Ensemble Based Hybrid Variable Selection Method for Heart Disease Classification

D. Rajeswari, K. Thangavel

Department of Computer Science, Periyar University, Salem 636 011, Tamil Nadu, India.

_____

Abstract: In this paper, we proposed an ensemble-based hybrid variable selection model that aggregates various variable selection methods results based on majority voting approach to select a risk features subset in the heart disease datasets. The performance of the devised framework is evaluated using Z-Alizadeh Sani heart disease dataset from the UCI repository. Besides, we also compare this devised method with three non-ensemble variable selection methods namely the Chi-square test, Recursive Feature Elimination, and L1-Regularization. The selection process of the devised method is validated through a random forest classifier, it performs better in terms of specificity, sensitivity, accuracy, precision, and f1-score. The proposed method significantly enhances the accuracy of the heart disease classification model.

Keywords: Ensemble Variable selection, Heart disease, Chi-square, Recursive Feature Elimination, L1 Regularization, Majority vote, Random Forest.

_____

1. Introduction

Medical records consist of a large dimension, superfluous information. To process these types of data becomes awkward. So, it is crucial to prepare the data before incorporating it into any classification algorithms. The prediction of illness becomes unambiguous if the data is explicit and free from the curse of dimensionality. In the medical domain, machine learning has an advent to analyze medical records. Variable Selection (VS) is one of the data preprocessing method in machine learning contributes more to find the hidden patterns from the raw data. Currently, a huge amount of health-related record is generating daily in the health care sector namely Electronic Health Records (EHRs) include clinical information of patients, the diagnostic study of patients, demographic behavior reports, medical history, physical examination, and information related to medications. There is a need to handle the large dimension curative care records by data preprocessing techniques in order to avoid misclassification results. The variable depletion process plays an important part in eradicating unnecessary variables; it helps in boosting the performance of machine learning algorithms. The importance of the VS process and the issues related to the data dimension were discussed in **(Oreski. D et al., 2017)**.

Heart Disease (HD) is a condition that affects the normal function of the heart organ it leads to heart failure due to the enriched oxygen does not flow properly in the blood arteries of the circulatory system, so it ruins our body and leads to unexcepted events like cardiac arrest, sudden death, and heartattack. So, early diagnosis of heart illness is a very essential role to avoid morbidity among patients. Heart disease prediction competes for a vital part in the curative care domain. Though, the health care sector needs prediction and decision support systems to detect HD at an early period, to develop a decision support framework that can offer good precision outcomes with a minimum number of informative variables. This work aims to choose the impact risk factors of heart disease. In this study, we devised an ensemble-based hybrid variable selector model that aggregates the outcomes of filter, wrapper, and embedded method to predict the risky factors of the HD.

One of the prolific fields in machine learning are ensemble learning. Generally, it has been used for classification problems, it aggerates the output of multiple learning models, it is better than using a single learning model and it provides better outcomes **(Bolon-Canedo & Alonso-Betanzos, 2019)**. However, an ensemble variable selection is the same as the ensemble approach **(Pardo. B.S. et al., 2017)**. The ensemble

_____

variable selection approach works based on fusing the results exploits from the several variable selectors. Variable Selection (VS) is the task of finding informative and eliminating superfluous variables from the original dataset, is called variable or attribute or subspace or feature selection or subset selection **(Guyon & Elisseeff, 2003)**. Variable selection is one of the key processes in a classification task.

The objectives of VS are it escalates the performance of the classifier, diminishes the computational cost, processing time, it avoids overfitting and underfitting in the learning model **(Fard. S.M.H. et al., 2013, Jain & Singh, 2018)**. There are two different ways of evaluating the variables of a data set subset evaluation and individual evaluation. The subset selection method estimates a group of variable subsets according to optimal criteria. In an individual evaluation approach, a rank score is assigning to each variable **(Pisica. I, 2013)**. Presently, VS techniques are divided into three groups: filter, wrapper, and embedded **(Tarek. S, 2017)**. Filter methods quantify the feature degree, which is totally independent of the classification model **(Duch. W. et al., 2002)**. The evaluation of each feature is measure through the correlation with the target class. It works based on entropy or statistic measures. This method is computationally fast; low complexity in computation and lightweight process as compared with other methods. This method works well in high-dimensional spaces. Generally, it ignores the optimal subsets of variables **(Tibshirani. R, 1994)**.

