

Early disease prediction algorithm for hypertension-based diseases using data aware algorithms

Yasmeen Shaikh¹, Vasudev Parvati², Sangappa Ramachandra Biradar²

¹Department of Computer Science and Engineering, KLS Vishwanathrao Deshpande Institute of Technology, Haliyal, India

²Department of Information Science and Engineering, Shri Dharmasthala Manjunatheshwara College of Engineering and Technology, Dharwad, India

Article Info

Article history:

Received Jul 24, 2021

Revised May 24, 2022

Accepted Jun 11, 2022

Keywords:

Early disease detection

Ensemble learning

Extended isolation forest

Hypertension

Voting ensemble learning

ABSTRACT

This paper implements a data aware early prediction of hypertension-based diseases. Automated data preprocessing method that adopts for both balanced and unbalanced data is the data aware method included in the disease classification algorithm. Proposed data aware data preprocessing method is evaluated on the ensemble learning based classification algorithm for early disease prediction. Data aware preprocessing method adopts isolation forest algorithm for outlier detection as part of the automation. Automated sampling method of applying the sampling corresponding to either balanced or unbalanced data is adopted. Performance evaluation of the proposed data aware algorithm using isolation forest algorithm for anomaly detection is experimented. Python based implementation of the proposed data aware classification algorithm inferred a better area under the curve (AUC) receiver operating characteristics (ROC) curve for isolation forest implementation in data preprocessing automation thus developed. While the individual classifiers multilayer perceptron classifier approached till 0.918 (AUC) in the ROC-AUC curve. The ensemble learning algorithm that included multilayer perceptron classifier, logistic regression classifier, support vector classifier and decision tree algorithm with the isolation forest-based anomaly detection algorithm performed better than the individual machine learning algorithm with 0.922 (AUC) in the ROC-AUC curve.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Yasmeen Shaikh

Department of Computer Science and Engineering

KLS Vishwanathrao Deshpande Institute of Technology

Haliyal, Karnataka, India

Email: y.s.shaikh1210@gmail.com

1. INTRODUCTION

Diabetes is a disease that leads to multiple other serious diseases including cardiovascular diseases, thyroid problems, bone related issues, heart attack due to cardiac arrest. Diabetes and blood pressure are related symptoms for numerous cardiovascular diseases. Thus, early detection of diabetes helps the patients to avoid getting into serious problems which are fatal in nature. According to a survey conducted by the International Diabetes Federation (IDF), in 2020, 77 million people from India are affected by diabetes from 88 million people affected in Southeast Asia [1]. Overall, 463 million people have diabetes in the world [1]. Type 1 diabetes is prevalent in children with India numbering two after the USA and highest in the Southeast region [2]. World Health Organization (WHO) records that among the total deaths 2% of deaths are due to diabetes in India [3]. Diabetes and hypertension is found to be the common non-communicable disease [4], [5]. Diabetes counts to 46.2% and hypertension counts to 4% of total deaths [6]. Metabolic

disorder that disturbs the production of the insulin due to blood glucose levels is called Type 2 diabetes [7], [8]. Stroke and mortality is evident in the people as risks increases with diabetes inherent in more populations in India [9]. By 2030 it is expected to reach 228 million in developing countries [10], [11]. 1 billion adults are estimated to suffer with hypertension in developing countries by 2025 [12] and one in every three people would die due to hypertension [13].

Machine learning algorithms are applied on prediction of hypertension and diabetes [14], [15]. Smart sensors and cloud computing based continuous monitoring of vital health signs like electrocardiogram (ECG), premature atrial contraction (PAC), alcohol consumption, smoking habits, caffeine intake help early detection of the cardiovascular diseases [16]. Heart rate variability (HRV) at different time zones of the day is observed to predict the hypertensive patients with higher risks. Decision tree algorithm is applied with random under-sampling boosting (RUSBOOST) to train the HRV and demographic features to predict high risk patients [17]. Type 2 diabetes and hypertension prediction using the ensemble learning algorithm is developed as a mobile application on four different datasets. Risk factor data is observed and intimated to the remote server [18]. Occupational related factors are included in the risk assessment of hypertension in the steel industry employees in China. Risk of hypertension is evaluated using learning vector quantization (LVQ) neural network algorithm and fisher-support vector machine (SVM) algorithm. Variation of accuracy with the input data sampling is evaluated to observe the 'tailing' process in both the machine learning algorithm [19]. A tree-based approach applies different machine learning algorithm to different subsets of feature space. Each subset of the feature set is associated to the node in the decision tree developed [20].

