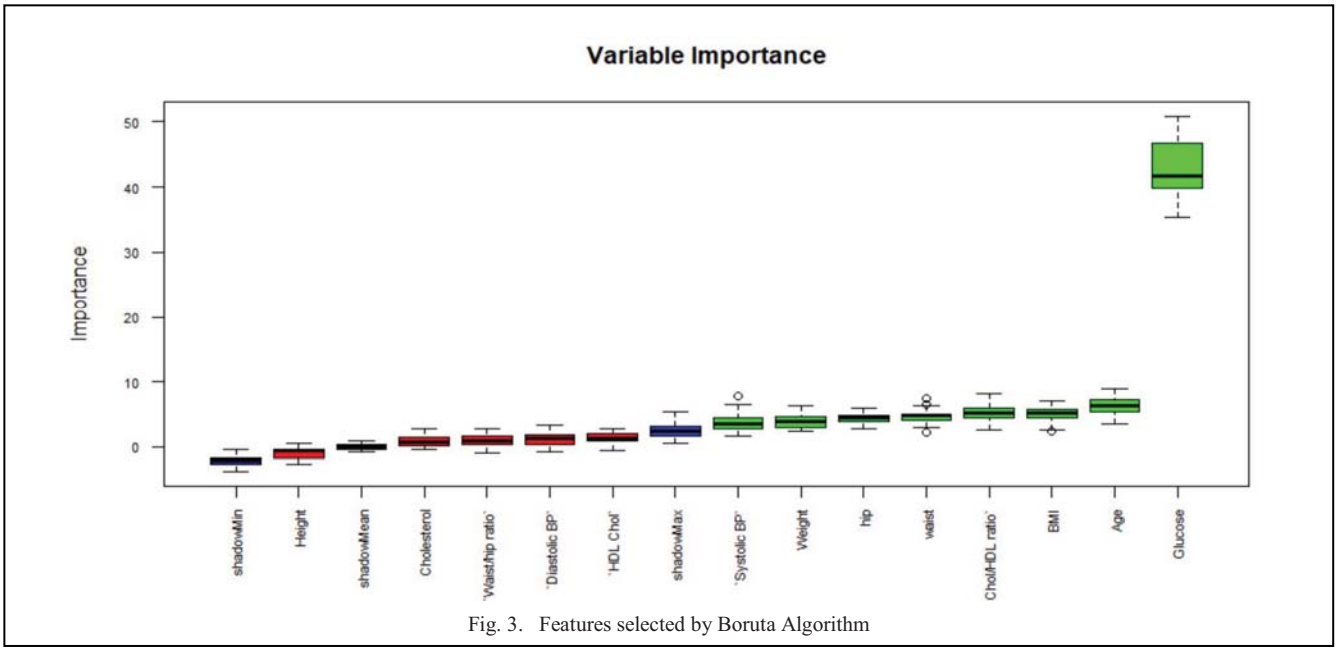

Fig. 3. Features selected by Boruta Algorithm

	attr_importance
Cholesterol	0.00000000
Glucose	0.23452500
HDL_CHOL	0.00000000
CHOL_HDL_RATIO	0.03686162
Age	0.05674891
Height	0.00000000
Weight	0.00000000
BMI	0.02274226
Systolic_BP	0.03155295
Diastolic_BP	0.00000000
waist	0.02567032
hip	0.00000000
Waist_Hip_ratio	0.00000000

Fig. 4. Information Gain of Features

with less information gain and is shown in Fig. 4. Combining the three feature selection methods discussed, the following features were selected: Glucose, Chol/HDL Ratio, Age, BMI, Systolic BP, Waist, Weight.

C. Cross validation

Validation is how a model's predicting ability is verified. Traditionally splitting the dataset into training or testing is used as validation but there are some inherent flaws to this approach as all the samples will not be tested fairly. The results are misleading and vary each time the dataset is split randomly. To perform validation in a fairer approach we use k-fold validation which creates folds of training and testing data with or without replacement allowing each sample to be present in training and testing data. K-fold assumes that class distribution is uniform, so it doesn't work on an imbalanced dataset. Therefore, we go for Repeated Stratified K-fold which preserves the class distribution in each fold of the training and validation test.

D. Sampling

Sampling has to be done on a training set and not on a validation set. If sampling is performed on the validation as well, then the artificially created samples will be predicted correctly thus leading to overfitting (poor generalization). The sampling techniques have been pipelined with the algorithm in the cross-validation setting.