Wrapper methods, which act as a black box it explores the importance of variable subsets along with the classification technique, in this method, the VS technique is embedded in the learning model. To evaluate the quality of the selected variable subset using a cross-validation technique. It is very expensive and works slowly as compare to the filter-based method. The issue in this approach is high computational cost due to repetitive execution with all combinations of the variables, it depends on the characteristic of the classifier model. It provides good classification accuracy **(Nahar. J, 2013)**.

The Embedded method, which is the combination of filter and wrapper methods. It evaluates the variable importance along with certain criterion that is engendered by the classification model **(Tarek.S.et al., 2017)**. Some advantages of this method are low computational cost; it works better for huge variable data samples; it automatically eliminates the non-informative variables and vulnerable to overfitting. The issue in this method is the choice variables subset depends on the characteristics of the classification technique **(Kohavi & John 1997)**.

To sum up, the filter method avoids the interaction measure among the independent features. The wrappers potentially bring out the compact features subset, but it is unfeasible to the search size space. Finally, the embedded methods process the dataset in the best way, but the decision depends on the classification model, and it suffers a hypothesis during the selection process and from the feature redundancy. In this work, we propose an aggregation of filter, wrapper, and embedded method to utilize the merits of each method and mitigating their weakness to find the risky variables which is the reason for the cause of heart disease.

The organization of the paper is as follows. Section 2 describes some of the related works. The dataset description and the proposed methodology are described in Sections 3 and 4 respectively. Section 5 deals with the experiment and result analysis and discussion, Section 6 explains the performance metrics. The conclusion part and future works are given in Section 7.

2. Related Works

This section describes previous research in the different VS techniques, several assessment criteria, and its significance in the classification process.

Wiharto.W. et al., (2016) proposed a framework which was an aggregation of the synthetic minority oversampling procedure, variable selection technique, and the C4.5 classifier, and executed implementation on the CHD dataset with 20 features from the University of California Irvine (UCI). The model achieved 84.2% accuracy, 74.7% sensitivity, and 93.7% specificity. **Babagolu,I et al., (2010)** developed an automatic detection of Coronary Artery Disease (CAD) using a combination of Principal Component Analysis (PCA) and Support Vector Machine (SVM). The proposed framework achieved an accuracy of 84.6%. **Vivekanandan,T., & Iyengar, S.N., (2017)**. contributed a modified Differential Evolution (DE) algorithm for selecting optimal features of cardiovascular disease and the prediction was carried out using Fuzzy AHP and Feedforward Neural Network. The proposed hybrid model achieved an accuracy of 83%. **Avci, E., (2009)** proposed a genetic support vector machine, structured with the composition of feature extraction and classification for the diagnosis of heart valve disease from doppler signals. The automatic heart disease diagnosis system was developed by **Amin, M.A., et al., (2014)** using Multilayer Perceptron and Adaptive Neuro-Fuzzy Inference Systems (ANFIS). The feature selection methods for the detection of cardiac arrhythmia were developed by **Mitra, M., et al., (2013)**, the composition of the correlation-based feature