Traditional learning methods that build the training model from the filtered labels and the noisy inferring labels is a straightforward learning method. Unlike the traditional method, the two-stage approach developed involves filtering true labels and building a training algorithm. Inference of the filtered labels reduces the accuracy of prediction since useful information may get lost. Bootstrapping method creates subsets from the total dataset and assigned with the class memberships of the multiple noisy labels. Base classifier trains the extended sub-dataset. Other unlabeled instances are predicted using the aggregation principle from the outputs of other M base classifiers. Advanced ensemble learning algorithm is found to be effective compared to the traditional methods [21]. Multi-tier weighted ensemble learning (MTWEL) is developed that optimizes the parameters of all the learning algorithms used as the ensemble using genetic algorithm (GA). Heart disease is predicted using the algorithm and good performance is observed [22]. Feature importance selection is applied on the features selected for prediction. A combination of k-nearest neighbor (KNN) and logistic regression is used for the ensemble learning framework that predict cardiovascular diseases. Data imbalance is managed via synthetic minority over-sampling technique (SMOTE) method [23]. Independent component analysis (ICA) is used for the dimensionality reduction to implement the lung cancer detection algorithm using the AdaBoost based ensemble learning method [24].

A meta-learning method combines multiple learning algorithms for a time series prediction algorithm using a combination policy of the ensembles. An actor-critic model is developed to optimize the weights in the deep reinforcement learning based ensemble learning framework [25]. Gaussian mixture model (GMM) based data preprocessing method is used for a power grid environment algorithm [26]. Data preprocessing to handle imbalance data for the industrial scenario using a time series prediction is dealt in [27]. The streamed data are immediately given to Z-score normalization method to homogenize the data range. Time series data is applied with sliding window method to sort before applying the machine learning algorithm. Sliding window is used to sort the time series data. A hybrid ensemble learning algorithm is introduced for the wind forecasting algorithm. Back propagation, least square support vector machine (LSSVM), adaptive neuro-fuzzy inference system (ANFIS) and Elman neural network (ENN) is applied as the ensemble learning programs and CLSJaya is used to optimize the weight values [28]. Data preprocessing automation at the data center is carried out. Missing value imputation, forecasting replacement value for missing values are carried out on the data for the data preprocessing algorithm [29]. Automation of data preprocessing, feature extraction and hyperparameter tuning is developed in [30].

Previous heart disease prediction algorithms need the data preprocessing automation to be implemented to enhance the prediction algorithm capability. An approach that carries out the data preprocessing algorithm by implementing the data aware approach is not carried out in the previous literature. Data preprocessing algorithms is implemented manually in the previous literatures and even the automated data preprocessing algorithms implemented are not a data aware technique in the previous literatures.

This paper discusses the disease prediction algorithm with the data aware data preprocessing approach. Automated data preprocessing approach includes the advanced feature importance detection using extended isolation forest (EIF) for anomaly detection. An algorithm that automatically detects whether the data is balanced or imbalanced and applies the data preprocessing according to the data available is developed in this paper. Extended isolation forest-based implementation is compared with the isolation forest-based ensemble learning algorithm and performance is compared. Section 2 discusses the proposed

data aware ensemble learning methodology with detailed flowchart and details; Section 3 discusses the database utilized for the early detection of diabetes; Section 4 discusses the result and discussion for the proposed algorithm and comparison results.

2. PROPOSED DATA AWARE ENSEMBLE LEARNING METHODOLOGY

Generalization in independent machine learning methods gets a hit due to different capability issues inherent to each method. The algorithm that could combine the advantages of different machine learning algorithms solves the generalization issue to a greater extent. Isolation forest algorithm is a good anomaly detection algorithm. Better anomaly detection implementation using the extended isolation forest algorithm is incorporated. The overall implementation details of the proposed implementation is as given in Figure 1. Disease prediction implementation is divided into four major parts: i) data preprocessing automation, ii) outlier detection and feature engineering, iii) training and testing, and iv) model evaluation. Complete data preprocessing process shown in Figure 1 is automated providing a human interference less machine learning paradigm. Ensemble learning algorithm increases the generality of the learning algorithm.

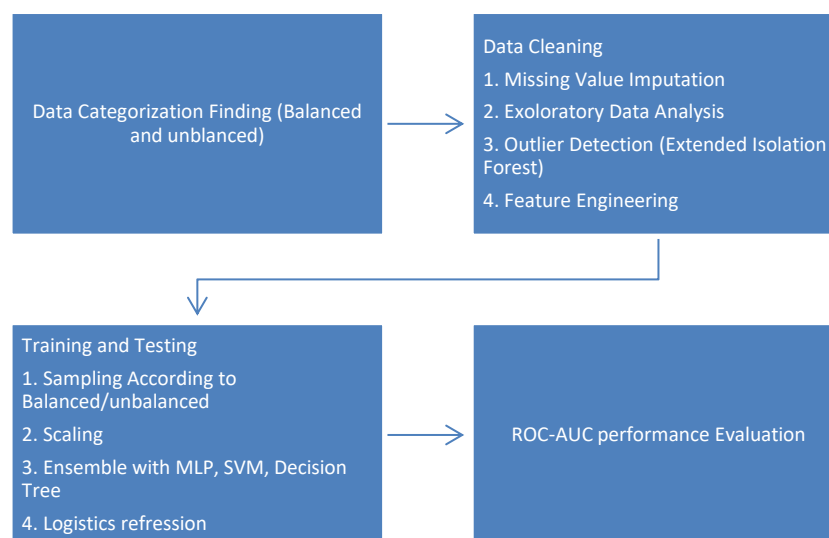


Figure 1. Overall block diagram-data preprocessing automated ensemble learning

2.1. Data preprocessing automation

Automation of data preprocessing involves categorization of data into balanced and imbalanced data, data cleaning and anomaly detection. This decides the sampling method of the proposed implementation. Missing value imputation by replacing the mean value, cleaning impurities are automated after the data is categorized.