1) SMOTE

Synthetic Minority Oversampling Technique (SMOTE) [9] is an over-sampling technique which can handle class imbalance by synthesizing new minority samples using existing minority data points. This technique was proposed to overcome the overfitting issue which is a major drawback in other oversampling techniques like Random Oversampling. New minority samples are synthesized in such a way that they are closer to the feature space of the existing minority samples. K-Nearest Neighbor algorithm is used to find the neighbors close to a randomly chosen minority sample. A random neighbor is then chosen and using linear interpolation a new minority sample is created between the two existing minority samples. This process continues until the distribution is balanced according to some preset parameter given by the user.

2) ADASYN

Adaptive Synthetic Sampling (ADASYN) [10] is an over-sampling approach. It is used to adaptively generate minority data samples. An example of the working of ADASYN is shown in Fig. 5.

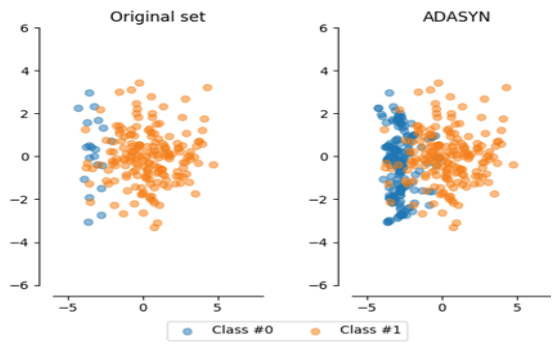

Fig. 5. ADASYN applied on a Dataset where Class #0 represents minority class and Class #1 represents majority class. Source: G. Lemaitre, F. Nogueira, D. Oliveira, C. Aridas, MIT

Generally, minority class instances that are tough to analyze and comprehend, are synthetically generated to help the model to learn better, which helps in reducing the learning bias introduced originally due to the imbalanced nature of the data distribution. For each minority sample, the density distribution is responsible for the generation of synthetic samples. ADASYN assigns weights to the minority class instances which help in generating synthetic sets for each sample.

SMOTE and ADASYN have a major difference in the generation of synthetic sample points for minority data points. In ADASYN, we consider a density distribution r_x which thereby decides the number of synthetic samples to be generated for a selected point, whereas in SMOTE, there's an identical weight for all minority points.

3) ENN AND ALLKNN (UNDERSAMPLING)

Edited Nearest Neighbors is employed for locating ambiguous and noisy examples in a dataset. It uses the K nearest neighbors' algorithm to determine which points to move out. ENN is performed on each sample belonging to the majority class. It removes all the samples which belong to minority class that are classified incorrectly, and those which are correctly classified remain as it is. It can also be used to remove points that are not present in its neighborhood. Repeated Edited Nearest Neighbors (RENN) runs ENN algorithm based on the user-specified number of times. At any time, the algorithm is rerun, more outlier points are removed. AllKNN is a modification of RENN which also enlarges the scale of the neighborhood being considered whenever it is run and has been shown in Fig. 6. [5]

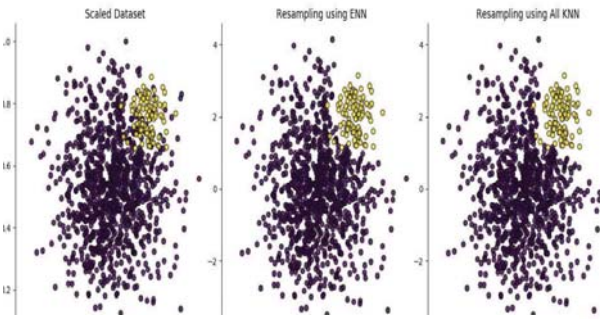

Fig. 6. ENN and ALLKNN applied on a dataset.

E. Algorithms

1. Logistic Regression

Logistic regression is a supervised learning algorithm that mostly focuses on classification problems. The main purpose of this technique is to find the best decision boundary to classify the sample points appropriately. It emphasizes in estimating the probability of the binary class (Non-diabetic (0) and Diabetic (1)) the data point belongs to by applying the sigmoidal activation function using a suitable threshold.