selection approach with incremental back propagation neural network, and Levenberg-Marquardt (LM) classification techniques have been deployed. An ANN-based diagnostic model for coronary heart disease proposed by **Yu, O., et al., (2012)** using genetic and non-genetic related features. The performance of feature selection methods is evaluated in **Pisica, I. et al., (2013)**, on the dataset of different characteristics like irrelevant, noisy. Several feature selection methods and their merits, demerits, and then analyze the classification models for the prediction of chronic disease are explained in **Kohavi, R & John, G., (1997)**. **Zhang, Z., et al., (2014)** proposed a novel feature selection method for heartbeat classification with ECG data, namely the OvO method, which can improve the performance of SVM. The classifier accuracy of the proposed feature selection obtained 86.66%, it performs well than the other feature selection techniques. A hybridized model was proposed by **Ayar,M., & Sabamoniri, S., (2018)** to select the optimal features using the Genetic Algorithm and Decision Tree with the C4.5 algorithm to classify cardiac arrhythmias into normal and abnormal cases using the ECG signal dataset. **Sarkar, C., et al., (2014)** proposed a technique that ensembles several variable selection methods based on the rank aggregation combination method. This proposed technique enhances the accuracy of the classifier by 3-4% than other conventional techniquesnamely information gain, chi-square, and symmetric uncertainty. The computational intelligence techniques were investigated by **Nahar, J., et al., (2017)** for the detection of HD using the Medical Knowledge-driven Feature Selection (MFS) process. **Oreski, D et al., (2017)** observed that the featureselection process directly impacts the quality of the classifier algorithm.

From the literature review, it is observed that many authors investigated the importance of ensemble techniques, feature selection processes and also, we noted that the researchers had used several evaluation criteria to reduce the variable size in the huge dimensional dataset to improve the quality of learning model. Motivated by the prior investigations in data mining, machine learning, the variable selection techniques with numerous calculation measures and the several searching approaches were used to accomplish the variable selection, and a classification technique is adopted to predict the HD using Z-Alizadeh Sani dataset. In this work, we aggregate VS techniques using the ensemble approach based on the majority voting scheme was proposed and evaluate using the random forest classifier.

3. Dataset Description

In this study, the dataset of heart failure patients is taken from the UCI repository namely the Z-Alizadeh Sani heart disease dataset **(https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani)** has been used. The characteristics of the dataset are shown in Table 1.

Table.1. Characteristics of Z-Alizadeh Sani HD dataset

Dataset	No of Samples	Total Features	Class Labels Counts
Z-Alizadeh Sani	303	56	2

Table 2. Z-Alizadeh sani dataset features

Variable Initial	Variable Name	Description
1	Age	Respondent Age
2	Weight	Respondent Weight
3	Length	Respondent Height
4	Sex	Respondent Gender
5	Bmi	Body Mass Index Kb/m2
6	Dm	Diabetes Mellitus
7	Htn	Hypertension
8	Current smoker	Smoking Status
9	Ex-smoker	Individual who quit cigarette consumption habit
10	Fh	Family History
11	Obesity	Stoutness of the body
12	Crf	Chronic Renal Failure
13	Cva	Cerebrovascular Accident

14	Ad	Airway Disease
15	Td	Thyroid Disease
16	Chf	Congestive Heart Failure
17	Dlp	Dyslipidemia
18	Bp	Blood Pressure (mm Hg)
19	Pr	Pulse Rate (ppm)
20	Edema	
21	Wpp	Weak Peripheral Pulse
22	Lr	Lung rales
23	Sm	Systolic Murmur
24	Dm	Diastolic Murmur
25	Tcp	Typical Chest Pain
26	D	Dyspnea
27	Fc	Function Class
28	At	Atypical
29	Ncp	Nonanginal Chest Pain
30	Ecp	Exertional Chest Pain
31	Low th ang	Low-Threshold Angina
32	R	Rhythm
33	Qw	Q Wave
34	Se	St Elevation
35	Sd	St Depression
36	T	Tinversion
37	Lvh	Left Ventricular Hypertrophy
38	Prp	Poor R Progression
39	Fbs	Fasting Blood Sugar
40	Cr	Creatine (mg/dL)
41	Tg	Triglyceride (mg/dL)
42	Ldl	Low-Density Lipoprotein (mg/dL)
43	Hdl	High-Density Lipoprotein (mg/dL)
44	Bun	Blood Urea Nitrogen (mg/dL)
45	Esr	Erythrocyte Sedimentation Rate (mg/dL)
46	Hb	Hemoglobin (g/dL)
47	K	Potassium (mEq/lit)
48	Na	Sodium (mEq/lit)
49	Wbc	White Blood Cell (cells/mL)
50	Lymph	Lymphocyte (%)
51	Neut	Neutrophil (%)
52	Plt	Platelet (1000/mL)
53	Ef-tte	Ejection Fraction (%)
54	Rr	Region with RWMA
55	Vhd	Valvular Heart Disease
56	Cath	Target Class Label

