Accepted Manuscript

A Random Forest based predictor for medical data classification using feature ranking

Md. Zahangir Alam, M. Saifur Rahman, M. Sohel Rahman

PII: S2352-9148(19)30019-X

DOI: https://doi.org/10.1016/j.imu.2019.100180

Article Number: 100180

Reference: IMU 100180

To appear in: *Informatics in Medicine Unlocked*

Received Date: 31 January 2019

Revised Date: 3 April 2019

Accepted Date: 5 April 2019

Please cite this article as: Alam MZ, Rahman MS, Rahman MS, A Random Forest based predictor for medical data classification using feature ranking, *Informatics in Medicine Unlocked* (2019), doi: https://doi.org/10.1016/j.imu.2019.100180.

A Random Forest based Predictor for medical data classification using Feature Ranking

Md. Zahangir Alam[a,1], M. Saifur Rahman[a], M. Sohel Rahman[a,*]

[a]*Department of CSE, BUET, ECE Building, West Palasi, Dhaka-1205, Bangladesh.*

Abstract

Medical data classification is considered to be a challenging task in the field of medical informatics. Although many works have been reported in the literature, there is still scope for improvement. In this paper, a feature ranking based approach is developed and implemented for medical data classification. The features of a dataset are ranked using some suitable ranker algorithms, and subsequently the Random Forest classifier is applied only on highly ranked features to construct the predictor. We have conducted extensive experiments on 10 benchmark datasets and the results are promising. We present highly accurate predictors for 10 different diseases, as well as suggest a methodology that is sufficiently general and is expected to perform well for other diseases with similar datasets.

Keywords:

Medical Data Classification, Feature Ranking, Random Forest, disease predictors

1. Introduction

In recent times, the application of computational or machine intelligence in medical diagnostics has become quite common. Machine intelligence aided decision systems are often being adopted to assist (but not to replace) a physician in diagnosing the disease of a patient. A physician typically accumulates her knowledge based on patient symptoms and the confirmed diagnoses. Thus diagnostic accuracy is highly dependent on a physician's experience. Since it is now relatively easy to acquire and store a large amount of information digitally, the deployment of computerized medical decision support systems has become a viable approach to assisting physicians to swiftly and accurately diagnose patients [1]. Such a system can be seen as a classification task as the goal is to make a prediction (i.e., diagnosis) on a new case based on the available records and features (of previously known cases). Such classification tasks are considered to be one of the most challenging tasks in medical informatics [2].

*Corresponding author. E-mail: msrahman@cse.buet.ac.bd; Tel.: +8801552389480; Fax.: +88029665612.

Email addresses: zahangirbd@gmail.com (Md. Zahangir Alam), mrahman@cse.buet.ac.bd (M. Saifur Rahman), msrahman@cse.buet.ac.bd (M. Sohel Rahman)

[1]Supported by an ICT (PhD) Fellowship

While various statistical techniques may be applied in medical data classification, the major drawback of these approaches is that they depend on some assumptions (e.g., related to the properties of the relevant data) for their successful application [3, 4]. To know the properties of the dataset is a difficult task and is sometimes is not feasible. On the other hand, soft computing based approaches are less dependent on such knowledge.

A number of soft computing based classifiers have been proposed and analyzed in the literature to classify medical data accurately. Abbass *et al.* proposed a system with the pareto-differential evaluation algorithm with a local search scheme, termed the Memetic Pareto-Artificial Neural Network (MPANN), to diagnose breast cancer [5]. Subsequently, Kiyan *et al.* [6] presented a statistical neural network-based approach to diagnose breast cancer. In [7], Karabatak *et al.* developed an expert system for detecting breast cancer, where, to reduce the dimensions of the dataset, Association Rules (AR) were used. Peng *et al.* proposed a hybrid feature selection approach to address the issues of high dimensionality of biomedical data, and experimented on the breast cancer dataset [8]. Fana *et al.* combined case- based data clustering and a fuzzy decision tree to design a hybrid model for medical data classification [9]. The model was executed on two datasets, WBC and liver disorders. Azar *et al.* proposed three classification methods, namely, radial basis function (RBF), multilayer perceptron (MLP), and probabilistic neural network (PNN), and experimented on a breast cancer dataset [10]. In their experiments, PNN showed better performance than MLP.

During the last three years, several works on medical data classification have been reported in the literature, albeit only on a breast cancer dataset. Examples include, but may not be limited to, the back propagation (BP-NN) approach [11], fuzzy-rough nearest neighbor method [12], PCA followed by Support Vector Machine (SVM) with Recursive Feature Elimination (SVM-RFE) [13], PCA in combination with a feed-forward neural network [14], ANN with MLP and also BP-NN [15], deep belief network (DBN) [16], SVM ensembles with bagging and boosting [17], knowledge-based system using Expectation Maximization (EM) clustering, noise removal, and Regression Trees (CART) [18]. Motivated by the promising results of [16], very recently, Karthik *et al.* [19] have worked on breast cancer classification using Deep Neural Networks (DNN) and achieved better results than others. On the other hand, Khan *et al.* have presented a model for breast cancer and Parkinson's disease prediction, in which ensembles of Evolutionary Wavelet Neural Networks have been used [20].

On the other hand, Anooj *et al.* employed weighted fuzzy rules to develop a clinical decision support system (CDSS) for heart disease prediction [21]. They first generated fuzzy rules based on historical data for better learning, and subsequently developed the CDSS based on those. Also, the fuzzy rules were weighted based on the importance of attributes. Samb *et al.* [22] proposed a modified SVM-RFE and conducted experiments on multiple medical datasets (e.g., SPECT Heart Data). They have also incorporated local search operators into their algorithm. Jaganathan *et al.* [23] employed feature selection using the concept of fuzzy entropy. An earlier work by Polat

and Gunes [24]

also had proposed a feature selection approach based on Kernel F-Score. Jabbar *et al.* [25] developed a hybrid approach using *K*-Nearest Neighbor (KNN) and Genetic Algorithm (GA). An intelligent medical decision method system on evolutionary strategy was developed in [26] using Neural network (NN), GA, SVM, KNN, MLP, RBF, PNN, self-organizing map (SOM), and Naive Bayes (NB) as classifiers.

Khanmohammadi *et al.* developed a CDSS [27] and experimented on ten medical datasets and based on their experiments reported SVM to be the most desirable classification algorithm for developing CDSS. Dennis *et al.* presented an efficient medical data classification system based on Adaptive Genetic Fuzzy System (AGFS) [28]. In this methodology, rules are first generated from data, and then optimized rules selection is performed using GA. Seera

et al. [29] proposed a hybrid intelligent system and conducted experiments on breast cancer, Diabetes, and Liver Disorders datasets. Alwidian *et al.* also have considered breast cancer prediction and have introduced WCBA, which is an efficient Weighted Classification (Based on Association rules) algorithm [30]. They experimentally have shown that WCBA, in most cases, outperforms the other Association Classification (AC) algorithms.

Most of the works discussed above have focused on a limited number of datasets (e.g., 1-3). On the other hand, in this paper, our focus has been on finding a general methodology for the medical datasets. The quest for a general methodology however is not new, and we do find a few attempts towards that direction in the literature. For example, a number of evolutionary Extreme Learning Machine (ELM) models have been reported for medical data classification in recent years, albeit with a slightly different focus. Mohapatra *et al.* first discussed the idea to classify binary medical dataset based on ELM [2] and very recently, Eshtay *et al.* have proposed an ELM based Competitive Swarm Optimization (CSO) technique for this task [31]. Both of these works have focused on experimentation by varying the number of hidden neurons (of the ELM). More will be discussed on this issue in a later section (Section 3.13).

The contribution of this paper is as follows.

- We propose a general methodology for medical data classification that employs a feature ranking and selection strategy followed by an appropriate training of a suitable classifier algorithm. We use a number of feature ranking strategies and the Random Forest algorithm as the final classifier for our predictors. We have conducted a thorough evaluation of our approach on 10 benchmark datasets and presented insightful discussions on the results.

- We present highly accurate predictors for 10 different diseases as well as suggest a generalized methodology that should perform well for other diseases with similar datasets. The proposed framework is also expected to be useful in any other domain that exhibits similar characteristics of features.

- We identify and report the most important features from the respective datasets based on how they contribute

in the prediction tasks, and present an insight on the results from a medical point of view. We also conduct an ablation study to verify the importance of the features selected for a model, and confirm the positive contribution thereof on the classification task.

2. Materials and Methods

2.1. Datasets

The datasets we have used was collected from University of California at Irvine (UCI) Machine Learning Repository [32]. In particular we use ten benchmark datasets, corresponding to ten diseases as described in Table 1. More details of the datasets are provided in the supplementary material. Since we want to do an independent testing of our model, for each disease, training and testing samples are separated, applying a random split following the strategy of [2] (please see Table 1). The datasets organized in training and testing samples can be downloaded from the following link to reproduce our experiments: https://github.com/zahangirbd/medical_data_for_classification.

Table 1: Brief description of the datasets used in this research

DataSet	ID	No. Of Features	Training Samples	Testing Samples
Wisconsin Breast Cancer	WBC	9	499	200
Pima Indians Diabetes	PID	8	576	192
Bupa	Bp	6	200	145
Hepatitis	Hp	19	80	75
Heart-Statlog	HtS	13	180	90
SpectF	SF	44	176	91
SaHeart	SHt	9	304	158
PlanningRelax	PRx	12	120	62
Parkinsons	PkS	22	130	65
Hepatocellular Carcinoma (HCC)	HCC	49	110	55

2.2. Model Construction Overview

A diagram of our model construction workflow is shown in Figure 1. For each disease, the same workflow is followed to create an independent model for that disease. We first check whether all features are important for the classification task. This is done using several feature ranking algorithms. Then based on the ranking, a subset of the top-ranked features are selected. Finally, the Random Forest algorithm is applied on the selected features to train and construct the final model. We apply 10-fold cross validation during the model training. Notably, cross validation is a method to evaluate a predictive model by partitioning the original sample into a training set to train the model, and a validation/test set to evaluate it. In 10-fold cross validation, the original samples are randomly partitioned into 10 equal sized subsamples, and among these subsamples a single subsample is retained as the

validation data for testing the model, while the remaining 9 subsamples are used as training data.