2.2. Outlier detection and feature engineering

Disease prediction algorithms being the most crucial medical applications, the foolproof nature of the early prediction algorithm is a primary indicator for the research thus carried out. A method that can handle a large amount of data (in millions) needs a method with higher generality and highly orthogonal input data. Possibility of higher correlation among the sample data insists on better feature engineering techniques for a better prediction performance. The data aware preprocessing algorithm is a challenge that needs to be obtained as the proposed objective. The block diagram insists on an extended isolation algorithm to be implemented on the medical diagnosis problem thus chosen. The outlier detection exhibits better orthogonality while the extended isolation forest is applied on the medical diagnosis implementation.

2.3. Training and testing

Complete dataset of training inputs and target pair is split into training and testing sets with 80% and 20% ratio respectively. The results obtained from different individual machine learning methods and the proposed ensemble learning method are compared for performance evaluation. Ensemble learning method

involves the multilayer perceptron (MLP), support-vector machine (SVM) and decision tree (DT) algorithm to obtain the first level of output and this output is given to the logistic regression to obtain the final classification output.

3. DATASET AND DATA PROCESSING USED FOR PROPOSED ENSEMBLE LEARNING

The UCI database [31] of chronic kidney disease is incorporated for the proposed implementation. Since diabetes is the early symptom of both of the kidney diseases it is obtained from the repository. The dataset has input and output as given in Table 1.

Table 1. Input and output data from UCI dataset [31]

Variable	Input/Target	Type	Unit of the variable
Blood Pressure	Input	Numerical	bp in mm/Hg
Age	Input	Numerical	in years
Red Blood Cells	Input	Nominal	Normal/abnormal
Specific Gravity	Input	Nominal	(1.005,1.010,1.015,1.020,1.025)
Pus Cell	Input	Nominal	(Present, Not Present)
Bacteria	Input	Nominal	(Present, Not Present)
Albumin	Input	Nominal	(0,1,2,3,4,5)
Hemoglobin	Input	Numerical	hemo in gms
Packed Cell Volume	Input	Numerical	
Blood Glucose Random	Input	Numerical	bgr in mgs/dl
Hypertension	Input	Nominal	Yes/no
White Blood Cell Count	Input	Numerical	wc in cells/cumm
Blood Urea	Input	Numerical	bu in mgs/dl
Appetite	Input	Nominal	(good,poor)
Serum Creatinine	Input	Numerical	sc in mgs/dl
Diabetes Mellitus	Input	Nominal	(yes,no)
Sodium	Input	Numerical	sod in mEq/L
Coronary Artery Disease	Input	Nominal	(yes,no)
Potassium	Input	Numerical	pot in mEq/L
Anemia	Input	Nominal	(yes,no)
Pedal Edema	Input	Nominal	(yes,no)
Red Blood Cell Count	Input	Numerical	rc in millions/cmm
Pus Cell clumps	Input	Nominal	present, notpresent
Class			ckd, notckd

4. RESULTS AND DISCUSSION

The disease prediction algorithm that is carried out for the cardiac disease is the ensemble learning algorithm with the extended isolation forest algorithm. The dataset for the proposed implementation is obtained from the UCI dataset [31]. For the given dataset effective data preprocessing automation is obtained along with the accurate early prediction than the individual learning algorithms. Extended isolation forest algorithm with the Ensemble learning algorithm is found to be classifying more accurately with the data aware preprocessing algorithm.

4.1. Preprocessing results

Toolbox named Pandas is used to understand the data profile and statistical information about the dataframe. Continuous variables are estimated for skewness, min, max, standard deviation, percentile is given by profile report that defines the statistical understanding of the data. Correlation of each independent variable with target variable is found. Reduced dataset that is chosen for training the different algorithms for the performance evaluation after preprocessing of the input data are as given in the Table 2.

Table 2. Input data for training

Parameter	Values
Number of variables	10
Number of observations	224
Total Missing (%)	0.0%
Total size in memory	17.6 KiB
Average record size in memory	80.6 B

The variable types in the data given include the numerical and the categorical data. Table 3 depicts the number of different variables. After the variables are converted to numerical variable and target is

mapped with the input using the ensemble learning method. Data visualization is carried out as the first step by the data aware algorithm. Outlier based analysis is carried out to provide the idea about the data. Box plot of each data from the dataset is as shown in the Figure 2. Box plot visualizes each variable in the data set. It depicts the range of values for each data in the data set.