4. Proposed Methodology

This section explains the devised framework methodology. Figure.1 illustrates the flowchart of the devised methodology. The proposed technique comprising of three individual base selectors including Chi-Square, RFE, and L1 regularization. The design of this combining strategy is that every individual feature selector utilizes its own nature of evaluation measure and may produce diverse scoring outcomes when implementing the same arbitrary dataset. The aim behind this is the aggregation of variable selector algorithms to integrate the consensus properties of chi-square, RFE, and L1 to obtain more stable results about the subset of informative features. The aggregation of the individual variable selection techniques used for merging the outcoming variable subsets generated by three different feature selection methods into a single subset of features. Here, we utilizing the majority voting technique for performing the aggregation.
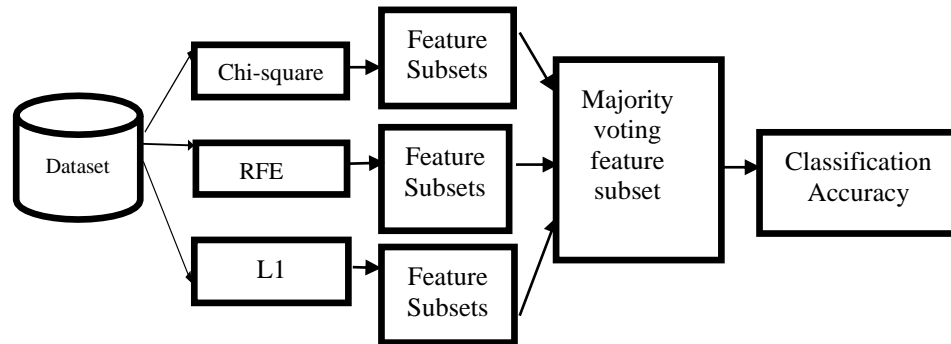

Figure 1. Overview of the proposed methodology

Algorithm 1. Ensemble based Hybrid Variable Selection Model

Input:
$S=\{s_1,s_2,s_3,s_4,\ldots\ldots,s_n\}$ set of n samples in the dataset
$V=\{v_1,v_2,v_3,v_4,\ldots\ldots\ldots,v_n\}$ set of n features
$C=\{c_1,c_2\}$ set of classes

Output:
Compact Feature Subset;

Data N: Number of Variable Selection Methods (VSM)
Data S: Select the optimal subset of features
1. For each n from 1 to N do
2. Implement the VSM to the original dataset
3. Select the variable V has a high rank
4. Obtaining the vote for high-rank variables
5. End
6. A= Calculate the total votes for each variable
7. At = Select the compact subset of variables have majority votes
8. Build classifier with the selected subset of variables At
9. Obtain Classification result C.

4.1 Chi-Square (χ^2)

Chi-Square (χ^2) is an information-theoretic function. This method assesses the connection between the predictor variables and response variable (**Bahassine. S et al., 2020**). It selects the optimal features (*of*) in associationwith the class value *C*. Chi-square measure can be expressed as follows:

$$x^2(of, C) = \frac{M(XN-PY)^2}{(X+P)(Y+N)(X+Y)(P+N)} \qquad (1)$$

In the above equation. (1), illustrate the chi-square test formula is related to the variable selection function to choose the optimal variable to connect to the class value. Here M is the number of instances, X is the total number of positive instances that present in the variable *of*, Y is the number of negative instances that present in the variable *of*, P is the number of positive samples that do not contain the variable *of*, N is the number of negative samples that do not contain variables *of*.

