Predicting Hypertension using Machine Learning: A Case Study at Petra University

Yasmin Sakka[1], Dina Qarashai[2], Ahmad Altarawneh[3]
Faculty of Administrative and Financial Sciences-Management Information Systems Department,
University of Petra, Amman, Jordan[1, 3]
Healthcare Center, University of Petra, Amman, Jordan[2]

***Abstract*—Hypertension is a key cardiovascular disease risk factor (CVD). Identifying these high-risk individuals is crucial since it would save time and money before using any sophisticated, invasive, or costly diagnostic technique. This endeavour may be accomplished in part with the use of modern machine learning techniques. Specifically, a prediction model may be created based on several easily-obtained, non-invasive, and inexpensive indicator characteristics of high-risk individuals. This research is an effort to forecast hypertension risks based on Petra University's population. This case study was done between 2019 and 2020 at Petra University. Using hospital-visited patients' medical records, the gathered data was used to develop a model. The research comprised a comprehensive dataset of 31500 patients, comprising 12658 hypertension cases and 18842 non-hypertensive cases. SMOTE was used as a dataset for the categorization of hypertension. The SMOTE-k-nearest neighbour prediction model performs exceptionally well, as evidenced by its excellent performance (83.9% classification accuracy, 85.1% specificity, 83.3% sensitivity, and 89.6% AUC) when compared to other classifiers using 10-fold cross-validation with full features and no oversampling on the hypertension dataset. The data extracted from Petra University Health Center is considered to be very helpful for ML and is availed to produce a decision tree to identify the data related to hypertension.**

Keywords*—Hypertension; machine learning; medical records; sensitivity; specificity

I. INTRODUCTION

There are 8.5 million fatalities worldwide attributed to hypertension, making it the leading cause of cardiovascular disease. Among these, 88% occur in low- and middle-income nations [1]. Access to unhealthy foods, sedentary lifestyles, and rural-urban migration are major contributors to the rising rate of hypertension in South Asia [2, 3]. The identification, treatment, and management of hypertension are similarly lowest in South Asia, and there has been little progress in these areas over the previous three decades [1]. Due to low levels of screening awareness, many cases of hypertension in South Asia go unreported [4]. Heart disease, stroke, renal failure, and premature death may all result from hypertension, although they are often avoidable with early diagnosis and treatment.

Physical inactivity, low levels of knowledge, smoking, an unhealthy diet, a lack of access to healthcare, and the high cost of drugs all play roles in explaining why hypertension is more common in South Asia [4-6]. However, most of the studies utilized unreliable methods to assess risk, had insufficient sample sizes and failed to adequately reflect the general population. The Framingham Risk Score for predicting coronary heart disease [7] and the American College of Cardiology/American Heart Association (ACC/AHA) Pooled Cohort Equations Risk Calculator are just two examples of risk prediction models that have been successfully used to identify and stratify patients according to risk factors and initiate preventative therapies [8].

When it comes to the creation of risk stratification tools for the diagnosis of cardiovascular illnesses, machine learning (ML) methods have recently been demonstrated to outperform classic statistical approaches [9,10,11]. The field of computer science known as "machine learning" allows computers to "learn" independently of explicit programming and handle massive datasets with complicated relationships. In contrast to more conventional statistical approaches, ML algorithms are not dependent on causal inference; nonetheless, this does not make them any less important for assessing causal effects in observational research. When compared to more conventional statistical methods, ML eliminates bias, the automated handling of missing variables with little changes to the original data, the mitigation of confounding factors, and the balancing of data [10]. More importantly, traditional statistical methods often fail when used in "big data" situations, whereas ML techniques succeed. Therefore, machine learning techniques may be used to create automated systems for illness prediction, decision support, and estimating hypertension prevalence in a community [10].

A dearth of research integrating socio-demographic and clinical data with signal processing, which might improve model performance, was observed in a recent review of ML approaches in hypertension identification [12]. ML algorithms were employed in prior work to automatically classify individuals with hypertension based on their unique phenotypes, but this analysis lacked socio-demographic information [13]. In a separate study, researchers in India used information gathered by community health workers from 2,278 individuals to create ML risk classification algorithms for diabetes and hypertension [14]. ML was employed in two different studies in China to analyze EHRs for signs of hypertension [15].

