Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest

Nur Ghaniaviyanto Ramadhan[1], Adiwijaya[2], Ade Romadhony[3]
School of Computing, Telkom University
Bandung, West Java
Indonesia

***Abstract*—Diabetes is a non-communicable disease that has a death rate of 70% in the world. Majority of diabetes cases, 90-95%, are of diabetes cases are type 2 diabetes which is caused by an unhealthy lifestyle. Type 2 diabetes can be detected earlier by using examination that contains diabetes-related parameters. However, the dataset does not always contain complete information, the distribution between positive and negative classes is mostly imbalanced, and some parameters have low importance to the decision class. To overcome the problems, this study needs to carry out preprocessing to improve detection precision and recall. In this paper, propose an approach on dataset preprocessing, which is applied to diabetes prediction. The preprocessing approach consists of the following process: missing value process, imbalanced data process, feature importance process, and data augmentation process. The data preprocessing process uses the median for missing value, random oversampling for imbalanced data, the Gini score in the random forest for feature importance, and posterior distribution for data augmentation. This research used random forest and logistic regression as classification algorithms. The experimental results show that the classification increased by 20% precision and 24% recall by applying proposed method and random forest method compared to without proposed method and random forest method.**

***Keywords*—Diabetes mellitus; data preprocessing; data augmentation; random forest; classification**

I. Introduction

Quoted from the 2016 WHO data, 70% of total deaths in the world are caused by diabetes, and 90-95% of diabetes cases are type 2 diabetes, which is mainly preventable because it is caused by an unhealthy lifestyle [1]. Diabetes mellitus is a chronic metabolic disorder caused by the pancreas not producing enough insulin or the body unable to use the insulin effectively [2]. In Indonesia, according to Basic Health Research (RisKesDas) in 2018 [2], people with diabetes from 2013 to 2018 increased gradually, where 6.9% of Indonesia population is diabetic. 69.6% of those with diabetes were undiagnosed, and 30.4% diagnosed. Meanwhile, in 2013, 5.7% were diabetic. As many as 73.7% of these people with diabetes, were undiagnosed and 26.3% were diagnosed. This data shows that diabetes mellitus is a dangerous disease since it can lead to various complications of other diseases, such as heart disease, kidney failure, stroke, and even paralysis and death [2].

The prevalence of diabetes mellitus (DM), based on a doctor's diagnosis in the population aged $\geqslant$ 15 years, is increased to 2% based on the report of Basic Health Research (RisKesDas) 2018 [2]. The largest DM sufferers are in the age range of 55-64 years and 65-74 years [2]. In 2018, the percentage of DM sufferers for female (1.8%) and male (1.2%) [2]. As for domicile areas, the percentage of DM sufferers in urban areas (1.9%) than in rural areas (1.0%) [2]. The highest estimate number of DM cases in Indonesia will occur in 2030, with a total population of 21.3 million [2]. Based on Basic Health Research (RisKesDas) diabetes data [2], undiagnosed patients can be detected beforehand. Diabetes detection could be performed by a doctor based on blood sugar and insulin levels or conducted automatically based on individual medical checkup data.

Prediction of diabetes diagnosis using data can determine whether the patients have diabetes or not. There are several studies that discussed diabetes diagnosis prediction based on data. Besides Pima Indian dataset [3-17], there is also data from Luzhou [4], Irvine [18], Kashmir [19,20], online questionnaire [21], and dr. Schorling [9,21]. There are various classification methods on diabetes diagnosis prediction like random forest, J48, naïve bayes (NB), support vector machine (SVM), logistic regression, neural network (NN), and K-Nearest Neighbors.

The explanation of paper contributions taken from some of the shortcomings of previous research is applied to diabetes prediction. In [8] discusses the process of missing value using the median in general and feature selection using this importance index and permutation importance index. Paper [10] discusses the problem of imbalanced data using general random oversampling. In [23] discusses data augmentation techniques for the problem of imbalanced data using a gaussian distribution.

In this paper, the contribution is firstly to replace the value of outliers using median for every six rows, secondly for imbalanced data using oversampling technique namely Random Oversampling by combining three imbalanced features, third for the selected feature process using feature importance technique in random forest model with Gini index value, fourth for data augmentation process using posterior distribution technique where latent data (Y) uses Karya Medika data. Comparison of the contribution of this study with several other studies can be seen in Table I. This paper aims to improve the precision and recall outcomes in diabetes prediction using data preprocessing.