Table 3. Variable types

Variable Type	Number of Variables
Numeric	5
Categorical	5
Boolean	0
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

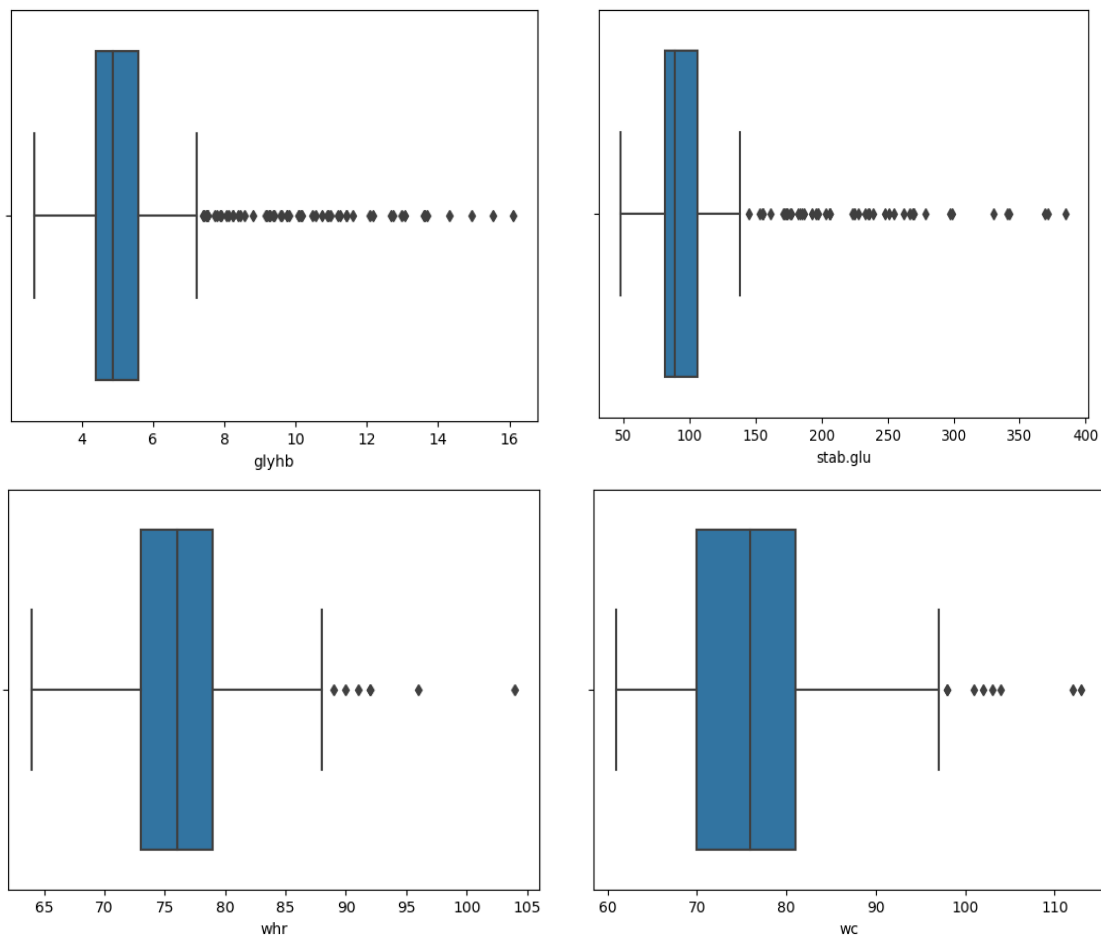


Figure 2. Outlier detection for different input variables

4.2. Outlier results

The outlier detection for diabetes dataset using isolation forest is shown in Figure 3. The normal and the predicted outliers from the extended isolation forest is as given in the Figure 3(a). The outliers found from this isolation forest algorithm will be removed. The 3D graph of the outliers in the variable space is as shown in Figure 3(b). It can be clearly seen that the data can be isolated as outlier and inliers as shown in Figure 3(b). The training can improve the classification performance as shown in the results obtained. Outliers detected from the isolation forest algorithm are conveniently removed to obtain better accuracy in training procedure. Isolation forests for both the male and the female datasets are obtained for visualization.

The outlier detection for female dataset using Isolation Forest is shown in Figure 4. Figure 4(a) shows the normal and the predicted outliers for the female data from the dataset. The 3D graph of the outliers in female in the variable space is as shown in Figure 4(b). The outlier detection for male dataset using Isolation Forest is shown in Figure 5. Figure 5(a) shows the normal and the predicted outliers for the male data from the dataset. The 3D graph of the outliers in male in the variable space is as shown in Figure 5(b). It can be clearly seen that the data can be isolated as outlier and inliers as shown in Figure 5(b). The training can improve the classification performance as shown in the results obtained. Anomaly detection from the isolation forest algorithm has clearly defined the inliers and outliers in the data and outliers are removed.