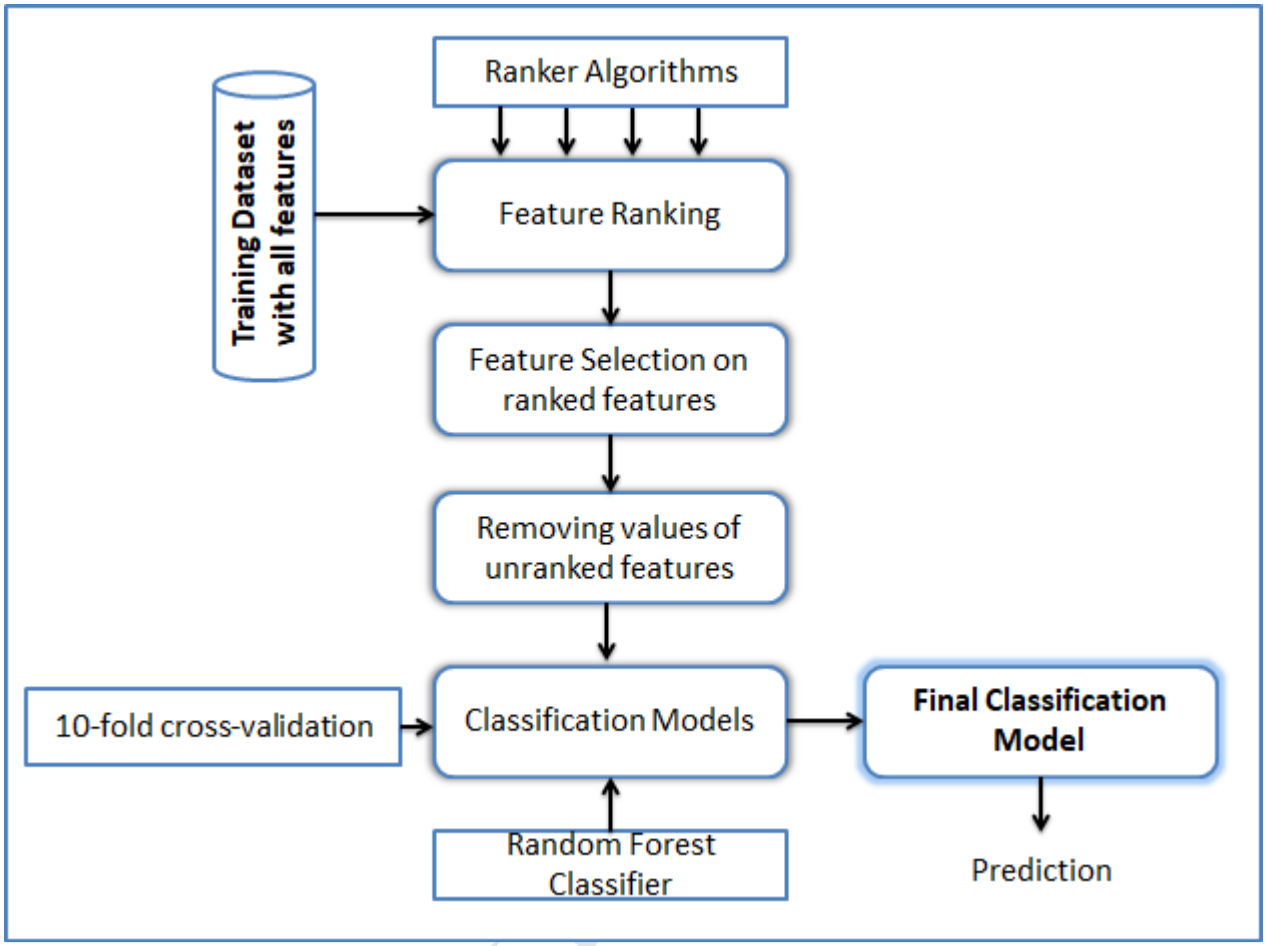


Figure 1: Model construction overview

2.3. *Feature Ranking and Selection*

Features are evaluated and ranked using some ranking algorithms. These algorithms evaluate/rank each of the features in the dataset in the context of the output variable (i.e., the *Class*). Many algorithms are available for feature ranking in the literature [33, 34, 35]. We use the following feature ranking options from the Waikato Environment for Knowledge Analysis (popularly known as *Weka*) tool [36]: InfoGainAttributeEval, GainRatioAttributeEval, CorrelationAttributeEval, OneRAttributeEval, ReliefFAttributeEval, RandomForest and SVM. More details of these algorithms are provided in the supplementary materials. Once the ranked features are identified, several combinations of features are selected from among the features based on their ranking scores.

2.4. Final Model Construction

For each disease, we construct multiple models. In particular, for each disease, we identify 3 separate subsets of highly ranked features, and thus in the sequel construct three separate models. We refer to these models using the following form: $\text{Model}^{<datasetID>}_{<I>}$, $\text{Model}^{<datasetID>}_{<II>}$ and $\text{Model}^{<datasetID>}_{<III>}$. Additionally, we have constructed a baseline model using all the features, i.e., ignoring the feature ranking exercise; this model is referred to as $\text{Model}^{<datasetID>}_{Base}$. For example, for Diabetes, we have four models, namely, $\text{Model}^{\text{PID}}_{\text{Base}}$, $\text{Model}^{\text{PID}}_{\text{I}}$, $\text{Model}^{\text{PID}}_{\text{II}}$ and $\text{Model}^{\text{PID}}_{\text{III}}$. We use The Random Forest algorithm as the classifier to train and construct the model and apply 10-fold cross validation.

3. Results

All experiments have been conducted on a CPU with Intel(R) Core(TM) i5-7200U (@ 2.50 GHz) having 8 GB of RAM and running the Windows 10 operating system. For applying various classification algorithms, Weka 3.8 [36] has been used. We evaluate the performance of the models using different popular metrics from the literature, namely, accuracy, sensitivity (or recall), precision, F-score, Area under Receiver Operating Characteristics Curve (AUROC), Area under Precision-Recall Curve (AUPR), and Root mean squared error (RMSE). Details of these metrics and relevant notions have been discussed in the supplementary material. All through the experiments, the 10-fold cross validation result is the average of the results.

For each disease, we proceed with the following experimental plan. Recall that, in addition to the base model (with all features), based on different ranking algorithms, we train three other models taking different subsets of the top-ranked features. We evaluate these four models and identify the one that performs the best. Then we take the base model and the best performing model and conduct independent testing on these two only. Finally, based on the independent testing results, we provide a comparison with the state-of-the-art, considering each disease separately.

The supplementary material additionally describes the mapping between feature names and its numbers, and reports the detailed ranking results for each dataset.

3.1. Results on Breast Cancer dataset

Based on the results of feature ranking, three subsets of features are finalized, and corresponding models are constructed as follows.

- $\text{Model}_{\text{I}}^{\text{WBC}}$: Based on the results produced by the ranker, *GainRatioAttributeEval*, Feature 1 is removed; the rest of the features are used to construct this model.
- $\text{Model}_{\text{II}}^{\text{WBC}}$: Considering the results of the rankers, *InfoGainAttributeEval* and *GainRatioAttributeEval*, Features 1, 9 are excluded while constructing this model.

- $Model_{III}^{WBC}$: Based on the outcome of the ranker, *InfoGainAttributeEval*, Features 1, 4, 5, 8, 9 are excluded and the other four features are used to construct this model.

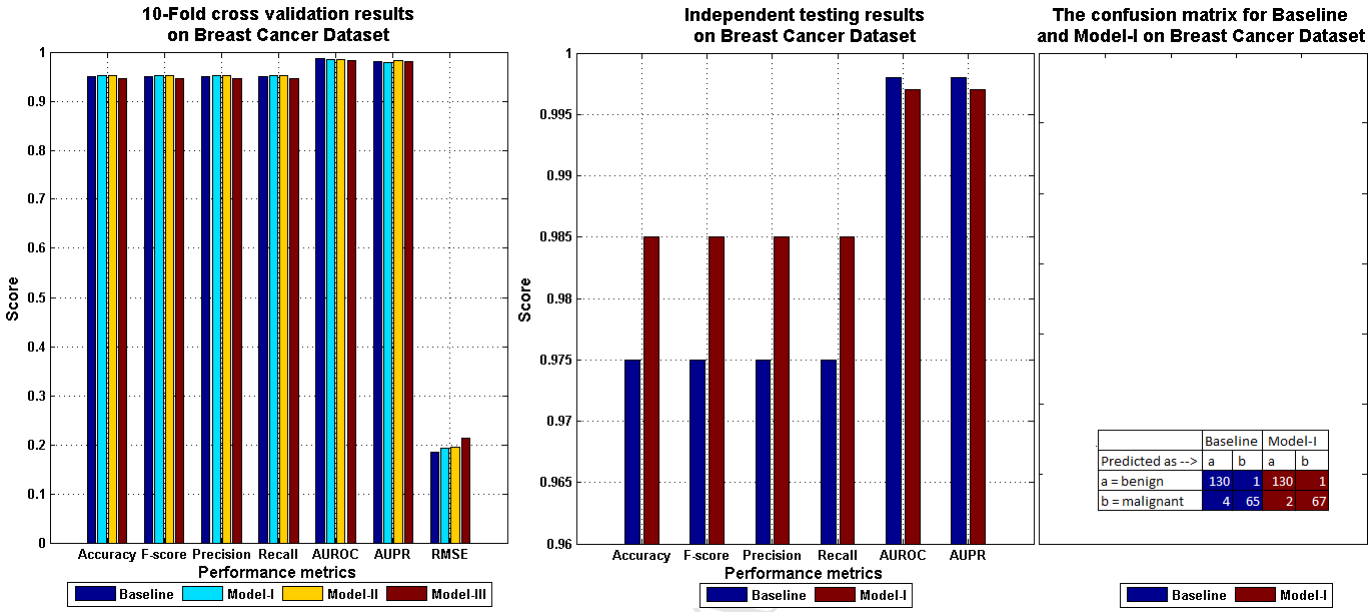


Figure 2: Results of the models on the Breast Cancer Dataset

Figure 2 (left panel) reports the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is evident that $Model_{I}^{WBC}$ and $Model_{II}^{WBC}$ perform better than $Model_{III}^{WBC}$ and $Model_{Base}^{WBC}$. Although the performances of $Model_{I}^{WBC}$ and $Model_{II}^{WBC}$ are similar, considering the RMSE value, $Model_{I}^{WBC}$ seems to have a slight edge over $Model_{II}^{WBC}$. As the best performer, we conduct independent testing for $Model_{I}^{WBC}$ and compare the results with the baseline ($Model_{Base}^{WBC}$). Figure 2 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.2. Results on Diabetes dataset

Based on the results of feature ranking, three subsets of features are finalized, and corresponding models are constructed as follows. The detailed ranking results are provided in the supplementary materials.