TABLE I. COMPARISON OF CONTRIBUTIONS

Author	Missing Value	Imbalanced Data	Feature Selection	Data Augmentation
[8]	Yes	No	Yes	No
[10]	No	Yes	No	No
[23]	No	No	No	Yes
This Study	Yes	Yes	Yes	Yes

II. LITERATURE REVIEW

In previous studies, the classification and prediction of DM with Pima Indian data have been carried out using several machine learning methods. However, only a few studies discussed about preprocessing on Pima Indian dataset. The problem of missing value is discussed in a limited number of papers [8,13,14,15,17]. The problem of imbalanced data [10,11,17] and of feature selection [5,9,10,14] have been discussed too. Several models have been used in data preprocessing, such as missing value using median [8], Interquartile Range [13,14], mean [15], and Naive Bayes [17]. In imbalanced data, there is Synthetic Minority Over-sampling [10,11], Random Oversampling [10], and Adaptive Synthetic Sampling [17]. Meanwhile, in feature selection, there is Principal Component Analysis [5,9], Maximum Relevance and Minimum Redundancy [5], Fisher Discriminant Ratio [9], Analysis of Variance [9], Information Gain [10], and forward backward [14] models.

According to several prior studies on diabetes prediction, important factors that contribute to classification accuracy are imbalanced data, the presence or absence of missing values, and features that affect the results [4,7,11,13-17,19-22]. In addition, paper explains that data augmentation can improve the accuracy of diabetes prediction [23]. Data augmentation is an algorithm used to augment the observed X data with a quantity of Y, referred to as latent data [24]. In the Pima Indian dataset, imbalanced data occurs in the class label. Imbalanced data is a problem related to the performance of learning algorithms faced with underrepresented data, and the slope of the class distribution is severe [25]. The missing value is a problem that replaces the null value in a variable [9]. The maximum limit for missing value varies from 5-10% and 50% [26]. Feature selection is an important problem in machine learning since it gets the most informative features [9].

III. METHODS OF RESEARCH

To improve the precision and recall outcomes in DM prediction analysis, this research proposed data preprocessing on the binary classification of DM type 2. Fig. 1 shows the proposed system diagram performed in this study whilst Fig. 2 shows the proposed system in more detail.

A. Dataset

This study used two different diabetes datasets, namely Pima Indian and Karya Medika. Kumar et al. provides Pima Indian dataset description [14].

For data augmentation, other data with the same characteristics with the Pima Indian data were used. In this paper, this research used datasets of DM from Karya Medika in January to April 2020. This dataset was taken from an individual sample of Indonesians from the Slawi region, Central Java with a sample size of 630 and has nine features include class labels. In Karya Medika dataset also has problems with preprocessing. Table II shows the dataset of Karya Medika, where the body mass index (BMI) value can be obtained using the formula (1). The BMI formula was used during the data augmentation process, which will become a new feature called BMI.

$$\text{BMI} = \frac{\text{Weight (kg)}}{\text{Height (meter)} * \text{Height (meter)}} \quad (1)$$

Table III presents a baseline of two different datasets, which used as a comparison. Same characteristics found in these two datasets are glucose level, diastolic blood pressure, BMI, age, and class types.

Table IV presents the features of missing values. Pima Indian has more outliers than Karya Medika dataset.

Table V compares features with an imbalanced value of two different datasets, where Karya Medika has more imbalanced features than Pima Indian.

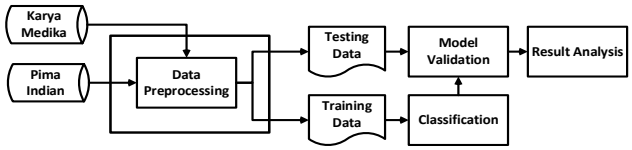

Fig. 1. Proposed System Diagram.