This method works based on statistical measures. It estimates the score to each variable and assigns it. Finally, all these score values are integrating into a final score of χ^2 (o*f, C*).

4.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is an optimization technique to determine the variable a subset, it determines the optimal subset of variables along with the classifier model **(Yana & Zhang 2015)**. It removes the least impact variables until the best subset variables are obtained. It obtains the optimal variable subset by using the cross-validation technique. RFE pruned the features from a constructed model by fitting the model iteratively and at each iteration, it removes the worst-performing feature, this process is recursively repeated till all the features in the dataset are exhausted, then it assigns weights to features.

4.3 L1 Regularization

L1 regularization is used for feature elimination. This method embeds a feature selection by applying the shrinking process to deal with an assessment on significant features, which coefficient value is equal to zero and this method uses the class separability as a criterion for variable selection, it can estimate the optimal solution of feature subset with a non-zero initial a point in order to decrease the regularized objective function **(Fonti & Belitser 2017)**. The goal of this method is to improvise the prediction accuracy and to minimize the prediction error, feature selection. It adds an L1 regularized penalty to the objective function **(Guo. S et al., 2017)**.

4.4 Classification Method

The efficiency of the proposed method is assessed through the Random Forest (RF) algorithm. The RF is based on the idea of ensemble learning method it constructs the multiple decision trees **(Breiman. L 2001)**, RF is a supervised machine learning algorithm based on the ensemble approach, is first proposed by L. Breiman, it is a group of decision trees wherein each tree is prepared by an alternate subset of the training samples is trained. This is conducted to boost the generalization potential of RF **(Subasi. A et al., 2019).** To build every one of the trees, RF picks a bootstrapped subset of the arbitrary training samples comprising about two-third of the training sample and the target value of the test set samples is categorized by the majority vote of the trees. In recent years numerous applications used this classification technique for the decision-making process, it develops a powerful model than other machine learning techniques such as logistic regression, classification and regression tree (CART) **(Chen. W et al., 2017)**. It performs well and yields the best outcome that is why it was used in this work. The aim of this work is to obtain the most desirable feature subset not to analyze the efficiency of the learning model.

5. Result Analysis

In this paper, we incorporated our proposed EHVS method to select an optimal subset of features for a random forest classifier algorithm that categorizes the sample as a heart disease patient on or non-heart disease patient. In order to assess the ability of the proposed framework we utilize the Z-Alizadeh Sani HD dataset as shown in Table.1. After implementing Algorithm 1. the outcomes of the three variable selection technique to obtain the impact subset of features, through which variable has the highest count of votes.

Table 3. shows the three different variable selection method uses different evaluation measures, some variables are similar among different VS techniques. Using a combination method, called majority voting (MV) approach the variables 1,51,43,3,22,37,18,17,27,5, and 48 (specified in boldface) similar among the three variable selection techniques; those are the final subset aggregation of delegate variables for the subsequent analysis for the classification model.

Table 3. Feature dimension reduction using variable selection methods

Variable Selection Methods	Variables selected
Chi-square	**22**,**48**,**1**,**18**,**43**,**5**,**51**,6,**3**,38,**37**,40,36,34,28,14,**17**,53,50,**27**
RFE	**1**,**3**,9,**18**,**17**,**22**,27,30,**27**,**37**,41,**48**,**43**,44,45,46,**5**,**51**,54
L1	**1**, **3**,**5**,**17**,**22**,**18**,**37**,39,**27**,**43**,**48**,**51**,53
EHVS	1,51,43,3,22,37,18,17,27,5,48

Table 3 show that 11 out of the 55 variables are selected from the Z-Alizadeh Sani dataset for the subsequent classification analysis using the EHVS method. This relevant subset of variables will subsequently be implied as input to the RF classifier. One more thing to observe from the above table is trying to focus on features {age, length, Hdl, Neut, Bp, Fc, Na, Bmi, Lvh, Lr, Dlp} are the attributes metntion in Table 2 to have high classification accuracy and those attributes are the valuable reason for heart illness arising.