4.3. AUC-ROC results

In order to compare the performance evaluation of the proposed algorithm, logistics, MLP classifier, decision tree classifier, switched virtual circuit (SVC) are 4 classifiers used and ensembled using stacking CV classifier. Imbalanced data is handled by using SMOTE which is decided due to the data categorization. Data is a balanced data. Classification algorithm relies on the area under the curve and receiver operating characteristic curve (AUC-ROC) curve for its performance evaluation. Degree of separability between the classes are defined by AUC values and ROC is the probability curve. AUC indicates how better it distinguishes between different classes.

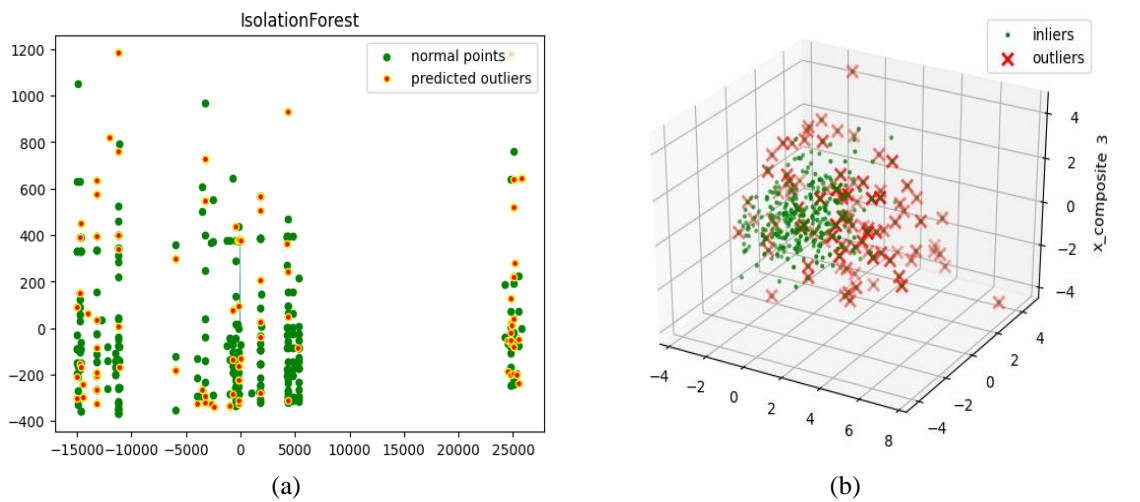


Figure 3. Outlier from isolation forest (a) isolation forest outliers 2D and (b) isolation forest outliers 3D

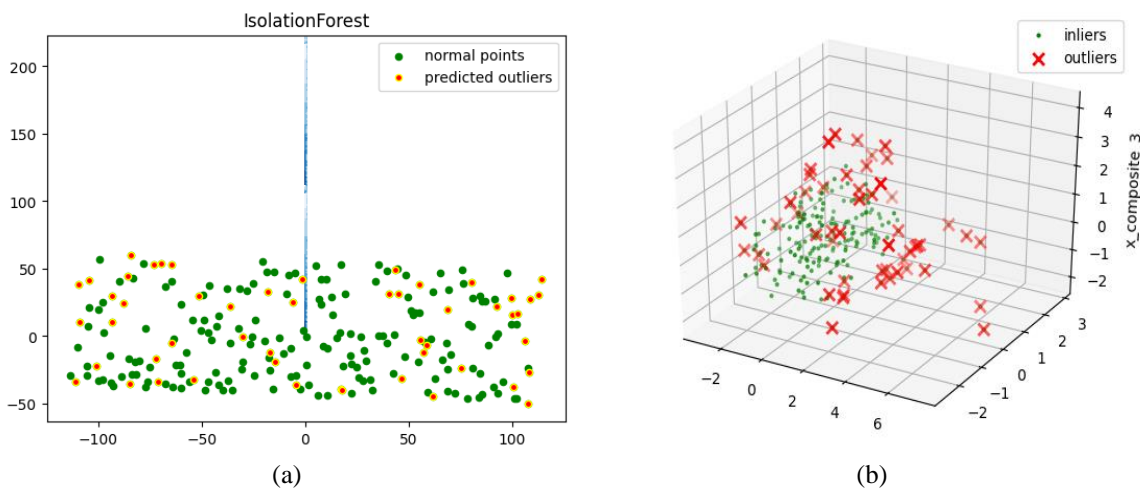


Figure 4. Outlier detected on female data (a). isolation forest outliers 2D and (b) isolation forest outliers 3D