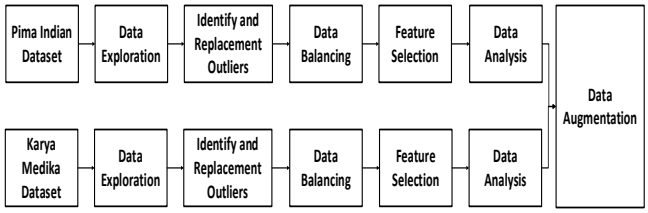

Fig. 2. Proposed System Diagram in Detail.

TABLE II. KARYA MEDIKA DATASET

No	Attribute Description	Unit	Attribute Type
1	Glucose	mg/dl	Numeric
2	Male (1) or Female (0)	-	Nominal
3	Systolic Blood Pressure	mmHg	Numeric
4	Diastolic Blood Pressure	mmHg	Numeric
5	Height	Kg	Numeric
6	Weight	cm	Numeric
7	Age	Year	Numeric
8	Fasting (1) or No-Fasting (0)	-	Nominal
9	Diabetes (1) or No-Diabetes (0)	-	Nominal

TABLE III. DATASETS STATISTIC COMPARISON

Dataset	Number of Observation	Number of Feature	Continuous Feature	Binary Feature	Categorical Feature	Class Type	Classes
Pima Indian	768	8	8	0	0	Binary	2
Karya Medika	630	8	6	2	0	Binary	2

TABLE IV. MISSING VALUES COMPARISON

Dataset	Feature of Missing Values					
Pima Indian	Pregnancies 14%	Glucose 0.65%	Diastolic Blood Pressure 4.5%	Skin Thickness 29.5%	Insulin 48.6%	BMI 1.4%
Karya Medika	Systolic Blood Pressure 10.63%	Diastolic Blood Pressure 10.63%	Height 8%	Weight 7.46%	Age 4.4%	

TABLE V. IMBALANCED DATA COMPARISON

Dataset	Feature of Imbalanced Data					
Pima Indian	Diabetes (268)			No Diabetes (500)		
Karya Medika	Male (264)	Female (366)	Fasting (452)	No Fasting (178)	Diabetes (290)	No Diabetes (340)

B. Outliers Identification and Replacement

In this process, identifying each dataset is whether each feature has a null value, as represented in NaN/{}/0. After determining the outliers, the value is calculated. The process of replacing the null value with a statistical method or machine learning model is carried out. This process can be referred to as missing imputation. In this study, the missing imputation process uses the median value. The median value is chosen since it only takes the middle value in the calculation process without considering other values. In this step, this research aim to find the median value with an even number of data because the imputation process will be carried out every six rows.

C. Data Balancing

In this step, this research uses the random oversampling (ROS) method, which will carry out the oversampling process for minor data to increase percentage. The ROS method was chosen because the data problem used occurred imbalanced in the minority class, which was suitable to use the oversampling method. Applying a re-sampling strategy to the pre-processing data process to obtain a more balanced data distribution is an effective solution to the imbalance problem [27]. ROS method also involves randomly duplicating samples from a minority class and adding them to the training dataset [27]. The process will also see the imbalanced ratio, which calculates the data set from two certain classes. The imbalanced ratio then can be calculated using formula (2) [12].

$$(\text{Instance Minority}/\text{Instance Majority}) \quad (2)$$

where, instance minority is the number of distributions of label class that is less, while instance majority is the number of distributions of label class that is more. So, to find the imbalanced ratio based on [12], namely, the distribution of minority divided by the distribution of the majority.

Table VI shows the imbalanced ratio in the two datasets of this study. The imbalanced ratio has a scale of 0-1, where if the result is close to the value of 1, then the class has only a few imbalanced data.

TABLE VI. IMBALANCED RATIO IN DATASET

Dataset	Feature	Imbalanced Ratio (%)
Karya Medika	Gender	0.72
	Fasting or No Fasting	0.39
	Class Label	0.85
Pima Indian	Class Label	0.53

D. Feature Selection

There are three feature selection techniques: univariate selection, feature importance, and correlation matrix with heat maps [28]. In this paper, this research performs the feature importance technique to solve predictive analysis problems [29]. This technique is carried out to provide a score for each feature against the label class, whether it has high or low attachment.

$$Gini = 1 - \sum_i^c P^2 i \quad (3)$$

where c is the number of values in the target attribute (number of classification classes) and P is the sample portion for the class i (diabetes and no diabetes).

In this paper, the feature importance technique uses the random forest model. Therefore, the calculation process uses the Gini function, as shown in equation (3) in the random forest model. The value of c is two classes, namely diabetes or no diabetes. Then Pi is the sample size for diabetes and no diabetes.

E. Data Augmentation

This study proposed other techniques in addition to using oversampling techniques on class balance problems. The proposed technique is data augmentation. This study uses data augmentation for the problem of lack of varied samples in the Pima Indian dataset, which will be done with additional data using dataset Karya Medika. The data augmentation process will provide a way to increase inference based on the posterior distribution [24]. The posterior distribution is shown in formula (4).

$$P(\vartheta|Y) = \int_X P(\vartheta|Y,Z)\, P(\vartheta|Z,Y) dZ \quad (4)$$

where P(ϑ|Y) denotes the posterior density of parameter ϑ given the dataset Pima Indian observation, P(ϑ|Z,Y) denotes the predictive density of the Karya Medika data Z given Pima Indian, and P(ϑ|Y,Z) denotes the conditional density of ϑ given the data augmented X=(Y,Z) namely augmented posterior [24].

This study will augment data from Pima Indian using Karya Medika data to produce data augmented (X) containing feature characteristics with similarities in both datasets. This study calculates the relative difference using equation (5) to calculate the increase in the original data changes with augmentation data.

$$RD = \left(\frac{\text{Result Augmentation} - \text{Result Original}}{\text{Result Original}}\right) \times 100\% \quad (5)$$

where Result Augmentation (RA) is the result after the Pima Indian dataset augmentation process with Karya Medika dataset. Result Original (RO) is the result before the dataset augmentation process is carried out.

Relative Difference (RD) is a measure that shows the percentage increase when an enlarged data set is used compared to the original data [23]. The instance value used from Karya Medika dataset for augmentation is 100%.

F. Classification

This process is data classification using supervised machine learning methods, namely random forest (RF) and logistic regression (LR), to see the precision and recall. This process also separates training data from data testing. This study split the dataset to train and to test dataset with the ratio of 75:25. Both models have been widely applied with success in various disciplines for classification and regression purposes [30]. The Random Forest used is entropy, as shown in equation (6), where *c* and *Pi* have been described above.

$$Entropy = \sum_i^c Pi \log 2\ Pi \quad (6)$$

IV. Result and Discussion

This section will discuss the results of the proposed method and analyze the results. Three experiments were conducted separately. First, using the Pima Indian dataset by applying the preprocessing algorithm and then conducting classification. Second, using the Karya Medika dataset by applying the preprocessing algorithm and then conducting classification. Third, using the augmented dataset by applying the preprocessing algorithm and then conducting classification.

As shown in Table VII, the Pima Indian dataset by applying the proposed preprocessing was compared to the original preprocessing increased by using RF and LR classification methods. In Karya Medika dataset by applying preprocessing proposal was compared to original preprocessing increased by using RF classification method compared to LR classification. The results indicated to be different in the Karya Medika dataset with the oversampling process of three features using LR experienced a decrease in precision of 7% and F1 score of 1% compared to the original data.

TABLE VII. Results

Dataset	Preprocessing	Classification	Result		
			Precision Diabetes (%)	***Recall Diabetes (%)***	***F1-Score Diabetes (%)***
Pima Indian	Original	Random Forest (RF)	70	53	61
	Median every six rows, Random Oversampling 1 Features Imbalanced (Class Label), Gini Index Rank 9 Features		83	88	86
	Original	Logistic Regression (LR)	77	55	64
	Median every six rows, Random Oversampling 1 Features Imbalanced (Class Label), Gini Index Rank 9 Features		78	74	76
Karya Medika	Original	Random Forest (RF)	88	83	85
	Median every six rows, Random Oversampling 3 Features Imbalanced (Gender, Fasting, and Class Label), Gini Index Rank 9 Features		98	99	98
	Median every six rows, Random Oversampling 1 Features Imbalanced (Class Label), Gini Index Rank 9 Features		92	98	95
	Original	Logistic Regression (LR)	88	81	84
	Median every six rows, Random Oversampling 3 Features Imbalanced (Gender, Fasting, and Class Label), Gini Index Rank 9 Features		81	84	83
	Median every six rows, Random Oversampling 1 Features Imbalanced (Class Label), Gini Index Rank 9 Features		91	84	87
Augmented Data	Median every six rows, Random Oversampling 1 Features Imbalanced (Class Label), Gini Index Rank 5 Features	Random Forest (RF)	94	96	95
	Median every six rows, Random Oversampling 1 Features Imbalanced (Class Label), Gini Index Rank 5 Features	Logistic Regression (LR)	82	67	74