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad (1)$$

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic} \quad (2)$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \quad (3)$$

2. Support Vector Machines

SVM is a powerful supervised learning algorithm that is used for both classification and regression problems. It is useful in cases where the datasets have high dimensionality. The motive of this algorithm is to apply kernel methods on the data points and then use these transformed data points to find the appropriate hyperplane which maximizes the separation between the two class distributions to obtain an efficient model. It makes our model less prone to errors.

3. K-Nearest Neighbors

K-Nearest Neighbor is a supervised learning non-parametric algorithm used in classification applications that assigns the class (Diabetic (1) and Non-Diabetic (0)) to the given data depending upon its similarity with the other points present in its proximity. The KNN method helps in checking the similarity of a point with the other points present in its neighborhood with the help of metrics such as the Euclidean and the Manhattan distance formula. This method is very efficient in dealing with noise present in the data.
Suppose the two points present in a plane are A (x, y) and B (x_0, y_0), then the Euclidean distance and the Manhattan distance between them is calculated as given in (4) and (5).

$$Z_0 = \sqrt[2]{(x - x_0)^2 + (y - y_0)^2} \quad (4)$$

$$Z_0 = |x - x_0| + |y - y_0| \quad (5)$$

4. Naive Bayes

Naive Bayes is a supervised classification algorithm that assigns class (Diabetic (1) and Non-Diabetic(0)) to a particular sample given by a vector having certain feature values. This algorithm assumes the predictors to be independent i.e., each variable is important and is majorly responsible in contributing to the overall probability for classification. Naive Bayes uses conditional probability that is calculated using the following Bayes Theorem which is formulated as in (6). For each class, the posterior probability is calculated and the outcome will be the class having the highest posterior probability.

$$P(y \mid X) = P(\frac{X}{y}) * P(y) / P(X) \quad (6)$$

where $X = (x_1, x_2, \ldots\ldots, x_n)$

5. Random Forest

Random forest is the collection of multiple decision trees and gives the class by taking the majority of votes or classes outputted from many individual trees. Multiple training sets which are generated randomly create these multiple trees. For each tree and at each node split, certain

features are chosen to produce a random and unrelated tree. These trees output classes and when a mode is applied, it gives the overall class for that sample.

6. Ensemble Learning

Ensemble methods use multiple classifiers at once to produce improved results. In general, they produce better solutions compared to using a single classifier. In our model, we have used Voting Ensemble learning which involves summing the predictions made by the above classification models and the classifier model with majority votes is taken. If none of the predictions get more than half of the votes, we may conclude that the ensemble method could not make a good prediction score for that instance. Particularly, we have implemented a hard-voting technique where we predict the class with the largest sum of votes from models. Using an ensemble approach has given improved results to our model. In particular, recall values have increased significantly after applying hybrid techniques on an ensemble consisting of Logistic Regression, Random Forest, Naive Bayes, KNN, SVM classifiers. We have obtained a recall of 0.88 after applying our hybrid approach to the dataset.

F. Performance Metrics

- Accuracy: It is defined as the ratio of the total number of accurately classified instances, which is the number of true positives and true negatives present to the total number of points present in the dataset.
- Precision: It is defined as the number of diabetic instances predicted accurately among all the diabetic instances predicted by the model.
- Recall: It is defined as the number of diabetic instances predicted accurately out of all the diabetic cases present in the dataset.
- F1 Score: It tries to stabilize the precision and the recall metric by calculating the harmonic mean between them.

III. RESULTS

The results were obtained by implementing the experiment in jupyter notebook using python sci-kit modules. The results have been tabulated depending on various algorithms, combinations for sampling for different evaluation metrics. We have compared the results initially for each algorithm i.e. Logistic Regression, Random Forest, Naïve Bayes, K-NN and SVM (Table III). We then proceeded with the Voting Ensemble model with a comparison of individual oversampling and undersampling methods (Table IV) and then combined oversampling and undersampling techniques (Table V). Oversampling and Undersampling have significantly improved the recall while precision has decreased maintaining a good F1 score. ADASYN has given the best results when applied individually. A combination of SMOTE and ENN has given the best recall of 0.88 compared to all other methods. The overall combination of oversampling and undersampling techniques pipelined with repeated stratified k-Fold cross-validation (splits = 10, repeats = 3) improved recall significantly compared to individual application of each technique, thus giving us a good ensemble model which dealt well with the imbalanced nature of the dataset. The results in each cell of the table are the average of the k-fold scores.