- $Model_{I}^{PID}$: Based on the outcome of the rankers, *InfoGainAttributeEval* and *GainRatioAttributeEval*, Feature 3 is excluded; the rest of the features are used to construct this model.

- $Model_{II}^{PID}$: Considering the ranked features produced by the rankers, *GainRatioAttributeEval* and *Correlation- AttributeEval*, Features 3, 4 are excluded while constructing this model.

- $Model_{III}^{PID}$: Features 3, 7 are excluded based on the outcome of the ranker, *InfoGainAttributeEval*; thus the model is constructed using the rest of the features.

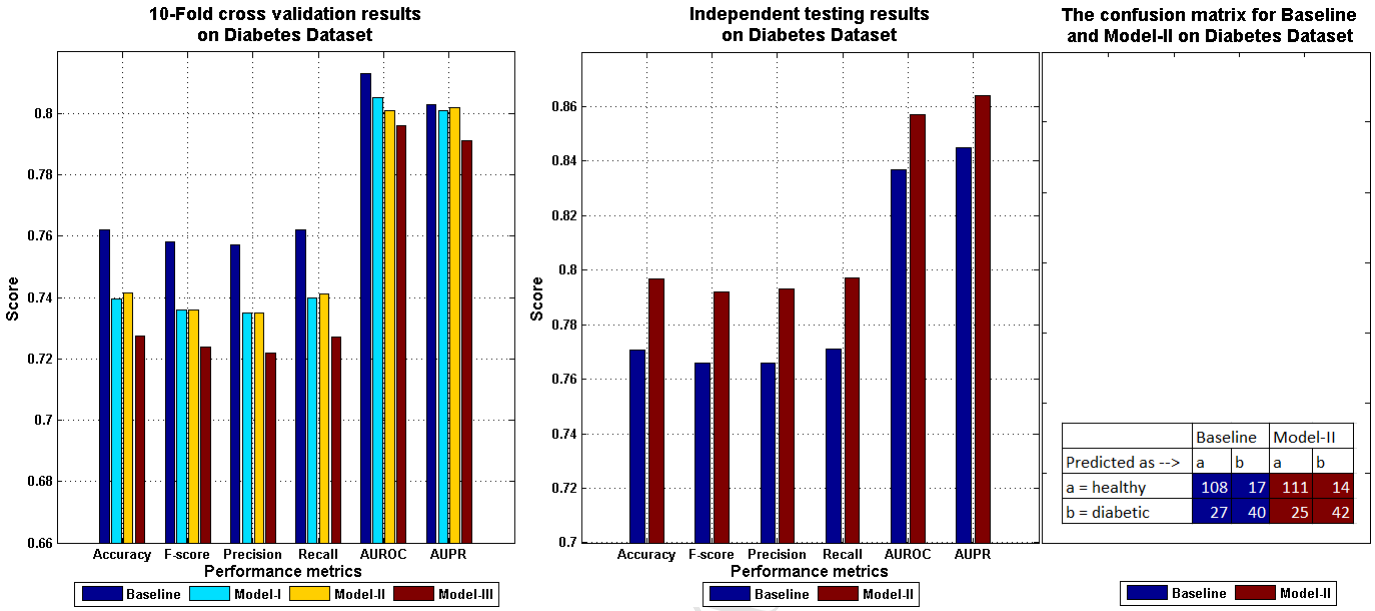


Figure 3: Results of the models on the Diabetes Dataset

Figure 3 (left panel) presents the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{II}^{PID}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{II}^{PID}$ and compare the results with the baseline ($Model_{base}^{PID}$). Figure 3 shows the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.3. Results on Bupa Dataset

Based on the results of feature ranking, three subsets of features are finalized and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{Bp}$: Based on the outcome of the rankers, *InfoGainAttributeEval* and *GainRatioAttributeEval*, Feature 1 is excluded; the rest of the features are used to construct this model.
- $Model_{II}^{Bp}$: Considering the ranked features produced by the ranker *CorrelationAttributeEval*, Feature 3 is excluded while constructing this model.
- $Model_{III}^{Bp}$: Features 1, 3 are excluded based on the outcome of the ranker, *GainRatioAttributeEval*; thus the model is constructed using the rest of the features.

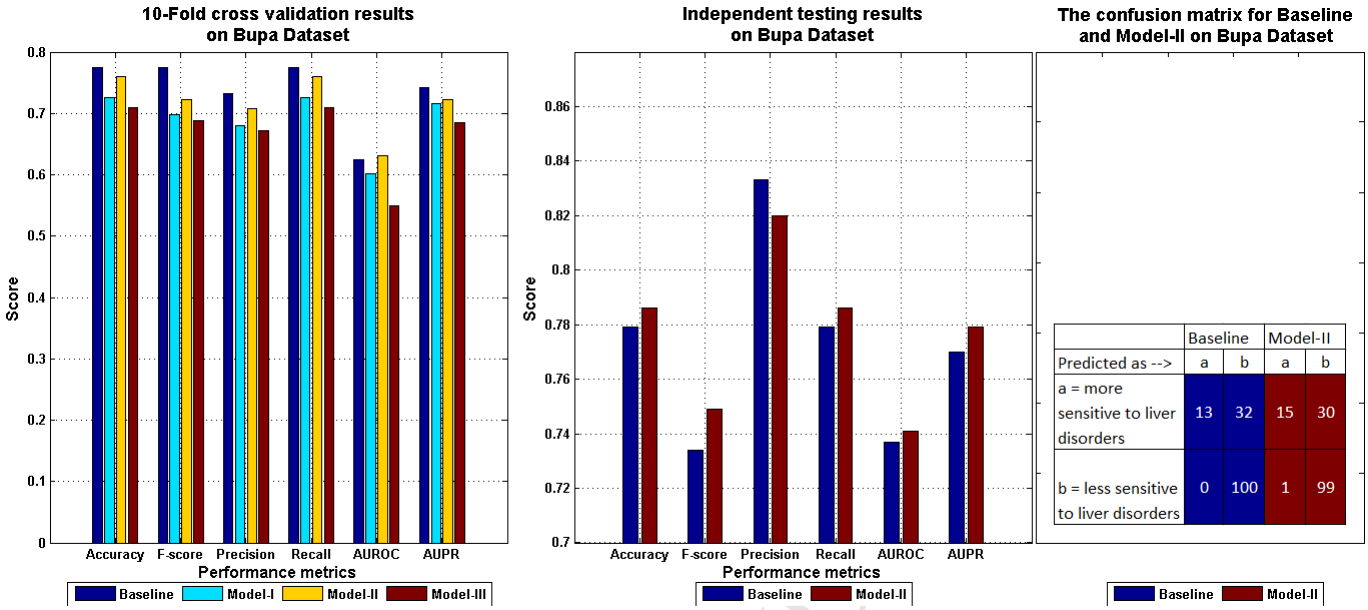


Figure 4: Results of the models on the Bupa Dataset

Figure 4 (left panel) displays the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{II}^{Bp}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{II}^{Bp}$ and compared the results with the baseline ($Model_{Base}^{Bp}$). Figure 4 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.4. Results on Hepatitis Dataset

Based on the results of feature ranking, three subsets of features are finalized and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{Hp}$: Based on the outcome of the ranker, *InfoGainAttributeEval*, Features 1, 3, 7-10, 13, 16 are excluded; the rest of the eleven features are used to construct this model.
- $Model_{II}^{HpHp}$: Considering the ranked features produced by the rankers *InfoGainAttributeEval* and *GainRatioAttributeEval*, Features 1, 8, 9, 16, are excluded while constructing this model.
- $Model_{III}^{HpHp}$: Features 1, 7-10, 13, 16 are excluded based on the outcome of the rankers, *InfoGainAttributeEval* and *GainRatioAttributeEval*; thus the model is constructed using the rest of the features.