In the previous studies related to the machine learning models, the success lies in the complex pattern which can predict the data which is not observed in any other model. The concept of machine learning develops the power of interpretability by introducing the strategies of the description, predictive relevant outcome and the predictive description of

relevant data. These are all presented in the form framework providing an interpretation of the discussion. The data obtained by machine learning provide the accuracy and the relevancy of the demand of human audience [16]. The study proposed by Alaa et al. [17] stated that people who are at a higher risk of cardiovascular attack need more preventative measures for the protection of their hearts. The clinical guideline has presented the model for risk prediction to identify the optimal performance of the patient groups. The data is collected from the machining process to improve the complex learning outcome and the complex interactions. Automatic machine learning (Autoprogress) is one of the techniques used to improve the traditional approaches and the risk prediction of CVD. This increases the accuracy of the data which is obtained from the large population.

However, no research has yet employed ML models to predict hypertension at the population level and verified the models using big datasets in South Asian nations, despite the progress in ML models for individual risk prediction for various illnesses. The present study aimed to determine the model's predictive abilities for hypertension and sought to employ ML techniques to discover characteristics related to hypertension diagnosis.

Therefore the remaining paper is based on an academic article structure which is as follows: in the first part, all the studies and models related to the study have been outlined. Then in methodology, the SMOTE model theory with its implication has been described; how the study has been processed. All the findings and the parameters have been mentioned in detail. Finally, for future research, the direction has been provided with the study limitation, conclusion and proposed ideas based on the study findings.

II. Related Work

Artificial intelligence has a great impact on the field of medicine, diagnosis of the disease or the treatment of diseases. Scientific development has introduced several strategies with the help of advancements in medical technologies and the proposed models. In this study, the use of machine learning and its algorithms with the induced models have displayed the data of the study. This method of study has embedded its pipelines in the data mining procedure. The rules of decision-making and the learning pattern of the algorithms can extract the data. The study uses the predictive models of the SMOTE model comprising of logistic regression, K-nearest neighbour, Naïve Bayes method, REP tree, Random forest, Artificial neural network, and Repeated incremental pruning to produce error reduction (RIPPER). Previous studies have been used to state the classical strategies of statistics. In the study of Dagliti et al. [2018]. The predictive constructive model which has been used is the model validation and the predictive model construction. The missing data have been handled by the Random Forest (RF) to correct the imbalances in the logistic regression for suitable strategies. In this study, the parameters of the feature selection of retinopathy, neuropathy, and neuropathy in different scenarios have been identified in diabetic hospitals. The considered variables are hypertension and hemoglobin, gender, age, and smoking habit. The compilation of this analysis has led to different models to translate easy clinical practices [18]. The major risk of cardiovascular disease is hypertension. The classification of hypertension in traditional Chinese medicine has too much effective methodology according to the syndromes of the patients. The data mining has the multi-learning model of labelling like BrSmote SVM which was developed to deal with the diagnosis and the class unbalancing of the data set. The experiment represents that it has diversified the evaluation of the average precision, one error, and the coverage with the loss ranking [19].

The classification of the multi-labelling is one of the problem classifications which is used for the data mining of the patient and the learning of the strategies of machine learning. This practice has developed the fast accumulation of clinic data [20]. The technique motivated the task of the medical diagnosis and the text categorization. These are divided into two main categories which are the transformation of the problem method and the second is the adaptation of the algorithm method. The problem transformation has more than one regression and hence they depended on the learning algorithm's methods where the algorithm method learning depends directly on the research data for the evaluation [21].