In augmented datasets increased by using RF classification method when compared with the Pima Indian original dataset and original dataset of Medika Works. Different results in augmented datasets using LR classification method when compared to the original dataset of Karya Medika experienced a decrease in results. However, if augmented datasets using LR classification method compared to the Pima Indian original dataset experienced an increase in results.

So, for the overall RF classification experimentation is superior to LR by applying the proposed method of preprocessing. This happens because the LR classification performed better when the number of noise variables was less than or equal to the number of explanatory variables. Therefore, if the LR classification results were going to be improved, it was necessary to note the importance of each variable used.

Based on augmented dataset results, it showed that Karya Medika data was able to make the predicted results of DM in the Pima Indian dataset increase. However, the Pima Indian dataset was unable to make the DM prediction results in the Karya Medika dataset increase. The F1 score showed that after the imbalanced data method was applied, the results for the minority class increased. Precision and recall results show that the importance of preprocessing the dataset in advance to improve the predicted results of diabetes mellitus.

For the most important preprocessing process to improve diabetes detection results is missing value and balancing class. This is because the missing value process is a built-in problem in which the data used there is a value of 0/NaN/{} in which the value must be replaced with a guess of value, if the missing value is not executed then there will be an error during the classification process. Meanwhile, the process of balancing the class has a great influence on the results of diabetes detection because the ratio of the class of diabetes used as a sample tends to be less than the class that is not diabetic. This is evident after the process of balancing the class of classification results obtained has increased significantly.

V. Conclusion

Based on the results of the implementation and analysis, it can be concluded that this study on the preprocessing process can improve the precision and recall results of the random forest classification model. The results indicate that the classification method using a random forest is superior to logistic regression. The proposed preprocessing method can also be applied to the other augmentation result data from two different datasets by looking at the data characteristics. For the most important preprocessing process to improve diabetes detection results is missing value and balancing class. Data augmentation can also improve the precision and recall results of each original data. This study found that the data quality used is better for Karya Medika dataset than Pima Indian.

Further works need to be conducted by adding some other parameters to the data with samples such as insulin levels, history of diseases suffered, family history of people with diabetes or not, and other parameters related to diabetes. In addition, further studies can also be done using other medical data such as patient data on cancer, heart disease, stroke, and others, or using other combinations of machine learning models in any preprocessing or classification process.

Acknowledgment

The authors would like to thank the Pima Indian and the Karya Medika Lab for the datasets on diabetes. The authors declare no competing financial interest.

References

[1] Ministry of Health RI. “CEGAH, CEGAH, dan CEGAH: Suara Dunia Perangi Diabetes, Accessed December. 12, 2019, https://www.kemkes.go.id/article/view/18121200001/prevent-prevent-and-prevent-the-voice-of-the-world-fight-diabetes.html, 2018 (In Indonesian).

[2] Khairani, InfoDatin (Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia), Hari Diabetes Sedunia, Accessed December, 12, 2019. PDF article, https://pusdatin.kemkes.go.id/download.php?file=download/pusdatin/infodatin/infodatin-Diabetes-2018.pdf, 2018 (In Indonesian).

[3] Vigneswari, D., et al, Machine Learning Tree Classifiers in Predicting Diabetes Mellitus, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE, 2019.

[4] Zou, Quan, et al, Predicting diabetes mellitus with machine learning techniques, Frontiers in genetics 9, 515, 2018.

[5] Mercaldo, Francesco, Vittoria Nardone, and Antonella Santone, Diabetes mellitus affected patients classification and diagnosis through machine learning techniques, Procedia computer science, 112, 2519-2528, 2017.

[6] Tafa, Zhilbert, Nerxhivane Pervetica, and Bertran Karahoda, An intelligent system for diabetes prediction, 2015 4th Mediterranean Conference on Embedded Computing (MECO), IEEE, 2015.