Figure 5 (left panel) shows the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model^{Bp}_{III}$ performs better than the other models. As the best

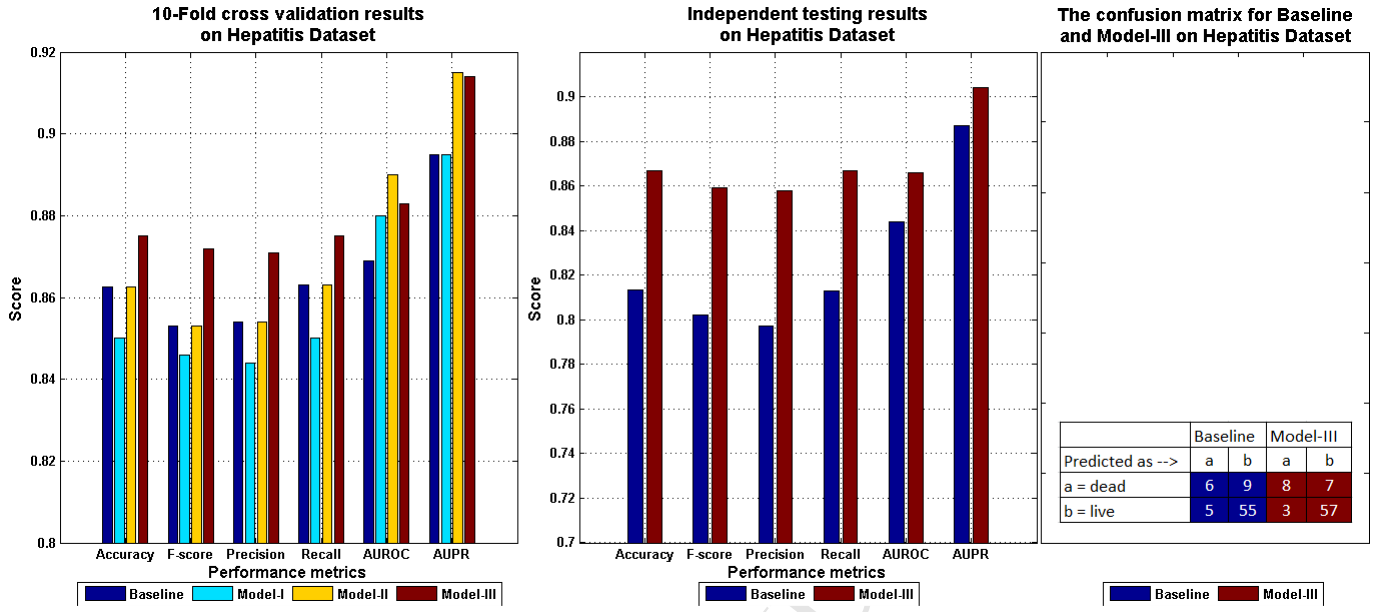


Figure 5: Results of the models on the Hepatitis Dataset

performer,
we conduct independent testing for $Model_{III}^{Hp}$ and compared the results with the baseline ($Model_{Base}^{Hp}$).

Figure 5 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.5. Results on Statlog (Heart) Dataset

Based on the results of feature ranking, three subsets of features are finalized and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{HtS}$: Based on the outcome of the ranker, *InfoGainAttributeEval*, Features 1, 4 are excluded; the rest of the features are used to construct this model.
- $Model_{II}^{HtS}$: Considering the ranked features produced by the ranker *CorrelationAttributeEval*, Features 4, 6 are excluded while constructing this model.
- $Model_{III}^{HtS}$: Feature 5 is excluded based on the outcome of the ranker, *ReliefFAttributeEval*; thus the model is constructed using the rest of the features.

Figure 6 (left panel) demonstrates the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{III}^{HtS}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{III}^{HtS}$ and compared the results with the baseline

($Model_{Base}^{HtS}$).

III Base

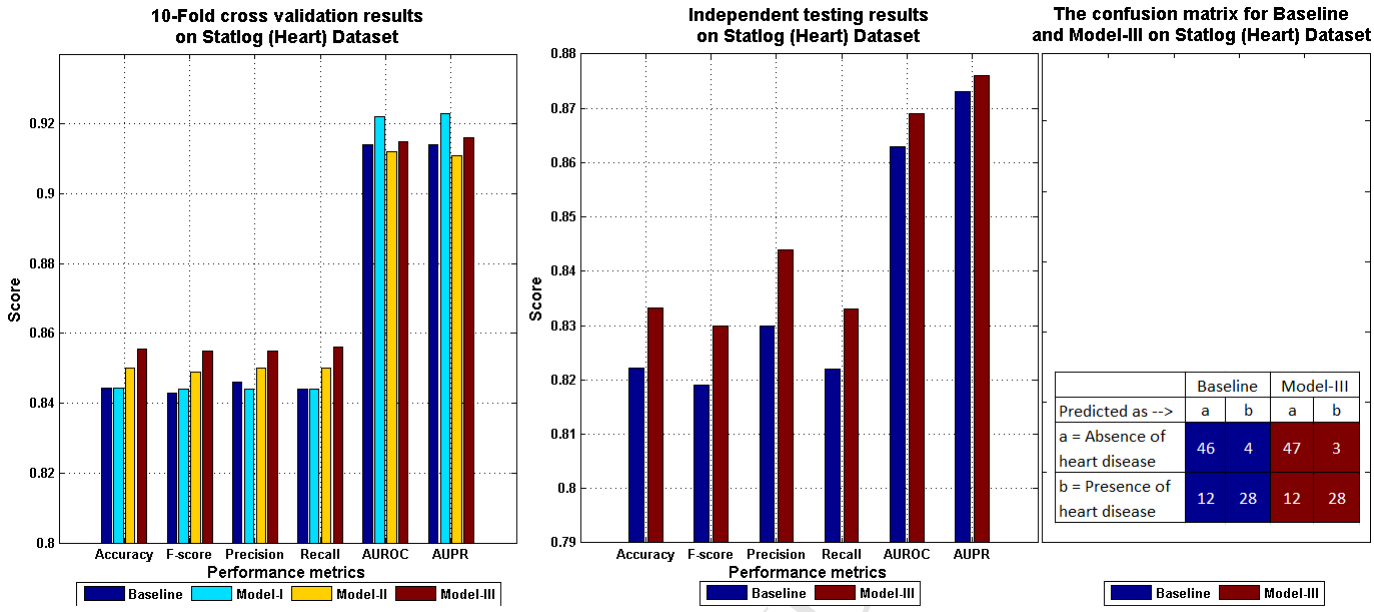


Figure 6: Results of the models on Statlog (Heart) Dataset

Figure 6 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.6. Results on SPECTF dataset

Based on the results of feature ranking three subsets of features are finalized and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{SF}$: Based on the results produced by the ranker, *InfoGainAttributeEval* & *InfoGainAttributeEval*, Features 1, 20, 21, 23, 24, 27, 31, 33, 38 are removed; the rest of the features are used to construct this model.
- $Model_{II}^{SF}$: Considering the results of the rankers, *RandomForest* Features 2, 4, 14, 20, 27, 28, 30, 36, 38, 40, 43 are excluded to construct this model.
- $Model_{III}^{SF}$: Based on the outcome of the ranker, *SVM*, Features 1, 24-33 are excluded and the rest features are used to construct this model.

Figure 7 (left panel) shows the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{I}^{SF}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{I}^{SF}$ and compared the results with the baseline ($Model_{Base}^{SF}$). Figure

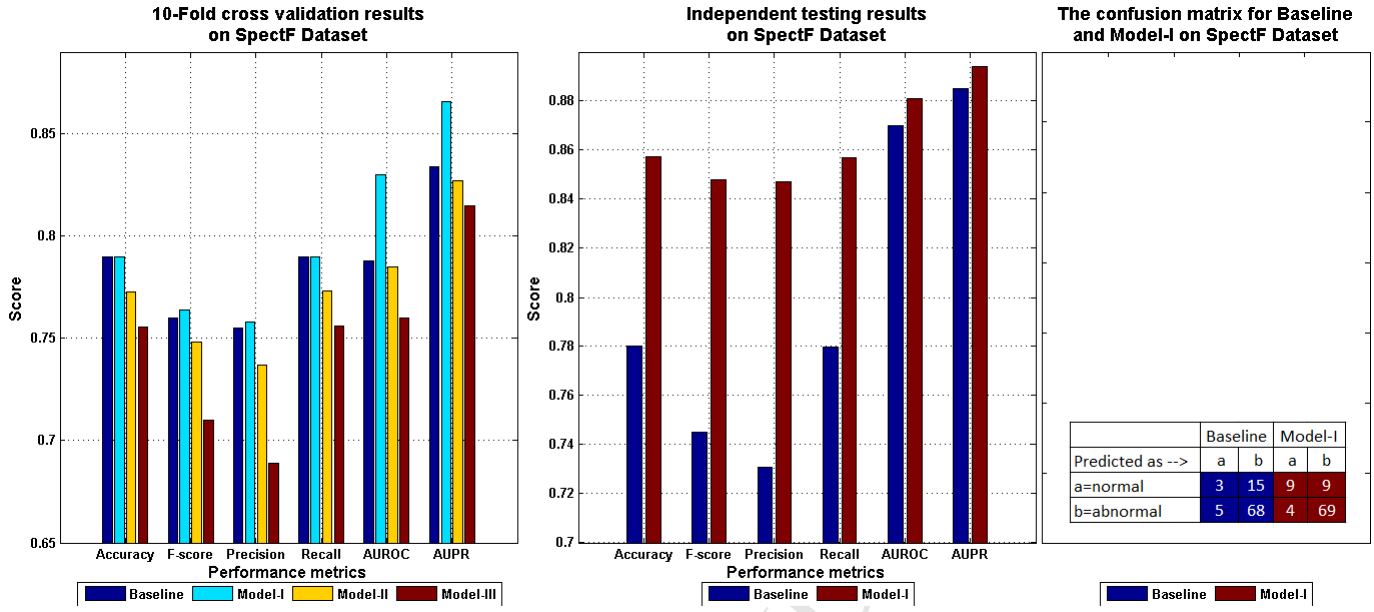


Figure 7: Results of the models on the SPECTF Dataset

7 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.7. Results on SaHeart dataset

Based on the results of feature ranking, three subsets of features are finalized, and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{SHt}$: Based on the results produced by the ranker, *ReliefFAttributeEval* Features 2, 3, 5, 7 are removed; the rest of the features are used to construct this model.
- $Model_{II}^{SHt}$: Considering the results of the rankers, *RandomForest* Features 4, 8 are excluded to construct this model.
- $Model_{III}^{SHt}$: Based on the outcome of the ranker, *SVM*, Features 1, 2, 5, 6, 7 are excluded and the rest are used to construct this model.

Figure 8 (left panel) demonstrates the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{II}^{SHt}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{II}^{SHt}$ and compared the results with the baseline

($Model_{Base}^{SHt}$).