III. Material and Methods

A. Study Design

This case study was done at Petra University from 2019 to 2020. Using medical records of people who went to the hospital, the data was collected to build a model. The study began after the ethics committee gave its approval, and it was done according to the rules of the Declaration of Helsinki. There were a total of 31500 complete cases in the study. Of these, 12658 had high blood pressure and 18842 did not. All of the people who signed up gave their medical record number and baseline and anthropometric information. During the study, patient information was kept secret at all times. The ML-based model which has been used is the SMOTE model consisting algorithmic method which is used to assemble all the predictive models by machine learning pipelines. These consist of feature processing, comprising data imputation, calibration and the classification of the algorithmic predictive models [17].

The basic and familiar risk factors are their age (in years), gender, blood pressure (in mmHg), blood glucose (in mmol/lit), urea (in mmol/lit), and creatinine (in umol/l) which were all used as parameters. We removed data where 30% of the values were missing. After the initial data screening, all of the records went through the labelling process so that the dataset could be cleaned up. This gave us the labelled standard records. This will be used to build a machine-learning model that will use the patient's blood biochemical tests to diagnose hypertension. The model's features are chosen based on how easy they are to use, to collect data, and they must be statistically significant for univariate logistic regression ($p < 0.05$). All the predictive parameters were assessed by using the area under the receiver operating curve of characteristic (AUC-ROC). The overall SMOTE model improved all the risk predictions by 10-fold cross-validation classification performance evaluation of different classifiers on the hypertension dataset on full features without oversampling and 10-fold cross-validation classification performance evaluation

using ten classifiers on the hypertension dataset on full features using oversampling. In this experiment, the dataset was run on ten machine learning classifiers using 10-fold cross-validation. 90% of the data was utilized to train the classifiers in the 10-fold cross-validation, whereas only 10% was used to test them.

B. Synthetic Minority Oversampling Approach (SMOTE)

The proposed disease prediction model, which was called the "Synthetic Minority Oversampling Approach (SMOTE)," was used to classify hypertension by applying it to a dataset. This technique is the most specific and powerful method used for sampling which follows the algorithm technique to calculate the distance of the space features in minority examples. This helps to develop the synthetic data within the premises of the minority example and required help from the neighbouring data. This technique is also known as borderline SMOTE [22]. The idea is generated from this method to produce the synthetic sample from near boundaries. These algorithms are more effective towards the binary classes having more than two features. This method is predicted by the over-sampling method produced by Chawla et al. [23]. This method has been used to increase the number of the sampling by interpolating the clustered samples which are in minority. The selection of accurate parameters has been mentioned to function correctly for the interpretation of the SMOTE algorithms. This model has three parameters which are: (k) for the neighbour which is very close, (perc. over) used to determine the extra cases of the minority classes, and accurate selection of the parameters. In this model, the method of optimization is one of the best findings for solving the evolutionary biological process [24]. The flow diagram of the over-sampling method (Fig. 1) and the improved hybrid model (Fig. 2) is presented.

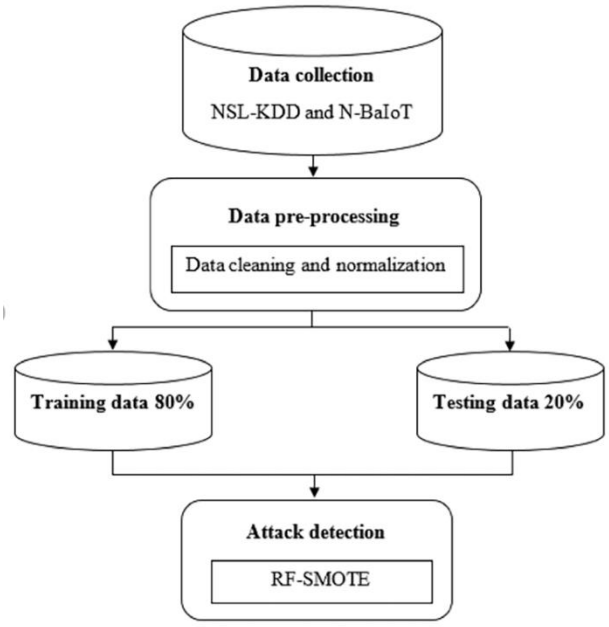

Fig. 1. Workflow radiofrequency SMOTE model. [25].