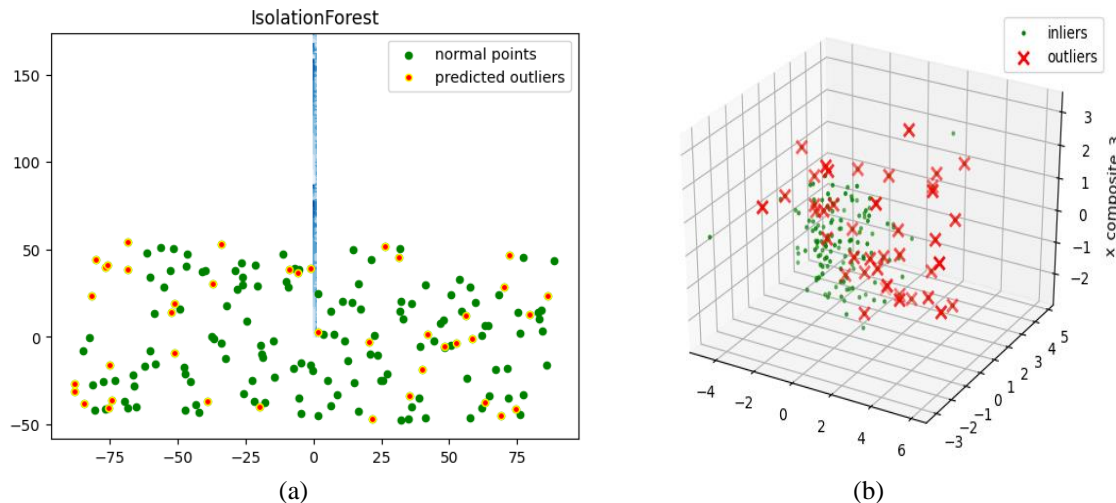


Figure 5. Outlier detected on male data (a) isolation forest outliers 2D and (b) isolation forest outliers 3D

Figure 6 shows the AUC ROC curves for linear regression (LR), MLP, decision tree, SVC and stack (Ensemble) method. It is observed from the Figure 6 that the AUC is as high as 0.922 compared to other machine learning algorithms. In the individual machine learning algorithm MLP scored better AUC with 0.918.

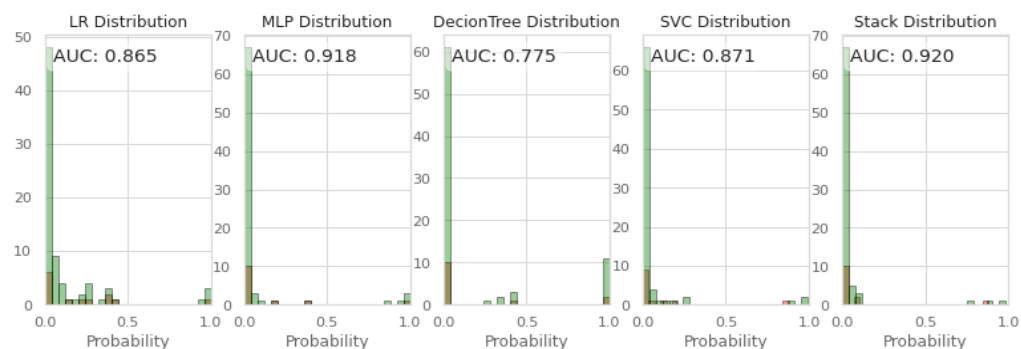


Figure 6. AUC ROC curves for LR, MLP, decision tree, SVC and stack (ensemble) method

5. CONCLUSION




The results obtained from the different outlier detection methods and the sampling methods are discussed and compared and performance evaluations of the proposed methods are discussed. A suggestion mechanism for the best ensemble classification for the medical diagnosis is implemented. Multilayer perceptron classifier approached till .918(AUC) in the ROC-AUC curve. The ensemble learning algorithm that included multilayer perceptron classifier, logistic regression classifier, support vector classifier and decision tree algorithm with the isolation forest-based anomaly detection algorithm performed better than the individual machine learning algorithm with .922 (AUC) in the ROC-AUC curve. For the given dataset effective data preprocessing automation is obtained using ensemble approach along with the accurate early prediction than the individual learning algorithms.

REFERENCES




- [1] S. Islam, N. Jahan, and M. E. Khatun, "Cardiovascular disease forecast using machine learning paradigms," in *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, Mar. 2020, pp. 487–490, doi: 10.1109/ICCMC48092.2020.ICCMC-00091.
- [2] A. Ramachandran, C. Snehalatha, and R. C. W. Ma, "Diabetes in South-East Asia: an update," *Diabetes Research and Clinical Practice*, vol. 103, no. 2, pp. 231–237, Feb. 2014, doi: 10.1016/j.diabres.2013.11.011.