II Base

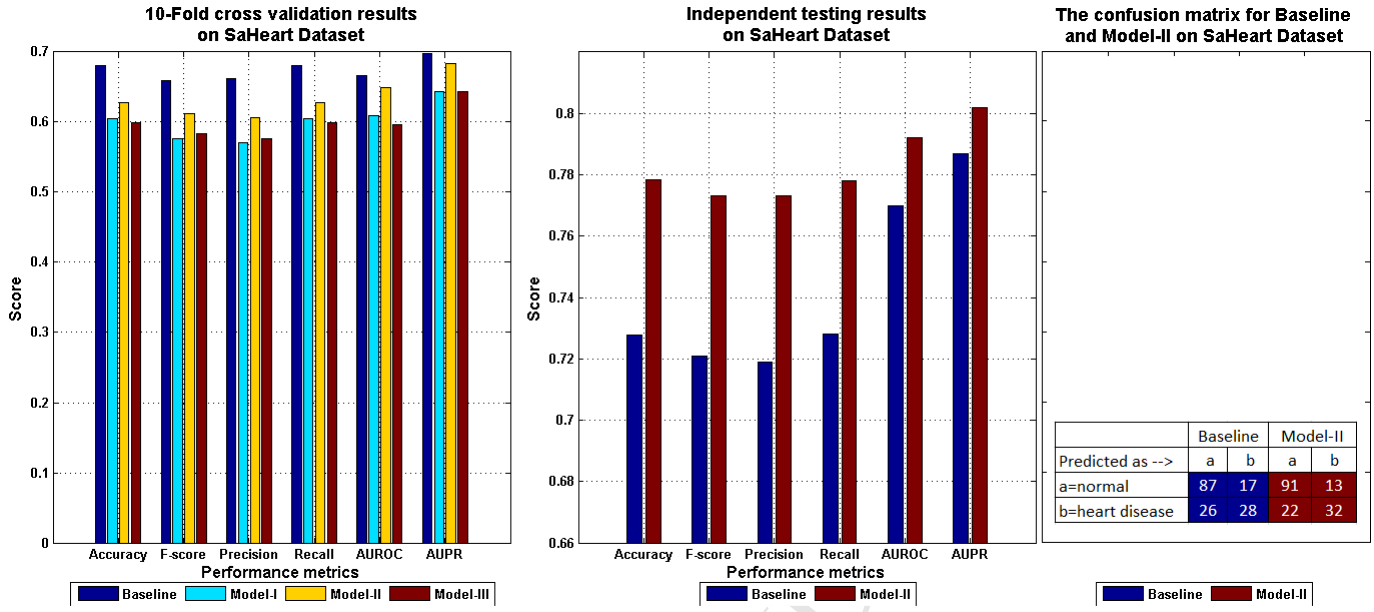


Figure 8: Results of the models on SaHeart Dataset

Figure 8 also displays the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.8. Results on PlanningRelax dataset

Based on the results of feature ranking, three subsets of features are finalized and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{PRx}$: Based on the results produced by the ranker, *CorrelationAttributeEval* Features 1, 2, 5, 6, 10, 11, are removed; the rest of the features are used to construct this model.
- $Model_{II}^{PRx}$: Considering the results of the rankers, *InfoGainAttributeEval*, *GainRatioAttributeEval* & *SVM* Features 1, 8, 9 are excluded to construct this model.
- $Model_{III}^{PRx}$: Based on the outcome of the ranker, *RandomForest*, Feature 6 is excluded and the rest are used to construct this model.

Figure 9 (left panel) displays the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{I}^{PRx}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{I}^{PRx}$ and compared the results with the baseline ($Model_{Base}^{PRx}$).

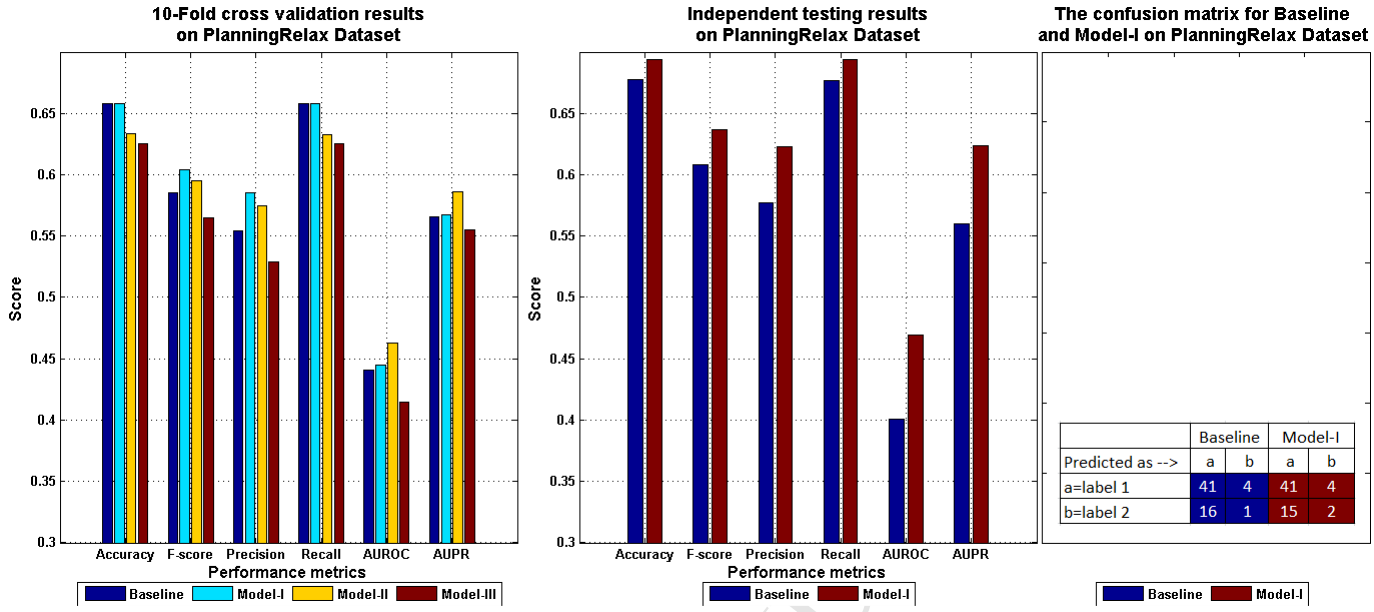


Figure 9: Results of the models on PlanningRelax Dataset

Figure 9 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.9. Results on Parkinsons dataset

Based on the results of feature ranking three subsets of features are finalized, and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{PkS}$: Based on the results produced by the ranker, *InfoGainAttributeEval* Features 17, 18 are removed; the rest of the features are used to construct this model.
- $Model_{II}^{PkS}$: Considering the results of the rankers, *InfoGainAttributeEval* Features 2, 4, 16-18, 20, 21 are excluded to construct this model.
- $Model_{III}^{PkS}$: Based on the outcome of the ranker, *SVM*, Features 1, 11-20 are excluded and the rest are used to construct this model.

Figure 10 (left panel) presents the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{III}^{PkS}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{III}^{PkS}$ and compared the results with the baseline ($Model_{Base}^{PkS}$).

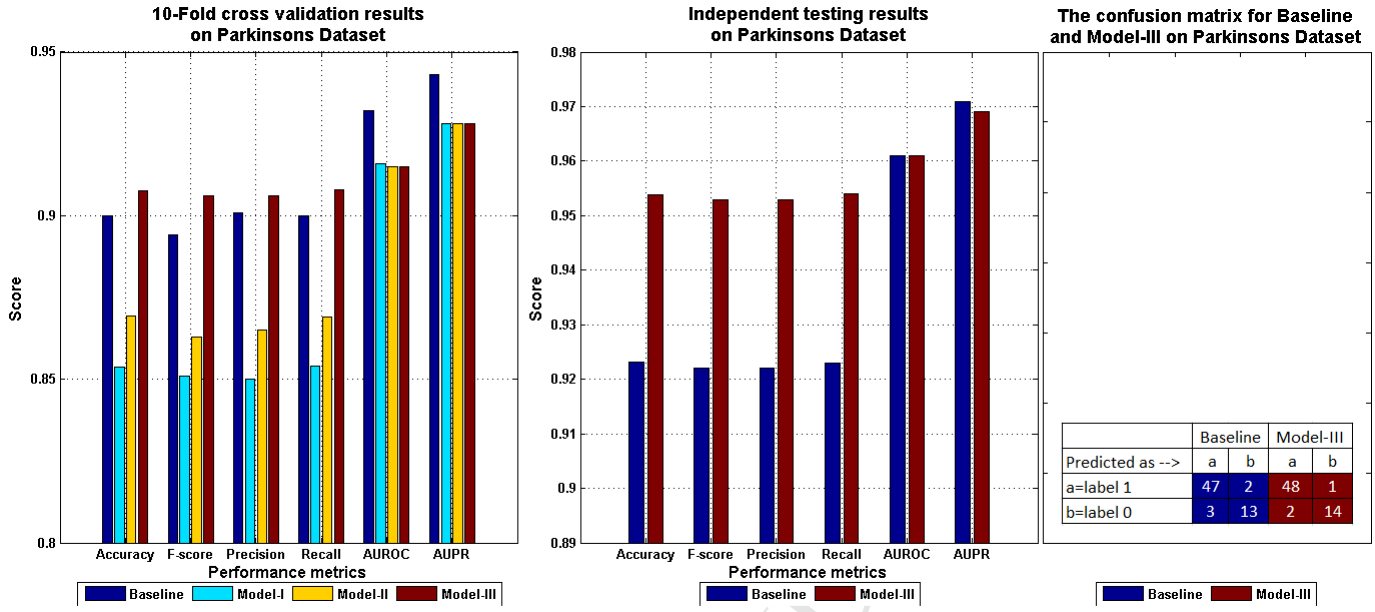


Figure 10: Results of the models on Parkinsons Dataset

Figure 10 also displays the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.10. Results on Hepatocellular Carcinoma (HCC) dataset

Based on the results of feature ranking, three subsets of features are finalized and corresponding models are constructed as follows. The detail ranking results are provided in the supplementary materials.

- $Model_{I}^{HCC}$: Based on the results produced by the ranker, *InfoGainAttributeEval* Features 1, 26, 28-30, 33-35 are removed; rest of the features are used to construct this model.
- $Model_{II}^{HCC}$: Considering the results of the rankers, *RandomForest* Features 24-26, 29, 30, 32, 36, 39, 40, 42, 44, 48 are excluded to construct this model.
- $Model_{III}^{HCC}$: Based on the outcome of the ranker, *SVM*, Features 7, 29, 41, 49 are excluded and the rest are used to construct this model.