6 Performance Metrics for Classification Effect

Table 4, shows the confusion matrix, consist of True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), False Negative (*FN*). In this paper, sensitivity, specificity, precision, accuracy, and F- measure are five evaluation metrics that were used to calculate the classifier performance based on the confusion matrix.

Table.4. Confusion matrix

Predictive values			
		Positive	Negative
Actual class	Positive	True Positives (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

where, TP - True Positive represents the instance number accurately predicted as a positive instance (patients with heart disease), TN – True Negative represents the instance number accurately classified as a negative instance (non-HD patients), FP- False Positive represents the instance number wrongly predicted as a positive instance, FN – False Negative represents the instance number wrongly predicted as a negative instance.

Equations (2), (3), (4), (5), and (6) represents the performance metrics:

$$Specificity = \frac{TN}{TN+FP} \times 100\% \qquad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \times 100\% \qquad (3)$$

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN} \times 100\% \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \qquad (5)$$

$$F1\ score = \frac{2tp}{2(tp+fp+fn)} \times 100\% \quad (6)$$

The proposed system yield high classification accuracy. In this study, we compare the efficiency of the devised EHVS framework with each conventional variable selection techniques in both before and after variable selection scenarios. Table 5. represents the outputs of the evaluation metrics of the RF algorithm utilizing the arbitrary dataset with 55 attributes, the chi-square method with 20 attributes, L1 regularization method with 15 attributes, and RFE with 20 attributes, and our developed EHVS method with 11 attributes.

Table.5. Classification Results of Feature Selection Methods with Random Forest Classifier

Variable Selection Methods	Sensitivity	Specificity	Precision	Accuracy	F1 Score
Full dataset	68.42%	83.33%	52%	80.22%	59.09%
Chi-Square	75.00%	89.55%	72%	85.71%	73.47%
RFE	78.26%	89.71%	72%	86.81%	75%
L1	76.92%	92.31%	80%	87.91%	78.43%
EHVS	80.77%	93.85%	84.0%	90.11%	82.35%

Figure 2. shows the evaluation measures of EHVS and other VS techniques. Our proposed model outperforms better in terms of specificity, sensitivity, precision, f1 score, and accuracy. It is clearly observed that the proposed model provides good classification accuracy of 90.11% for the Z-Alizadeh Sani HD dataset with the selected subset of features.

Table 6. shows the number of variables reduced by the devised method and other single feature selection techniques from the full set of variables
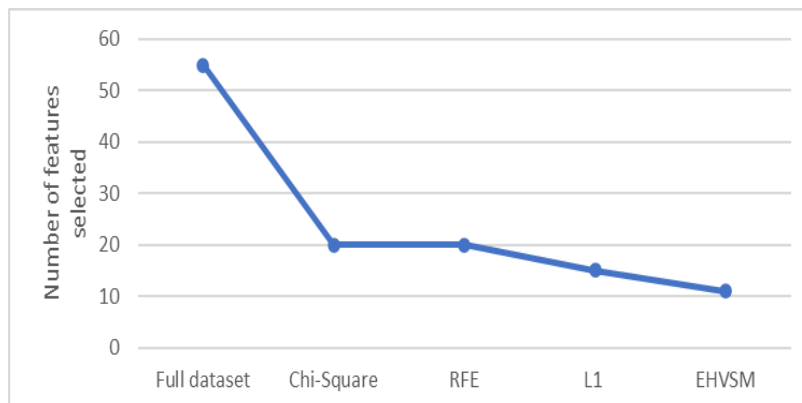
Table.6. Total number of variables selected by the devised EHVS method and non-ensemble feature selection methods