- [3] H. R. H. Al-Absi, M. A. Refaee, A. U. Rehman, M. T. Islam, S. B. Belhaouari, and T. Alam, "Risk factors and comorbidities associated to cardiovascular disease in Qatar: a machine learning based case-control study," *IEEE Access*, vol. 9, pp. 29929–29941, 2021, doi: 10.1109/ACCESS.2021.3059469.
- [4] M. S. Capehorn, D. W. Haslam, and R. Welbourn, "Obesity treatment in the UK health system," *Current obesity reports*, vol. 5, no. 3, pp. 320–326, Sep. 2016, doi: 10.1007/s13679-016-0221-z.
- [5] S. Hariharan, R. Umadevi, T. Stephen, and S. Pradeep, "Burden of diabetes and hypertension among people attending health camps in an urban area of Kancheepuram district," *International Journal Of Community Medicine And Public Health*, vol. 5, no. 1, p. 140, Dec. 2017, doi: 10.18203/2394-6040.ijcmph20175771.
- [6] J. R. Petrie, T. J. Guzik, and R. M. Touyz, "Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms," *Canadian Journal of Cardiology*, vol. 34, no. 5, pp. 575–584, May 2018, doi: 10.1016/j.cjca.2017.12.005.
- [7] K. G. Alberti and P. F. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation.," *Diabetic Medicine*, vol. 15, no. 7, pp. 539–553, 1998.
- [8] American Diabetes Association, "Standards of medical care in diabetes–2006," *Diabetes Care*, vol. 29, no. suppl_1, pp. s4–s42, Jan. 2006, doi: 10.2337/diacare.29.s1.06.s4.
- [9] N. N. Tun, G. Arunagirinathan, S. K. Munshi, and J. M. Pappachan, "Diabetes mellitus and stroke: A clinical update," *World Journal of Diabetes*, vol. 8, no. 6, p. 235, 2017, doi: 10.4239/wjd.v8.i6.235.
- [10] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: estimates for the year 2000 and projections for 2030," *Diabetes Care*, vol. 27, no. 5, pp. 1047–1053, May 2004, doi: 10.2337/diacare.27.5.1047.
- [11] F. Rubino, "Is type 2 diabetes an operable intestinal disease? A provocative yet reasonable hypothesis," *Diabetes care*, vol. 31 Suppl 2, no. Supplement_2, pp. S290–S296, Feb. 2008, doi: 10.2337/dc08-s271.
- [12] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, P. K. Whelton, and J. He, "Global burden of hypertension: analysis of worldwide data," *The Lancet*, vol. 365, no. 9455, pp. 217–223, Jan. 2005, doi: 10.1016/s0140-6736(05)17741-1.
- [13] F. Ullah, A. H. Abdullah, O. Kaiwartya, and S. Prakash, "Patient data dissemination in wireless body area network: A qualitative analysis," in *ACM International Conference Proceeding Series*, 2016, vol. 04-05-March-2016, pp. 1–6, doi: 10.1145/2905055.2905278.
- [14] X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93–99, Feb. 2013, doi: 10.1016/j.kjms.2012.08.016.
- [15] H. F. Golino *et al.*, "Predicting increased blood pressure using machine learning," *Journal of Obesity*, vol. 2014, pp. 1–12, 2014, doi: 10.1155/2014/637635.
- [16] A. Forkan, I. Khalil, and Z. Tari, "Context-aware cardiac monitoring for early detection of heart diseases," *Computing in Cardiology*, vol. 40, pp. 277–280, 2013.
- [17] M. Alkhodari, D. K. Islayem, F. A. Alskafi, and A. H. Khandoker, "Predicting hypertensive patients with higher risk of developing vascular events using heart rate variability and machine learning," *IEEE Access*, vol. 8, pp. 192727–192739, 2020, doi: 10.1109/ACCESS.2020.3033004.
- [18] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," *IEEE Access*, vol. 7, pp. 144777–144789, 2019, doi: 10.1109/ACCESS.2019.2945129.
- [19] J. H. Wu *et al.*, "Risk assessment of hypertension in steel workers based on LVQ and fisher-SVM deep excavation," *IEEE Access*, vol. 7, pp. 23109–23119, 2019, doi: 10.1109/ACCESS.2019.2899625.
- [20] J. Yoon, W. R. Zame, and M. V. Der Schaar, "ToPs: ensemble learning with trees of predictors," *IEEE Transactions on Signal Processing*, vol. 66, no. 8, pp. 2141–2152, Apr. 2018, doi: 10.1109/TSP.2018.2807402.
- [21] J. Zhang, M. Wu, and V. S. Sheng, "Ensemble learning from crowds," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 8, pp. 1506–1519, Aug. 2019, doi: 10.1109/TKDE.2018.2860992.
- [22] J. Dhar and N. A. Ayele, "Multi-tier ensemble learning model with neighborhood component analysis to predict health diseases," *IEEE Access*, vol. 9, pp. 138677–138715, 2021, doi: 10.1109/ACCESS.2021.3117963.
- [23] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.
- [24] V. A. Binson, M. Subramoniam, G. K. Ragesh, and A. Kumar, "Early detection of lung cancer through breath analysis using adaboost ensemble learning method," in *ACCESS 2021 - Proceedings of 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems*, Sep. 2021, pp. 183–187, doi: 10.1109/ACCESS51619.2021.9563337.
- [25] A. Saadallah and K. Morik, "Online Ensemble aggregation using deep reinforcement learning for time series forecasting," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics, DSAA 2021*, Oct. 2021, pp. 1–8, doi: 10.1109/DSAA53316.2021.9564132.
- [26] W. Li, X. Zhang, Y. Dong, Y. Liu, and K. Chao, "Cloud platform operating data preprocessing method based on I-GMM algorithm," in *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Mar. 2021, pp. 859–863, doi: 10.1109/IAEAC50856.2021.9390724.
- [27] I. Pisa, I. Santin, J. L. Vicario, A. Morell, and R. Vilanova, "Data preprocessing for ANN-based industrial time-series forecasting with imbalanced data," in *European Signal Processing Conference*, Sep. 2019, vol. 2019-September, pp. 1–5, doi: 10.23919/EUSIPCO.2019.8902682.
- [28] Z. Tang, G. Zhao, G. Wang, and T. Ouyang, "Hybrid ensemble framework for short-term wind speed forecasting," *IEEE Access*, vol. 8, pp. 45271–45291, 2020, doi: 10.1109/ACCESS.2020.2978169.
- [29] S. N. Haider, Q. Zhao, and B. K. Meran, "Automated data cleaning for data centers: A case study," in *Chinese Control Conference, CCC*, Jul. 2020, vol. 2020-July, pp. 3227–3232, doi: 10.23919/CCC50068.2020.9189357.
- [30] M. Gada, Z. Haria, A. Mankad, K. Damania, and S. Sankhe, "Automated feature engineering and hyperparameter optimization for machine learning," in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Mar. 2021, pp. 981–986, doi: 10.1109/ICACCS51430.2021.9441668.
- [31] UCI Machine Learning Repository, "Chronic_Kidney_Disease Data Set," 2015. <https://archive.ics.uci.edu/ml/datasets>.