TABLE II INDIVIDUAL CLASSIFIERS WITHOUT SAMPLING

Classifiers	Recall	Precision	F1 - Score	Accuracy
Logistic Regression	0.33	0.8	0.47	0.88
Random Forest	0.5	0.67	0.57	0.884
Naïve Bayes	0.75	0.64	0.69	0.897
KNN	0.42	0.83	0.56	0.897
SVM	0.5	0.75	0.60	0.897

TABLE III VOTING ENSEMBLE WITH EACH SAMPLING ALGORITHM

Sampling Techniques	Recall	Precision	F1 - Score	Accuracy
SMOTE	0.761	0.70	0.73	0.89
ADASYN	0.844	0.58	0.68	0.869
AllKNN	0.738	0.68	0.71	0.90
ENN	0.70	0.76	0.73	0.914

TABLE IV VOTING ENSEMBLE WITH COMBINATION OF SAMPLING ALGORITHMS

Sampling Techniques	Recall	Precision	F1- Score	Accuracy
SMOTE + ENN	0.88	0.555	0.680	0.86
ADASYN + ENN	0.862	0.545	0.667	0.90
ADASYN + AllKNN	0.86	0.555	0.676	0.898

IV. CONCLUSION

We have presented a hybrid model which uses sampling and ensemble learning for relatively imbalanced data. The proposed hybrid approach in our model emphasizes on getting a better recall since decreasing false negatives is the main focus in medical applications. This has been achieved by applying a combination of different undersampling and oversampling techniques (SMOTE, ADASYN, ENN), stratified k-fold, and voting ensemble thus helping us getting a very good recall value of 0.88 which is very high compared to other datasets such as Pima Indian (34.89% of the minority classes) and other diabetic datasets which are

balanced. The dataset has outliers which have a significant impact on the stochastic nature of the model (variation of scores when the algorithm is evaluated multiple times). Sampling methods and Stratified K-fold validation have been used to tackle this issue. Different sampling methods have been investigated on their performance on imbalanced data with outliers. The proposed model focuses on imbalanced datasets as oversampling techniques are not needed for balanced datasets. Future research can include methods to deal with outliers to improve the robustness, efficiency, and reduce the stochastic effect of the model.

REFERENCES

[1] WHO Diabetes Report, 8 June 2020. [Online] Available: https://www.who.int/news-room/fact-sheets/detail/diabetes/

[2] Rahul Joshi, Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm:Ensemble Approach," International Research Journal of Engineering and Technology, vol. 04, no. 10, Oct 2017.

[3] Ijaz, Fazal & Alfian, Ganjar & Syafrudin, Muhammad & Rhee, Jongtae. (2018). "Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest," Applied Sciences. 10.3390/app8081325.

[4] W. Chen, S. Chen, H. Zhang and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2017, pp. 386-390, doi: 10.1109/ICSESS.2017.8342938.

[5] Miron B. Kursa, Witold R. Rudnicki , "Feature Selection with the Boruta Package," Journal of Statistical Software, vol. 36, pp. 1-13, 2010.

[6] M. A. Arefeen, S. T. Nimi and M. S. Rahman, "Neural Network-Based Undersampling Techniques," IEEE Transactions on Systems, Man, and Cybernetics: Systems, doi: 10.1109/TSMC.2020.3016283.

[7] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI),Udupi, 2017,pp.7985, doi: 10.1109/ICACCI.2017.8125820.

[8] Yusof, Rozianiwati & Kasmiran, Khairul & Mustapha, Aida & Zin, N.A.M., "Techniques for handling imbalanced datasets when producing classifier models," Journal of Theoretical and Applied Information Technology, vol. 95, pp. 1425-1440, 2017.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal Of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[10] H. He, Y. Bai, E. A. Garcia, and S. Li, ''ADASYN: Adaptive synthetic sampling approach for imbalanced learning,'' in Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell., pp. 1322–1328, Jun. 2008.

[11] Vanderbilt Biostatistics Wiki, Dr John Schorling, Department of Medicine, University of Virginia School of Medicine. Williams JP, Saunders JT, Hunt DE, Schorling JB, "Prevalence of coronary risk factors among rural blacks: A community based study," Southern Medical Journal, vol. 90, pp. 814-820, 1997 [Online] Available: https://data.world/informatics-edu/diabetes-prediction/