[7] Saru, S., and S. Subashree, Analysis and Prediction of Diabetes Using Machine Learning, International Journal of Emerging Technology and Innovative Engineering, 5, 4, 2019.

[8] Maniruzzaman, Md, et al, Accurate diabetes risk stratification using machine learning: role of missing value and outliers, Journal of medical systems, 42, 5, 92, 2018.

[9] Ijaz, Muhammad, et al, 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, 8, 8, 1325, 2018.

[10] Shi, Zhan, Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification, IOP Conference Series: Materials Science and Engineering, IOP Publishing, Vol. 719, No. 1, 2020.

[11] Tyagi, Shivani, and Sangeeta Mittal, Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning, Proceedings of ICRIC 2019, Springer, Cham, 209-221, 2020.

[12] Devi, R. Delshi Howsalya, Anita Bai, and N. Nagarajan, A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms, Obesity Medicine, 17, 100152, 2020.

[13] Nnamoko, Nonso, and Ioannis Korkontzelos, Efficient treatment of outliers and class imbalance for diabetes prediction, Artificial Intelligence in Medicine, 104, 101815, 2020.

[14] Raghavendra, S., and J. Santosh Kumar, Performance evaluation of random forest with feature selection methods in prediction of diabetes, International Journal of Electrical & Computer Engineering, 10, 2088-8708, 2020.

[15] Rajni, Amandeep, RB-bayes algorithm for the prediction of diabetic in PIMA Indian dataset, International Journal of Electrical and Computer Engineering (IJECE), 9, 6, 4866-4872, 2019.

[16] Azrar, Amina, et al, Data mining models comparison for diabetes prediction, Int J Adv Comput Sci Appl, 9, 2018.

[17] Wang, Qian, et al, DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values, IEEE, Access 7, 102232-102238, 2019.

[18] Kumar, N. Komal, et al, An Optimized Random Forest Classifier for Diabetes Mellitus, Emerging Technologies in Data Mining and Information Security, Springer, Singapore, 765-773, 2019.

[19] Mirza, Shuja, Sonu Mittal, and Majid Zaman, Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree, International Journal of Applied Engineering Research, 13, 11, 9277-9282, 2018.

[20] Shuja, Mirza, Sonu Mittal, and Majid Zaman, Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE, Advances in Computing and Intelligent Systems. Springer, Singapore, 195-211, 2020.

[21] Wu, Han, et al, Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, 10, 100-107, 2018.

[22] Abd Rahman, Muhammad Hafiz Fazren, Wan Wardatul Amani Wan Salim, and Mohd Firdaus Abd Wahab, Risk Prediction Analysis For Classifying Type 2 Diabetes Occurrence Using Local Dataset, Biological and Natural Resources Engineering Journal, 3, 1, 48-61, 2020.

[23] Moreno-Barea, Francisco J., José M. Jerez, and Leonardo Franco, Improving classification accuracy using data augmentation on small data sets, Expert Systems with Applications, 161, 113696, 2020.

[24] Tanner, Martin A, Tools for statistical inference: observed data and data augmentation methods, Springer Science & Business Media, Vol. 67, 53, 2012.

[25] He, Haibo, and Edwardo A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and data engineering, 21.9, 1263-1284, 2009.

[26] Madley-Dowd, Paul, et al, The proportion of missing data should not be used to guide decisions on multiple imputation, Journal of clinical epidemiology, 110, 63-73, 2019.

[27] Branco, Paula, Luis Torgo, and Rita Ribeiro, A survey of predictive modelling under imbalanced distributions, arXiv preprint, arXiv:1505.01658, 2015.

[28] Verma, Anurag Kumar, and Saurabh Pal, Prediction of skin disease with three different feature selection techniques using stacking ensemble method, Applied Biochemistry and Biotechnology, 1-20, 2019.

[29] Kuo, Kuang-Ming, et al, A multi-class classification model for supporting the diagnosis of type II diabetes mellitus, PeerJ, 8, e9920, 2020.

[30] Couronné, Raphael, Philipp Probst, and Anne-Laure Boulesteix, Random forest versus logistic regression: a large-scale benchmark experiment, BMC bioinformatics, 19, 1, 270, 2018.