Figure 11 (left panel) displays the 10-fold cross validation performance of the above three models along with the baseline model. From these results it is observed that $Model_{I}^{HCC}$ performs better than the other models. As the best performer, we conduct independent testing for $Model_{I}^{HCC}$ and compared the results with the baseline ($Model_{Base}^{HCC}$).

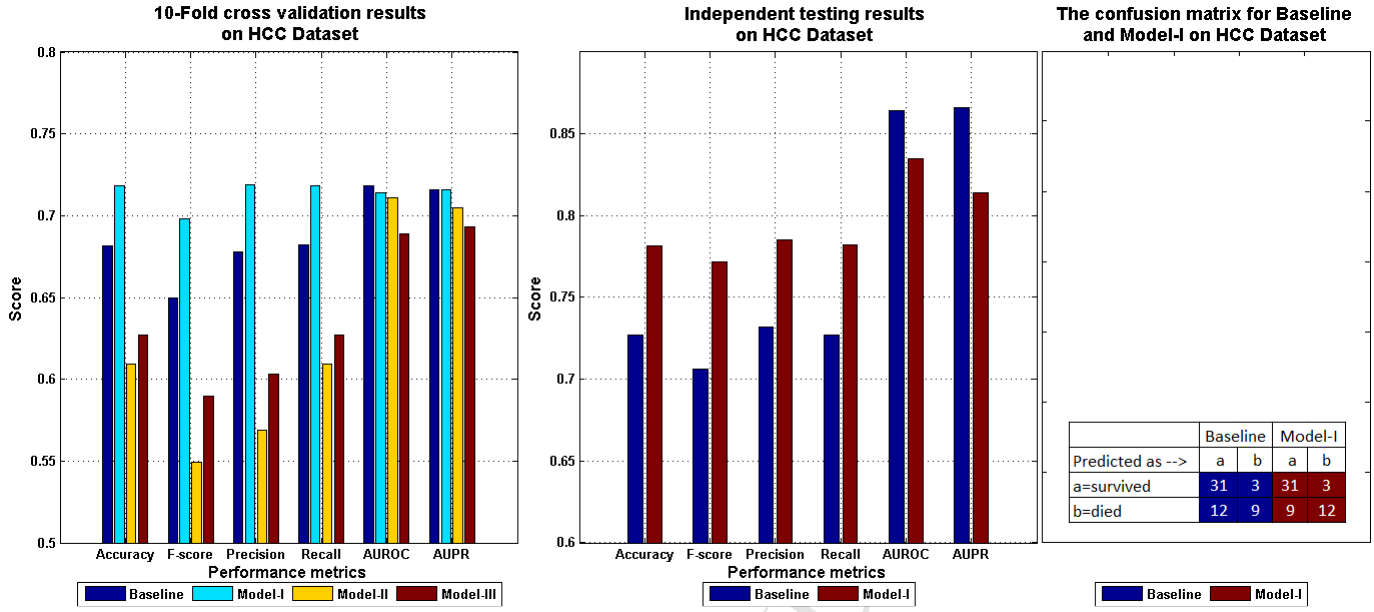


Figure 11: Results of the models on HCC Dataset

Figure 11 also presents the independent testing results (mid panel) and the corresponding confusion matrix (right panel). The contribution of feature ranking and selection is evident from the results.

3.11. Comparison

The feature ranking based models proposed in this paper have achieved better training (cross-validation) accuracy as well as better independent testing accuracy for medical data classification as compared with the baseline (i.e., without feature ranking). Their performance is also quite promising in comparison with the state-of-the-art. Figure 12 presents the testing accuracy and F-score of the best models reported in this paper and the state-of-the-art; results of other methods have been taken from the respective papers and the results that are not available in the literature are left missing in the bar-charts.

Now as we can see, in testing accuracy, our best model outperforms all previous approaches for most of the datasets. In the Breast Cancer dataset, the recent work by Karthik *et al.* [19], which employs a deep neural network (DNN) following a recursive feature elimination step is ahead of us in accuracy, albeit only slightly. Karthik *et al.* also presented the sensitivity (i.e., recall) and specificity (i.e., True Negative Rate = 1 - False Positive Rate) with the help of a bar chart (Fig. 5 of [19]). From the bar chart we see that the sensitivity is slightly less than 0.98 whereas our best model's sensitivity (i.e., recall) is 0.985 (cf. Figure 2). The specificity of DNN [19] is slightly better than that of our best model. Overall, DNN [19] performs slightly better than ours but notably, the approach of [19] isonly

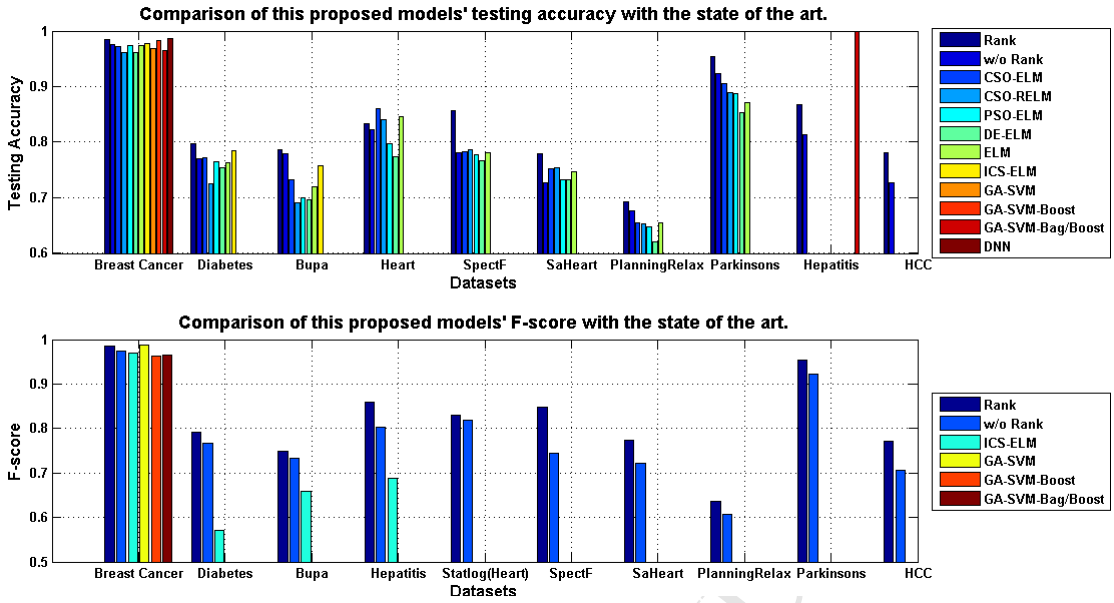


Figure 12: Comparison of this proposed models' testing accuracy and Fscore with the state-of-the-art. w/o Rank means the baseline model for that disease and Rank means the best model constructed after feature selection and ranking. We do not report base ELM results from the study [2] as the work [31] is very recent.

designed for the Breast Cancer dataset.

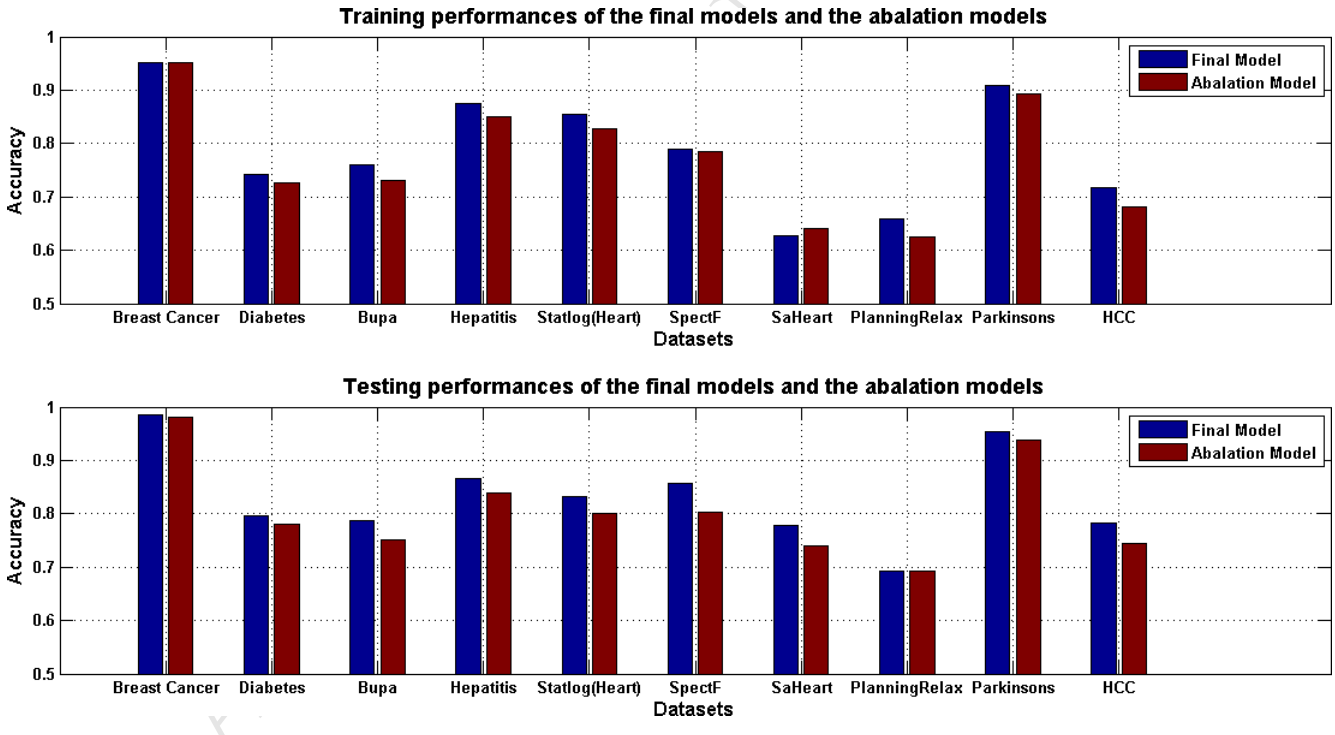
In the Hepatitis dataset, our testing accuracy is only second to the claimed 100% accuracy of ICSELM [2]. However, this performance should be checked with the F-scores presented in Figure 12: the F-score of ICSELM is quite low (0.6875) as compared to our best model (0.859), suggesting that our model is more robust than ICSELM.