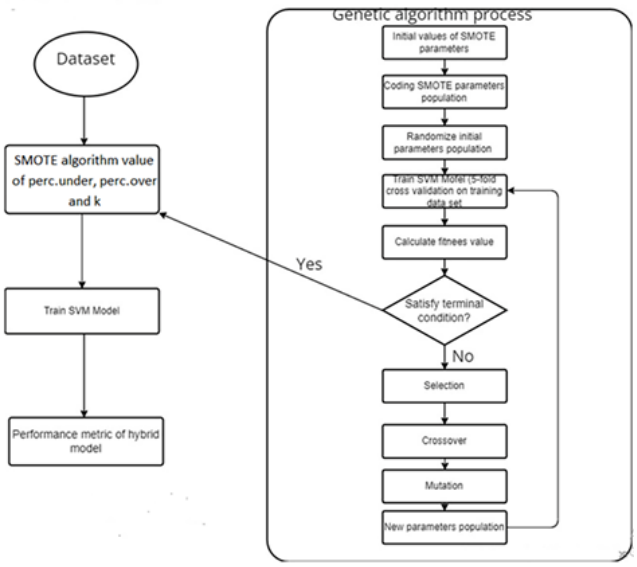

Fig. 2. Improved SMOTE model. [24].

The model's features are chosen based on how easy they are to use to collect data, and they must be statistically significant for univariate logistic regression ($p < 0.05$). The inputs were the number of nearest neighbours, the number of SMOTEs, and the number of minority class samples (MCS). The outputs were synthetic MCS to solve the problem of imbalance in classification [26].

IV. RESULTS

The suggested model uses the Synthetic Minority Oversampling Approach (SMOTE), which is based on the oversampling technique and generates synthetic samples for the minority class. This method assists in addressing the overfitting problem brought on by random oversampling. By overlaying several good examples, it focuses on the dataset to generate new cases [27, 28]. As a result, while evaluating the suggested model, the AUC value is taken into account. Table I displays the recommendations for evaluating any classifier using AUC [29].

In this experiment, the dataset was run on ten machine learning classifiers using 10-fold cross-validation. 90% of the data was utilized to train the classifiers in the 10-fold cross-validation, whereas only 10% was used to test them. The results of ten classifiers' 10-fold cross-validation are shown in Tables II and III.

TABLE I. AUC VALUES DESCRIPTION [3]

AUC Range	Description
AUC = 0.50	Bad classification (no discrimination)
0.50 < AUC < 0.70	Poor classification (poor discrimination)
0.70 ≤ AUC < 0.80	Acceptable classification (acceptable discrimination)
0.80 ≤ AUC < 0.90	Excellent classification (excellent discrimination)
AUC ≥ 0.90	Outstanding classification (outstanding discrimination)

TABLE II. 10-FOLD CROSS-VALIDATION CLASSIFICATION PERFORMANCE EVALUATION OF DIFFERENT CLASSIFIERS ON THE HYPERTENSION DATASET ON FULL FEATURES WITHOUT OVERSAMPLING

Predictive model	Classifiers' performance evaluation metrics			
	Accuracy (%)	*Precision (%)*	*Recall (%)*	*AUC (%)*
Logistic regression	76.5	74.4	76.5	79
K-nearest neighbour k=3	78	76.7	78	81.5
PART (rule-induction algorithms)	77.5	78	77.5	71.2
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)	77.5	77.3	77.5	65
NaiveBayes	79	77.8	79	81.7
Decision Tree (J48)	75.5	74.1	75.5	74
REPtree	76	75.1	76	72.2
Support Vector Machine (SVM)	76.5	74.1	76.5	62.4
Random forest	73	70.8	73	79
Artificial Neural Networks (ANN)	74	74.4	74	75.1

Therefore, Table II presents ten classifiers for the hypertension dataset with complete features and no oversampling. Accuracy, precision, recall, and AUC were some of the assessment performance metrics utilized to rate the suggested hypertension prediction model. It became clear that most classifiers can discover trustworthy results using these metrics. The greatest AUC value, equivalent to 81.7, was obtained experimentally for hypertension prediction (high blood pressure detection) using the Naïve Bayes classifier without oversampling.