BIOGRAPHIES OF AUTHORS

Prof. Yasmeen Shaikh    is an Assistant Professor in the Department of Computer Science and Engineering, at KLS's Vishwanathrao Deshpande Institute of Technology, Haliyal, Karnataka, India. She obtained her Bachelor of Engineering in Computer Science and Engineering from Hirasugar Institute of Technology, Nidasoshi, Karnataka, India. She received her master's degree in technology from SDM College of Engineering and Technology, Dharwad, Karnataka, India. She is pursuing Ph.D from VTU Belagavi. She has guided more than 30 UG projects (2 per year). She has published 03 papers in International Journal and 03 papers at International Conferences. Her research interests include the applications of machine learning, big data analytics, internet of things, cloud computing, networking. She can be contacted at email: y.s.shaikh1210@gmail.com.



Prof. Vasudev Parvati    is an Associate Professor in the Department of Information Science and Engineering, at S.D.M. College of Engineering and Technology, Dharwad, Karnataka, India. He obtained his Bachelor of Engineering from SDM College of Engineering and Technology, Dharwad, Karnataka, India. He received his Master's degree in Technology from VTU, Belagavi, Karnataka, India. He is pursuing Ph.D. from Karnataka University, Dharwad, Karnataka, India. He guided U.G projects (02 each year) and 05 PG students projects. He has published 05 papers at National and 06 international journals. He is a member of BOE in CS, KUD. Ex-member of BOS in CS/IS in NIE MYSORE & BOE in ISE, SDMCET, Dharwad. His research interests include networking, machine learning, and internet of things. He can be contacted at email: vkparvati@gmail.com.



Dr. Sangappa Ramachandra Biradar    is a Professor in the Department of Information Science and Engineering, at S.D.M. College of Engineering and Technology, Dharwad, Karnataka, India. He obtained his Bachelor of Engineering from BLDEA's College of Engineering & Technology, BIJAPUR. He obtained his Master of Technology from M.I.T., Mahe, Manipal. He received his Ph.D. from Jadhavpur University, Kolkatta, India. He is guiding 06 Ph.D. students at Visvesvaraya Technological University, Belagavi, Karnataka and guided U.G. (2 per year) and 15 PG. students projects. He has received the International Travel Support (ITS) from DST, Govt of India, to attend the 2011 World Congress in Computer Science, Computer Engineering, and Applied Computing, Las Vegas, Nevada, USA. He has also won third prize in a paper in International Conference on Computers and Devices for Communication, Kolkatta, 2006. He has published 20 papers at International Journal and 32 at International and national conferences. He is a life member of ISTE, ACM and IAENG. His research interests include networking, machine learning, cloud computing, big data analytics, and internet of things. He can be contacted at email: srb.sdm@gmail.com.