Variable selection methods	Total features
Chi-Square	20
RFE	20
L1	15
EHVS	11
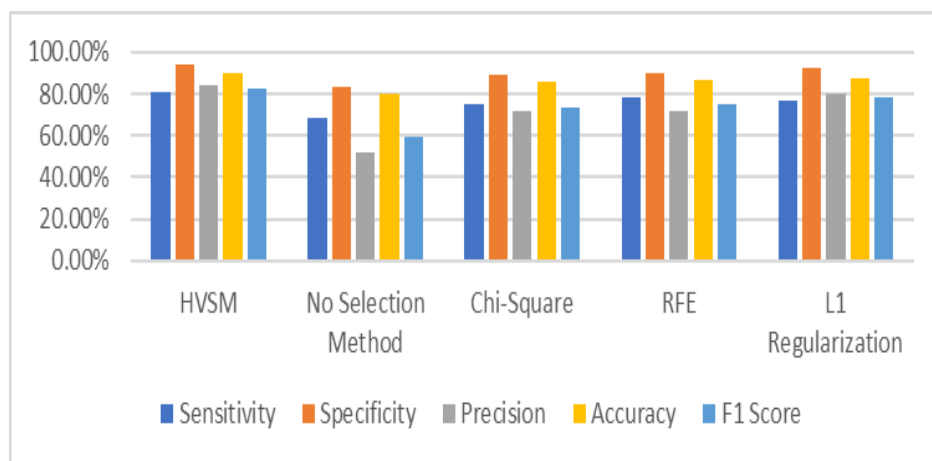
Figure 2. Performance of devise EHVS method and different variable selection methods

Figure 3. Feature reduction dimension by VS methods

Figure 3. depicts the dimension reduction of different VS techniques, the number of variables chosen by the four different VS methods, and the original dataset. The devised method selects the smaller number of variables and the significant variables which have a high count of votes, it helps to reduce the complexity of the learning model.

7 Conclusion

In this paper, we present the Ensemble-based Hybrid Variable Selector Model (EHVS) integrates several assessment criteria of the VS methods to quantify the characteristics of the highlight variables. Furthermore, Random Forest, a powerful technique was utilized to categorize the HD. From the findings, the VS techniques could diminish the dimension of the dataset and improve a classifier performance. On this basis, EHVS model merged the various VS techniques based on the idea of the ensemble method. The proposed method exploits the merits of the diversity of base selectors, and it measures the delegate features. Finally, the proposed system achieved good results in the classification performance and stability in the HD dataset. Additionally, the proposed framework is a non-invasive method, simple, provides better precision outcome, and cost-effective approach. Thusly, it tends to be utilized as a decision support system for analysis of HD which can accomplish early treatment and early prediction of HD. Generally, this approach is appropriate for the large dimensional dataset. In future work, more various classifiers can be used as a base classification algorithm. In addition, this framework can be applied to other benchmarks, particularly multiclass datasets.

Acknowledgement

References

[1]. Amin, M.A, Alqudah, A, Adwan, O., (2014). Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches.*Journal of Software Engineering and Applications*, 7, 1055-1064.

[2]. Avci,E., (2009). A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. *Expert Systems with Applications*, 36, 10618-10626.

[3]. Ayar, M, Sabamoniri, S., (2018). An ECG-based feature selection and heartbeat classification model using ahybrid heuristic algorithm. *Informatics in Medicine Unlocked*, 13, 167-175.

[4]. Babaoglu, I, Fndk, O, Bayrak, M., (2010). Effects of principle component analysis on assessment of coronary artery diseases using support vector machine. *Expert Systems with Applications*, 37, 2182-2185.

[5]. Bahassine, S, Madani, A, Sarem, M. Al, Kissi, M., (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University – Computer and Information Sciences*, 32(2), 225-231.