The minority class of the family history of high blood pressure dataset is oversampled using the SMOTE method. Investigated is the class imbalance caused by the oversampling of the family history of high blood pressure dataset.

The hypertension dataset's entire characteristics were used in the tests, which made use of all classifiers, as shown in Table III. The findings show that all classifiers that used accuracy, precision, recall, and AUC values generated positive outcomes. Based on oversampling AUC values of 89.6 for the k-nearest neighbor, outstanding results were obtained. So, without oversampling, the suggested SMOTE-k-nearest neighbor prediction model outperformed the classifiers. As a result, the SMOTE-k-nearest neighbor prediction model performs exceptionally well, as evidenced by its excellent performance in Table III (83.9% classification accuracy, 85.1% specificity, 83.9% sensitivity, and 89.6% AUC) when compared to other classifiers using 10-fold cross-validation on the hypertension dataset with full features and no oversampling.

The AUC curve, which is a visual depiction of a classification model's true-positive and false-positive rates, is shown in Fig. 3. AUC around one indicates a superior classification with a good class separability metric. AUC which is around 0 indicates a poor model with no distinguishing class. AUC of 0.5 means that the model is unable to distinguish between classes. As shown in Table III, the majority of classifiers provided good classifications, with the k-nearest neighbor getting the greatest AUC rate, which is equivalent to 89.6 and denoting excellent classification, as shown.

TABLE III. 10-FOLD CROSS-VALIDATION CLASSIFICATION PERFORMANCE EVALUATION USING TEN CLASSIFIERS ON HYPERTENSION DATASET ON FULL FEATURES USING OVERSAMPLING

Predictive model	Classifiers' performance evaluation metrics			
	Accuracy (%)	*Precision (%)*	*Recall (%)*	*AUC (%)*
Logistic regression	77.9	78.0	77.9	84.1
K-nearest neighbor k=4	83.9	85.1	83.9	89.6
PART (rule-induction algorithms)	80.72	80.9	80.7	79.8
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)	80.7	81.8	80.7	80.5
NaiveBayes	78.7	78.5	78.7	85.6
Decision Tree (J48)	82.7	83	82.7	79.6
REPtree	80.3	80.2	80.3	80.2
Support Vector Machine (SVM)	77.51	77.6	77.5	76.6
Random Forest	81.5	81.9	81.5	86.5
Artificial Neural Networks (ANN)	78.8	79.5	78.7	83.4

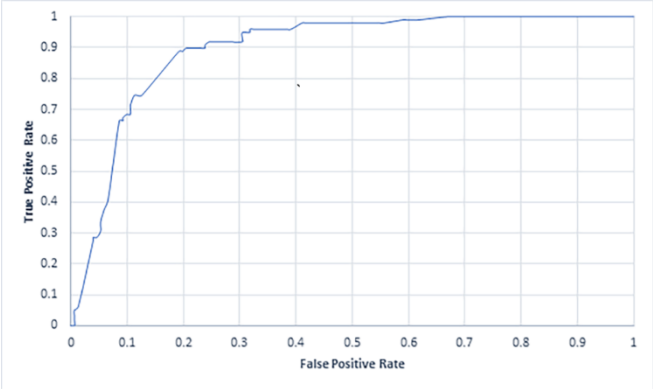

Fig. 3. AUC curves for the family history of high blood pressure dataset.

V. DISCUSSION AND CONCLUSION

The data which has been extracted from Petra University Health Center is considered to be very helpful for ML. It has been availed to produce a decision tree to identify the data related to hypertension. There are many implications to modifying the solutions which can perform preventive measures. These solutions have been identified by using projecting models. According to the previous studies the SMOTE model is proven to be a successful model with the combination of others like BrSmote model for the analysis of the classification of the accuracy into five multi-label classifiers and the ten multi-label classifiers. The average and loss ranking is more sensitive and accurate in SMOTE model as compared to BrSmote model. The diseases like hypertension and other syndromes can be classified to improve the diagnosis of the patients [19]. These models can be very predictive for

those people who can develop hypertension at very high risk. [16]These predictive models provide better risk communication. They can guide people who are concerned about their health decision and diagnosis of disease. This awareness in the community develops positive impact.[25] Furthermore, the lowest classification of the number is based on the class interest and the parameter which are required for the diagnosis. In the real world, the classification varies with the nature of the risk management, detection of fraud and the diagnosis of the disease in the medical history [30]. This positive mediation helps to decide the significant level of understanding. In the future, this research can improve the reliability of the predictive models. It can minimize the count of unhealthy effects on a large population by using accurate algorithms. There are many assessing tools for predicting different algorithms. These predictors are boosting gradient machines, supporting vector machines, classifiers of naive Bayes, and neural artificial networks.

ACKNOWLEDGEMENT

The author is thankful to all the associated personnel who contributed to this study by any means.

REFERENCES

[1] B. Zhou, P. Perel, G.A. Mensah and M. Ezzati, "Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension," Nat. Rev. Cardiol. vol. 18, pp. 785-802, November 2021. https://doi.org/10.1038/s41569-021-00559-8.

[2] W.H.O, "Global status report on noncommunicable diseases 2010," World Health Organization, 2011.

[3] S.M.S. Islam, T.D. Purnat, N.T.A. Phuong, U. Mwingira, K. Schacht and G. Fröschl, "Non-Communicable Diseases (NCDs) in developing countries: a symposium report," Glob. Health. vol. 10, pp. 1-8, July 2014. https://doi.org/10.1186/s12992-014-0081-9.

[4] K.T. Mills, A.Stefanescu and J. He, "The global epidemiology of hypertension," Nat. Rev. Nephrol. vol. 16, pp. 223-237, September 2020. https://doi.org/10.1038/s41581-019-0244-2.

[5] S. Basu and C. Millett, "Social epidemiology of hypertension in middle-income countries: determinants of prevalence, diagnosis, treatment, and control in the WHO SAGE study," Hypertension, Standford CA, vol. 62, pp. 18-26, November 2013. https://doi.org/10.1161/hypertensionaha .113.01374.

[6] Krishnan and R. Garg, "Hypertension in the South-East Asia region: an overview," In Reg. Health Forum. India, vol. 17, pp. 7-14, September 2013.

[7] R.B. D'Agostino Sr, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro and W.B. Kannel, "General cardiovascular risk profile for use in primary care: the Framingham Heart Study," Circulation, vol. 117, pp. 743-53, April 2008. https://doi.org/10.1161/ CIRCULATIONAHA.107.699579.

[8] D. C. Goff, Jr, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Jr, Smith, P. Sorlie, N. J. Stone, P. W. Wilson, H. S. Jordan, L. Nevo and J. Wnek, "2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," Circulation, vol. 129, pp. S49-73, November 2014. https://doi.org/10.1161/01.cir. 0000437741.48606.98.

[9] J.B. Echouffo-Tcheugui, G.D. Batty, M. Kivimäki and A.P. Kengne, "Risk models to predict hypertension: a systematic review," PloS one. California, vol. 8, pp. e67370, July 2013. https://doi.org/10.1371/journal.pone.0067370.

[10] Q. Bi, K.E. Goodman, J. Kaminsky and J. Lessler, "What is machine learning? A primer for the epidemiologist," Am. J. Epidemiol. Baltimore, vol. 188, pp. 2222-2239, December 2019. https://doi.org/10.1093/aje/kwz189.

[11] J.J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado and M.F. Landecho, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)," J. Biomed. Inform. Spain, vol. 97, pp. 103257. July 2019. https://doi.org/10.1016/j.jbi.2019.103257.

[12] E.A. Martinez-Ríos, M.R. Bustamante-Bello and L.A. Arce-Sáenz, "A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data," Biomed. Signal Process Control. Mexico, vol. 68, pp. 102813, July 2021https://doi.org/10.1016/j.bspc.2021.102813.

[13] M. Nour and K. Polat, "Automatic classification of hypertension types based on personal features by machine learning algorithms," Math. Probl. Eng. Saudia Arabia, pp. 1-13, January 2020. https://doi.org/10.1155/2020/2742781.

[14] J.J. Boutilier, T.C. Chan, M. Ranjan and S. Deo, "Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis," JMIR. Hyderabad, India, vol. 23, pp. e20123, January 2021. https://doi.org/10.2196/20123

[15] X. Diao, Y. Huo, Z. Yan, H. Wang, J. Yuan, Y. Wang, J. Cai and W. Zhao, "An application of machine learning to the etiological diagnosis of secondary hypertension: retrospective study using electronic medical records," JMIR. Med. Inform. Beijing, China, vol. 9, pp. e19739, January 2021. https://doi.org/10.2196/19739.

[16] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences. 2019 Oct 29;116(44):22071-80. https://doi.org/10.1073/pnas.1900654116.

[17] AM. Alaa, T. Bolton, E. Di Angelantonio, JH. Rudd, M. Van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," PloS one. vol. 14, pp. e0213653.May 2019. https://doi.org/10.1371/journal.pone.0213653.

[18] Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. De Cata, L. Chiovato, R. Bellazzi, "Machine learning methods to predict diabetes complications," J. Diab. Sci. Tech. vol. 12. pp. 295-302., March 2019 https://doi.org/10.1177/1932296817706375.

[19] GZ. Li, Z. He, FF. Shao, AH. Ou, XZ. Lin, "Patient classification of hypertension in Traditional Chinese Medicine using multi-label learning techniques," BMC Med. Gen. vol. 8. pp.1-6, December 2015. https://doi.org/10.1186/1755-8794-8-S3-S4.

[20] GP. Liu, GZ. Li, YL. Wang, YQ. Wang, "Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning," BMC Med. vol. 10. pp. 1-2. December 2010 https://doi.org/10.1186/1472-6882-10-37.

[21] YM. Huang, CM. Hung, HC. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem," Nonlinear Analy. Wor. App. vol. 7. pp.720-47, September 2006. https://doi.org/10.1016/j.nonrwa.2005.04.006.

[22] H. Han, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1 2005 (pp. 878-887). Springer Berlin Heidelberg.).

[23] NV. Chawla, KW. Bowyer, LO. Hall, WP. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research," vol.1.pp. 321-57, June 2002.

[24] P. Akın, "A new hybrid approach based on genetic algorithm and support vector machine methods for hyperparameter optimization in synthetic minority over-sampling technique (SMOTE)," AIMS Mathematics. vol. 8. pp.9400-9415, 2023.

[25] MG. Karthik, MM. Krishnan, "Hybrid random forest and synthetic minority over sampling technique for detecting internet of things attacks." J.A.I.H.C. pp.1-1, March 2021.

[26] Azad, B. Bhushan, R. Sharma, A. Shankar, K.K. Singh and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus,"

Multimed. Syst. Switzerland, vol. 28, pp. 1289-1307, June 2022. https://doi.org/10.1007/s00530-021-00817-2.

[27] J. Zhai, M. Wang and S. Zhang, "Binary imbalanced big data classification based on fuzzy data reduction and classifier fusion," Soft. Comput. Switzerland, vol. 26, pp. 2781-2792, March 2022. https://doi.org/10.1007/s00500-021-06654-9.

[28] Data, Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu and Y. Zhou, "A novel ensemble method for classifying imbalanced data," Pattern. Recognit. China, vol. 48, pp. 1623-1637, November 2015. https://doi.org/10.1016/j.patcog.2014.11.014.

[29] Jr, D.W. Hosmer, S. Lemeshow and R.X. Sturdivant, "Applied logistic regression." John Wiley & Sons, vol. 398, 2013.

[30] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks. 2008 Mar 1;21(2-3):427-36.