In all other datasets, considering the accuracy, our model outperforms CSO-ELM [31] (and also CSO-RELM) excepting in the Heart dataset; for this dataset the accuracy reported for CSO-ELM and CSO-RELM in [31] is 0.8612 and 0.8402 respectively. Our best model is not far behind having an accuracy of 0.8333 with an F-score of 0.830 suggesting the robustness of our predictor. Unfortunately, Eshtay et al. [31] did not report any other metric nor did they provide the confusion matrix for further evaluation/comparison. On another note, both CSO-ELM and CSO-RELM exploit a metaheuristics technique called Competitive Swarm Optimization (CSO). Due to the inherent stochastic nature of CSO, Eshtay *et al.* [31] conducted 30 independent runs. According to their report, the worst case accuracy of CSO-ELM across these 30 runs is 0.8152 which is lower than our accuracy; for CSO-RELM this is even worse, 0.7826. With respect to the ELM based works of [2, 31], we have some more observations as discussed in a later section (Section 3.13).

3.12. Ablation Study

Figure 13: The results of Ablation study on the best-performing model of each dataset

We have conducted an *Ablation study* on the best-performing model for each dataset. In particular, for each disease, we have removed the lowest ranked feature from the best-performing model which leads to the formation of another model, called, the *ablation model*. For example, $Model_I^{HCC}$ is the best-performing model for HCC dataset with Feature 32 being the lowest ranked feature thereof. Now, through removing Feature 32 we get the ablation model. Then the performances of the best-performing model and the corresponding ablation model are compared with each other. Figure 13 presents the training and testing accuracies of the best-performing model and the corresponding ablation model for each of the datasets.



From these results presented in Figure 13, it is evident that when the least-ranked feature is removed from the best-performing model, the performance degrades. There is only one exception to this finding and that is for the SaHeart dataset, albeit only in the context of training accuracy; for testing accuracy we do find that the ablation model performs worse. Thus we can be confident that all selected features (through feature ranking and selection) are indeed

contributing towards the robustness of the predictor for each disease.

3.13. Discussions

We have proposed a general methodology (Figure 14) for medical data classification that employs a feature ranking and selection strategy followed by model training and construction using a suitable classifier algorithm. We believe that this general methodology would be useful for any medical dataset for prediction/diagnosis tasks. To judge the

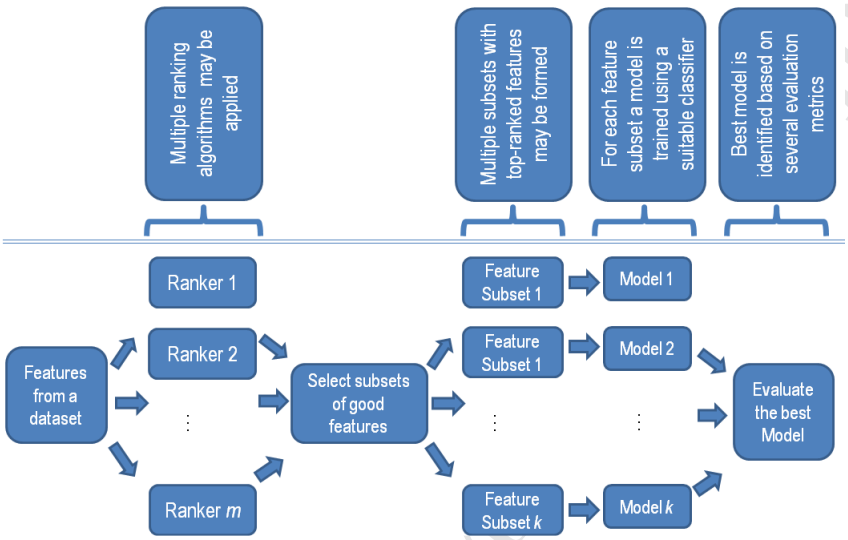


Figure 14: General Methodology

efficacy, robustness and generality of our approach, we have experimented on 10 medical datasets using a number of feature ranking strategies, and applied the Random Forest algorithm as the final classifier for our predictors. Our experiments have clearly suggested that feature ranking and selection is useful.

We have also conducted an ablation study on each dataset to verify the importance of the features selected for the model and to confirm the positive contribution thereof on the classification task. Section 3.12 has demonstrated these results. These findings further strengthen our claim that our general methodology of feature ranking and selection does play a strong role in a robust model construction.

As a by-product, we also have a suggestion concerning the feature importance in each dataset (from classification point of view). For example, for the Diabetes dataset, our experiments suggest that Features 3 and 4 have less contribution/importance in the context of disease prediction. This is interesting, as Feature 3 is 'diastolic blood pressure

(mm Hg)' and Feature 4 is 'triceps skin fold thickness (mm)'. On the other hand, across all ranking algorithms, Feature 2 ('plasma glucose concentration a 2 hours in an oral glucose tolerance test') has been ranked as the most important feature, which is in accord from a medical perspective. As another example, for the Heart Disease dataset,

all rankers have ranked Feature 12 as the most important feature; Feature 12 represents the ‘number of major vessels colored by fluoroscopy’ and hence is indeed a very important feature from clinical point of view.

From another angle, as has been briefly indicated before, recent works of [31, 2] have focused on experimentation by varying the number of hidden neurons (of the ELM). However, those studies remain inconclusive, as sometimes the testing accuracy was found better with a higher number of hidden neurons, and sometimes with a lower number. For example, for the Breast Cancer dataset, the testing accuracy is higher for lower number of hidden neurons- 97% for 10 hidden neurons and 83.87% for 300- and for the Hepatitis dataset, the relation is reversed: 86.12% for 10 and 100% for 300 [2]. Similar issues are also observed in [31]. Hence, for a new medical dataset, it is difficult to make even an educated guess as to what should be the ideal number of neurons to obtain best results. On the other hand, from our experiments it is evident that the feature ranking and selection strategy provides better results than the baseline, and thus we present a general methodology that is expected to perform well across all medical datasets (which have similar characteristics in the data pattern) consistently.

We have also conducted some experiments using other classifier algorithms (in place of Random Forest). It has been observed that unlike Random Forest, other classifiers do not perform consistently well across all datasets. For example, the testing accuracies of SVM and Bayes Net on Breast Cancer Dataset are 0.99 and 0.975 respectively. Thus SVM actually outperforms Random Forest (accuracy for our best model, $Model_{HI}^{WBC}$, is 0.9850) on Breast Cancer Dataset. But for the Bupa dataset both SVM and Bayes Net fall significantly short with testing accuracies of only 0.689655 and 0.689655 respectively (the accuracy of our best model, $Model_{HI}^{Bp}$ is 0.786207).

4. Conclusion

Medical data classification is one of the complex and challenging tasks in medical informatics. Due to its complex nature, various methods have been proposed in the literature. In this paper, we have revisited this challenge and have made an effort to present a generalized methodology for this classification task. In particular, we have proposed a feature ranking based methodology that employs a feature selection strategy, and train and construct the model based on only the highly ranked features. To elaborate, we first check whether all features are important for the classification task. This is done by applying several feature ranking algorithms through 10-fold cross validation. Then based on the ranking, a subset of the top-ranked features is formed based on 10-fold cross validation performance. Finally, the Random Forest algorithm is applied on the selected features to train and construct the final model. In fact we form several subsets of top-ranked features and corresponding to each subset, train a model. Thus we have several models to compare and choose from each other.

We have conducted extensive experiments on 10 benchmark datasets and our results are promising. Our feature ranking and selection based models have performed consistently better than the baseline (without feature ranking) model. Our best models are also found to be competitive with the state-of-the-art. We have also done limited

ablation study to verify the importance of the features selected for the model, and to confirm the positive contribution thereof on the classification task. To conclude, we not only have developed highly accurate predictors for 10 different diseases, but also have presented a general methodology that should perform well for other diseases which have similar characteristics in the data pattern.

While the use of feature ranking and selection strategy is not new in the applied machine learning literature, to the best of our knowledge, this had not been comprehensively investigated in the context of Medical Data Classification before the current research work. From this angle, this can be seen as the first attempt where the feature ranking and selection scheme has been applied in the context of Medical Data Classification using the Random Forest as the final classifier. Also, the ablation study conducted here is new in this application domain. Thus, while this paper does not provide any new theoretical advancement in the context of machine learning, we believe that our rigorous experiments along with an appropriate ablation study have advanced the state-of-the-art. In fact, through extensive experiments on 10 different benchmark datasets, we have been able to show that our approach is indeed useful and sufficiently general. It is also worth mentioning that the top-ranked features have been found to be important from a medical point of view. Therefore, while one may argue that feature ranking and selection should always lead to a better predictor/classifier, in our case, we have the added value of explaining the phenomenon in real (clinical) context. Thus we expect that a medical practitioner would be more confident (and comfortable, so to speak) in using our predictor.

Although not reported here, we have also conducted extensive experiments with other classifiers, such as: Support Vector Machine (SVM), Bayes Network, Multilayer Perceptron, etc. and found the better contribution of Random Forest across all datasets as compared to other classifiers. The use of 10 datasets is also a strong feature of this research as only one study [31] in the literature, so far as we know, has used all of these datasets for experimentation, albeit with a different goal: they have worked on the evolutionary Extreme Learning Machine (ELM) model, focusing on compacting networks by reducing the number of neurons in the hidden layer. This is clearly in contrast with our focus of finding a general methodology for medical data classification. Notably, their study remained inconclusive as sometimes the testing accuracy was found better with higher number of hidden neurons and sometimes with lower number. On the other hand, we present a general methodology that is expected to perform well across all medical datasets (which have similar characteristics in the data pattern) consistently.

Supplementary materials:

The supplementary material.pdf: This document describes following things:

- Detailed description of the datasets

- Rankers
- Evaluation Metrics
- Feature name mapping with its number and feature ranking results for each dataset

Acknowledgment

This work was funded by ICT Division, Government of the People's Republic of Bangladesh, website: http://www.ictd.gov.bd/

[1] F. Chabat, D. M. Hansell, G.-Z. Yang, Computerized decision support in medical imaging, IEEE Engineering in Medicine and Biology Magazine 19 (5) (2000) 89–96.

[2] P. Mohapatra, S. Chakravarty, P. Dash, An improved cuckoo search based extreme learning machine for medical data classification, Swarm and Evolutionary Computation 24 (2015) 25–49. doi:10.1016/j.swevo.2015.05.003.
URL https://www.sciencedirect.com/science/article/pii/S2210650215000413

[3] R. O. Duda, P. E. Hart, Pattern Classification and Scene Analysis, A Wiley-Interscience Publication, John Wiley & Sons, NY, USA, 1973.

[4] D. Coast, R. Stern, G. Cano, S. Briller, An approach to cardiac arrhythmia analysis using hidden markov models, IEEE Trans Biomed Eng. 37 (9) (1990) 826–36. doi:10.1109/10.58593.
URL https://www.sciencedirect.com/science/article/pii/0031320394900310

[5] H. A. Abbass, An evolutionary artificial neural networks approach for breast cancer diagnosis, Artificial Intelligence in Medicine 25 (3) (2002) 265281.
URL https://www.sciencedirect.com/science/article/pii/S001048251300303X

[6] T. Kiyan, T. Yildirim, Breast cancer diagnosis using statistical neural networks, Journal of Electrical and Electronics Engineering 4 (2) (2004) 11491153.

[7] M. Karabatak, M. C. Ince, An expert system for detection of breast cancer based on association rules and neural network, Expert Systems with Applications 36 (2) (2009) 34653469. doi:doi.org/10.1016/j.eswa.2008.02.064.
URL https://www.sciencedirect.com/science/article/pii/S0957417408001103

[8] Y. Peng, Z. Wu, J. Jiang, A novel feature selection approach for biomedical data classificatio, Journal of Biomedical Informatics 43 (2010) 1523. doi:doi.org/10.1016/j.jbi.2009.07.008.
URL https://www.sciencedirect.com/science/article/pii/S1532046409001014

[9] C.-Y. Fana, P.-C. Changb, J.-J. Linb, J. Hsiehb, A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification, Applied Soft Computing 24 (2011) 632644. doi:doi.org/10.1016/j.asoc.2009.12.023.
URL https://www.sciencedirect.com/science/article/pii/S1568494609002774

[10] A. T. Azar, S. A. El-Said, Performance analysis of support vector machines classifiers in breast cancer mammography recognition, Neural Computing and Applications 4 (5) (2014) 11631177. doi:10.1007/s00521\-012\-1324\-4.
URL https://link.springer.com/article/10.1007/s00521\discretionary{-}{}{}012\discretionary{-}{}{}1324\discretionary{-}{}{}4

[11] A. Bhattacherjee, S. Roy, S. Paul, P. Roy, N. Kausar, N. Kausar, Classification approach for breast cancer detection using back propagation neural network: a study, Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes (2015) 12doi:10.4018/978-1-4666-8811-7.ch010.

[12] A. Onan, A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer, Expert Systems with Applications 42 (20) (2015) 68446852. doi:doi.org/10.1016/j.eswa.2015.05.006.
URL https://www.sciencedirect.com/science/article/pii/S0957417415003267

[13] Z. Yin, Z. Fei, C. Yang, A. Chen, A novel svm-rfe based biomedical data processing approach: Basic and beyond, in: IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2016, p. 71437148. doi:10.1109/IECON.2016.7793954.
URL https://ieeexplore.ieee.org/document/7793954

[14] S. Jhajharia, H. K. Varshney, S. Verma, R. Kumar, A neural network based breast cancer prognosis model with pca processed features, in: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2016, p. 18961901. doi:10.1109/ICACCI.2016.7732327.
URL https://ieeexplore.ieee.org/document/7732327

[15] H. Jouni, M. Issa, A. Harb, G. Jacquemod, Y. Leduc, Neural network architecture for breast cancer detection and classification, in: IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), IEEE, 2016, p. 3741. doi:10.1109/IMCET.2016.7777423.
URL https://ieeexplore.ieee.org/document/7777423/

[16] A. M. Abdel-Zaher, A. M. Eldeib, Breast cancer classification using deep belief networks, Expert Systems with Applications 46 (2016) 139144. doi:doi.org/10.1016/j.eswa.2015.10.015.
URL https://www.sciencedirect.com/science/article/pii/S0957417415007101

[17] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, C.-F. Tsai, Svm and svm ensembles in breast cancer prediction, PloS one 12 (1) (2017) e0161501.

[18] M. Nilashi, A. P. D. O. Ibrahim, H. Ahmadi, L. Shahmoradi, A knowledge-based system for breast cancer classification using fuzzy logic method, Telematics and Informatics 34 (4). doi:10.1016/j.tele.2017.01.007.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161501

[19] S. Karthik, R. S. Perumal, P. V. S. S. R. C. Mouli, Breast cancer classification using deep neural networks, Knowledge Computing and Its Applications (2018) 227–241doi:10.1007/978\-981\-10\-6680\-1_12.
URL https://link.springer.com/chapter/10.1007/978\discretionary{-}{}{}981\discretionary{-}{}{}10\discretionary{-}{}{}6680\discretionary{-}{}{}1_12

[20] M. M. Khan, A. Mendes, S. K. Chalup, Evolutionary wavelet neural network ensembles for breast cancer and parkinson's disease prediction, PLoS ONE 13 (2) (2018) e0192192. doi:10.1371/journal.pone.0192192.
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0192192

[21] P. K. Anooj, Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules, Journal of King Saud University-Computer and Information Sciences 11 (1) (2012) 2740. doi:doi.org/10.1016/j.jksuci.2011.09.002.
URL https://www.sciencedirect.com/science/article/pii/S1319157811000346

[22] M. L. Samb, F. Camara, S. Ndiaye, Y. Slimani, M. A. Esseghir, A novel rfe-svm-based feature selection approach for classification, International Journal of Advanced Science and Technology 43. doi:10.1.1.641.826.
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.641.8266

[23] P. Jaganathan, R. Kuppuchamy, A threshold fuzzy entropy based feature selection for medical database classification, Computers in Biology and Medicine 43 (12) (2013) 27–38. doi:10.1016/j.compbiomed.2013.10.016.
URL https://www.sciencedirect.com/science/article/pii/S001048251300303X

[24] K. Polat, S. Gnes, A new feature selection method on classification of medical datasets: Kernel f-score feature selection, Expert Systems with Applications 36 (2009) 1036710373.

[25] M. A. Jabbar, B. Deekshatulu, P. Chandra, Classification of heart disease using k- nearest neighbor and genetic algorithm, Elsevier Procedia Technology 10 (2013) 85–94. doi:doi.org/10.1016/j.protcy.2013.12.340.
URL https://www.sciencedirect.com/science/article/pii/S2212017313004945

[26] F. Gorunescu, S. Belciug, Evolutionary strategy to develop learning-based decision systems. application to breast cancer and liver fibrosis stadialization, Journal of Biomedical Informatics 49 (2014) 112118. doi:doi.org/10.1016/j.jbi.2014.02.001.
URL https://www.sciencedirect.com/science/article/pii/S1532046414000173

[27] S. Khanmohammadi, M. Rezaeiahari, Ahp based classification algorithm selection for clinical decision support development, Elsevier Procedia Computer Science 36 (2014) 328–334. doi:doi.org/10.1016/j.procs.2014.09.101.
URL https://www.sciencedirect.com/science/article/pii/S1877050914013507

[28] B. Dennis, S. Muthukrishnan, Agfs: Adaptive genetic fuzzy system for medical data classification, Applied Soft computing 24 (2014) 242–252. doi:doi.org/10.1016/j.asoc.2014.09.032.
URL https://www.sciencedirect.com/science/article/pii/S1568494614004852

[29] M. Seera, C. P. Lim, A hybrid intelligent system for medical data classification, Expert Systems with Applications 41 (2014) 22392249. doi:10.1016/j.eswa.2013.09.022.
URL https://www.sciencedirect.com/science/article/pii/S0957417413007562

[30] J. Alwidian, B. H. Hammo, N. Obeid, Wcba: Weighted classification based on association rules algorithm for breast cancer disease, Applied Soft Computing 62 (2018) 536549. doi:10.1016/j.asoc.2017.11.013.
URL https://www.sciencedirect.com/science/article/pii/S1568494617306762

[31] M. Eshtay, H. Faris, N. Obeid, Improving extreme learning machine by competitive swarm optimization and its application for medical diagnosis problems, Expert Systems With Applications 104 (2018) 134152. doi:10.1016/j.eswa.2018.03.024.
URL https://www.sciencedirect.com/science/article/pii/S0957417418301696

[32] University of california at irvine (uci) machine learning repository, https://archive.ics.uci.edu/ml/datasets.html.

[33] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Analysis 1 (1-4) (1997) 131–156. doi:10.1016/S1088-467X(97)00008-5.
URL http://www.lsi.us.es/~riquelme/publicaciones/kes03.pdf

[34] R. Ruiz, J. C. Riquelme, J. S. Aguilar-Ruiz, Fast feature ranking algorithm, V. Palade, R.J. Howlett, and L.C. Jain (Eds.) KES 2003 (LNAI 2773) (2003) 325331.
URL http://www.lsi.us.es/~riquelme/publicaciones/kes03.pdf

[35] J. NOVAKOVI, P. STRBAC, D. BULATOVI, Toward optimal feature selection using ranking methods and classification algorithms, Yugoslav Journal of Operations Research (21). doi:10.2298/YJOR1101119N.
URL http://elib.mi.sanu.ac.rs/files/journals/yjor/41/yujorn41p119\discretionary{-}{}{}135.pdf

[36] Waikato environment for knowledge analysis (weka), https://www.cs.waikato.ac.nz/ml/weka/.

The first author is supported by an ICT Fellowship. No external funding are available for this research.

Ethical Statement:

A. No financial and personal relationships with other people or organizations exists that could inappropriately influence (bias) this work.
B. The authors confirm that this manuscript reports original research works of the authors and has not been published or submitted for consideration elsewhere.