[6]. Bolon-Canedo, V, Alonso-Betanzos, A., (2019). Ensemble for feature selection: A review and future trends. *Information Fusion*, 52(1-12).

[7]. Breiman, L., (2001). *Random Forests. Machine Learning*, Statistics Department, University of California, Berkeley.

[8]. Chen, W, Xie, X, Wang, J., (2 0 1 7). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA*, 151, 147-160..

[9]. Duch, W, Grabczewski, K, Winiarski, T, Biesiada, J, Kachel, A., (2002). Feature selection based on information theory, consistency and separability indices", in: International Conference on Neural Information Processing, ICONIP '02.

[10]. Fard, S.M.H, Hamzeh, A, Hashemi, S., (2 0 1 3). Using reinforcement learning to find an optimal set of features. *Computers and Mathematics with Applications*, 66, 1892-1904.

[11]. Fonti, V, Belitser, E., (2017). Feature Selection using LASSO. Research Paper in Business Analytics,Vrije Universiteit Amsterdam, 1-26.

[12]. Guo, S, Guo, D, Chen, L, Jiang, Q. I., (2 0 1 7). A L1-regularized feature selection method for local dimension reduction on microarray data. Computational Biology and Chemistry, 67, 92-101.

[13]. Guyon, I, Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of MachineLearning Research*, 3, 157-1182.

[14]. Jain, D, Singh, V., (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal*, 19(3), 1-14.

[15]. Kohavi, R, John, G., (1997). Wrappers for feature selection. Artificial Intelligence", 97(1-2), 273–324.

[16]. Mitra, M, Samantha, R.K., (2013). Cardiac arrhythmia classification using neural networks with selected features. *Procedia Technology*, 10, 76-84.

[17]. Nahar, J, Imam, T, Tickle, K.S, Chen, Y.P.P., (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, 40(1), 96-104.

[18]. Oreski, D, Oreski, S, Klicek, B., (2017). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, 52, 109-119.

[19]. Pardo, B.S, Porto-Diaz, I, Canedo, V.B., Betanzos, A.A., (2017). Ensemble Feature Selection: Homogeneous and Heterogeneous Approaches. *Knowledge-Based Systems*, 118, 124-139.

[20]. Pisica, I, Taylor, G, Lipan, L., (2013). Feature selection filter for classification of power system operating states. *Computers and Mathematics with Applications*. 66, 1795–1807.

[21]. Sarkar, C, Cooley, S, Srivastava, J., (2014). Robust Feature Selection Technique Using Rank Aggregation. *Applied Artificial Intelligence*, 28(3), 243-257.

[22]. Subasi, A, Ahmed, A, Alickovic, E, Hassan, A.R., (2019). Effect of photic stimulation for migraine detection using random forest and discrete wavelet transform. *Biomedical Signal Processing and Control*, 49, 231-239.

[23]. Tarek, S, Elwahab, R.A, Shoman, M., (2017). Gene expression-based cancer classification. *Egyptian Informatics Journal*, 18, 151-159.

[24]. Tibshirani, R., (1994). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 267-288.

[25]. Vivekanandan, T, Iyengar, S.N., (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Computers in Biology and Medicine*, 90, 125-136.

[26]. Wiharto, W, Hari, K, Herianto, H., (2016). Interpretation of clinical data based on C4.5 algorithm for the diagnosis of coronary heart disease. *Healthcare Informatics Research*, 22(3), 186-195.

[27]. Yana, K, Zhang, D., (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators*, 212, 353–363.

[28]. Yu, O, Svetlanaet, G, Alexandr., (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of Cardiology*, 59(2), 190- 194.

[29]. Zhang, Z, Dong, J, Luo, X, SzeChoi, K, Wu, X., (2014). Heartbeat classification using disease-specific feature selection", *Computers in Biology and Medicine*, 46, 79-89.

[30]. https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani