JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 791-803 (2022)
DOI: 10.6688/JISE.202207_38(4).0006

Univariate and Multivariate Filter Feature Selection for Heart Disease Classification

HOUDA BENHAR[1], MOHAMED HOSNI[1,2] AND ALI IDRI[1,3,+]
[1]*Software Project Management Research Team, ENSIAS*
Mohammed V University in Rabat
Rabat, 10100 Morocco
[2]*ENSAM-Meknes*
Moulay ISMAIL University
Meknes, 50050 Morocco
[3]*MSDA, Mohammed VI Polytechnic University*
Benguerir, 43150 Morocco
E-mail: {houda_benhar; ali.idri}@um5.ac.ma; hosni.mohamed1@gmail.com

Feature selection (FS) is a data preprocessing task that can be applied before the classification phase, and aims at improving the performance and interpretability of classifiers by finding only a few highly informative features. The present study aims at evaluating and comparing the performances of six univariate and two multivariate filter FS techniques for heart disease classification. The FS techniques were evaluated with two white-box and two black-box classification techniques using five heart disease datasets. Furthermore, this study deals with the setting of the hyperparameters' values of the four classifiers. This study evaluates 600 variants of classifiers. Results show that white-box classification techniques such as K-Nearest Neighbors and Decision Trees can be very competitive with black-box ones when hyperparameters' optimization and feature selection were applied.

Keywords: classification, filters, feature selection, data preprocessing, heart disease

1. INTRODUCTION

Heart disease (HD) is one of the most prevalent diseases and considered among the leading causes of death worldwide. It is therefore, considered as one of the main priorities in medical informatics research [1]. Therefore, data mining (DM) techniques have been used to extract useful predictive and descriptive knowledge from large HD datasets [2].

Feature selection is a data preprocessing task that aims to improve the performance of DM-based decision support systems for HD classification [1]. FS algorithms generally fall into four categories: filters, wrappers, embedded, and hybrid models. Filters, in contrast to embedded and wrapper techniques, select features without optimizing the performance of a DM technique [3]. Hybrid models are mainly based on the combination of the three aforementioned types. These techniques can be either univariate or multivariate [4]. Univariate techniques, also known as feature rankers, consist of ranking features individually based on some performance measures and the final features subset can be determined by setting a cutoff threshold or specify how many features to retain; while multivariate techniques evaluate an entire feature subset based on a specific search strategy using some performance measures and select the best features subset.

According to authors' knowledge, no work has evaluated and compared univariate

Received March 26, 2021; revised July 5, 2021; accepted October 15, 2021.
Communicated by Tzung-Pei Hong.
[+] Corresponding author.

and multivariate filter FS techniques using different classifiers over multiple heart disease datasets. Moreover, most of the existing studies have focused on enhancing the classification accuracy and have neglected the interpretability issues of the classifiers used which were in general black-box models [3].

The present study has a twofold objective: (1) evaluate and compare the impacts of six univariate filters: ReliefF (RF), Linear Correlation (LC), Info Gain (IG), Signal-to-noise ratio (SN), minimum Redundancy Maximum Relevance (MR), and t-test (TT), and two multivariate filters: Correlation-based feature subset selection and Consistency-based subset selection on the performance of two white-box classifiers: K-Nearest Neighbors (KNN) and Decision Trees (DTs); and two black-box classifiers: Support Vector Machines (SVM) and Multilayer Perceptron (MLP) for heart disease diagnosis, and (2) investigate to what extent hyperparameters' (HP) tuning affects the performance of heart disease classification. This study uses the grid search (GS) optimization technique (OT) to tune the parameters of the four classifiers investigated. Note that classifiers based on the same parameters over all data subsets using Weka uniform configuration (Weka-UC) were also used in order to conduct a comparison.

The experiments were performed using the Weka 3.8.3 tool [5]. The classifiers were evaluated using a 10-fold cross validation method and three performance criteria: accuracy, kappa statistic, and area under the ROC curve (AUC). Overall, this study evaluates 600 variants of classifiers: 600 = (4 classifiers) * (6 univariate-filters * 2 selection-thresholds + 2 multivariate-filters + original features set) * (5 datasets) * (2 optimization techniques), and aims at addressing the following research questions: **RQ1:** Does optimizing the parameters of classification techniques improve their performance regardless of the feature selection technique used? **RQ2:** Do multivariate filters outperform univariate ones when used for heart disease classification? **RQ3:** Is there any feature selection technique that distinctly outperforms the others? And **RQ4:** Is there any combinations of OT, FS and classifier that outperform others?

The remainder of this paper is organized as follows: Section 2 describes the experimental design followed in this study as well as the datasets used. Results are presented and discussed in Sections 3 and 4 respectively. Finally, the conclusions and future works are presented in Section 5.

2. MATERIAL AND METHODS

2.1 Datasets

The present study used five datasets related to heart disease: Statlog Heart data (ST), Heart failure dataset (HF), processed Cleveland Heart Disease (CV), unprocessed Cleveland Heart Disease (ORGCV), and Arrhythmia (ARR) datasets, which are available from the UCI Machine Learning Repository [6]. In this paper, we aim to simply distinguish between the absence and presence of a heart disease. Therefore, all class values indicating the presence of heart disease in the CV, ORGCV, and ARR datasets were replaced by 1 while class 0 indicates the absence of heart disease. ST, HF, CV and ORGCV are considered as small datasets, while ARR is considered as a medium-sized dataset.

2.2 Methodology

Performances of the four classifiers were evaluated using a 10-fold cross validation strategy [7]. The methodology applied is as follows:

Step 1: Each dataset is checked for missing values and unimportant features.
Step 2: Feature selection techniques are applied for each dataset:

- Multivariate techniques (CFS and CON) return a feature subset. In total, we have 10 feature subsets;
- Univariate techniques (RF, LC, IG, SN, MR, and TT) return a list of ranked features. Based on a previous study [8], the thresholds 40% and 50% were used with the univariate techniques to select the final feature subsets. In total 60 features subsets were obtained.

Step3: For each dataset, the original feature set as well as each selected feature subset of Step 2 were investigated with KNN, SVM, MLP and DT using GS optimization and Weka-UC. The predefined search spaces of GS for each classification technique are listed in Table 1. In total, we obtain 120 variants of the four classifiers per dataset.

Table 1. Search spaces of HPs values used by Grid Search for the four classification techniques.

Classifiers	HPs
KNN	K = [1, 20] with an increment of 1
SVM	Kernel = {RBF, Poly}; C = [1, 200] with an increment of 5; Exponent Poly = [1, 5] with an increment of 1; Gamma RBF= [0.001, 0.1] with an increment of 0.001
MLP	Hidden layers = [1, 16] with an increment of 1; Learning rate = [0.01, 1] with an increment of 0.01; Momentum = [0.1, 1] with an increment of 0.1; Epochs = [100, 2000] with an increment of 100
DT	Leaf = [1, 20] with an increment of 1; Confidence = [0.01, 0.7] with an increment of 0.01

Step 4: For each dataset, cluster the constructed techniques of Step 3 using the Scott-Knott (SK) statistical based on the kappa score. The SK algorithm is a hierarchical cluster analysis approach used to partition treatments into distinct groups. The SK test deals with multiple comparisons problems
Step 5: Rank the classifiers belonging to the best SK cluster by means of Borda Count voting system based on accuracy, kappa and AUC scores to gain more insight into the results.

In order to simplify the naming of the constructed models, we use the following abbreviations:

- The first number of the threshold 40% or 50% is used along with the abbreviation of the feature rankers to indicate a selected feature subset. For instance, RF4 describes the subset of ReliefF with the threshold 40%.
- ORG describes the original feature set for each dataset.
- G denotes the Grid search optimization while U denotes Weka-UC.

The constructed models are then abbreviated as follows:

- **OT-ClassifierFeatureSubset**

For instance, U-MLPLC4 refers to MLP classifier with the uniform HP configuration of Weka applied to the feature subset selected by means of Linear Correlation and the threshold 40%.

3. RESULTS

The empirical results are depicted in this section. A software prototype with Java programming language and Weka API was developed to carry out the experiments. FS techniques not available in Weka were performed using Python's Scikit-learn library [9]. The Scott-Knott statistical test was performed using R Software. Features with high percentages of missing values or containing same values over all instances were removed from the datasets. Thereafter, instances with missing values were removed.

3.1 Feature Selection Results

Applying the univariate filters over ST and CV datasets resulted in the selection of 5 and 6 attributes with 40% and 50% thresholds respectively. For ORGCV, the thresholds 40% and 50% selected 14 and 18 features respectively. A total of 111 and 139 features were selected with the thresholds 40% and 50% respectively in ARR dataset. For HF dataset the 40% and 50% thresholds selected 4 and 6 features respectively. As for multivariate techniques, CFS selected a total of 7 features in both ST and CV dataset, 15 attributes in ORGCV, 38 attributes in ARR, and 4 attributes in HF dataset. Also, a total of 11 features in ST and CV, 6 in ORGCV, 21 in ARR, and 10 in HF dataset were selected using CON.

3.2 Classification Results

The results of the 120 variants of the four classifiers in terms of kappa were compared with SK test for each dataset.

The SK test for ST dataset identified 4 clusters. Of 120 variants, a total of 99 variants belong to the best SK cluster. The majority of the best cluster variants (55 classifiers) were trained using GS optimization. It is to be noted that, with exception of G-MLPSN4, U-MLPSN4, G-MLPSN5, G-SVMSN4, and U-SVMSN4, all MLP and SVM-based variants were present in the best cluster. Moreover, none of DT and KNN classifiers trained with subsets selected with SN4 or SN5 appeared in the best cluster except for G-KNNSN5. Moreover, all classifiers trained with the original set appear in the best cluster except for U-KNNORG and U-DTORG.

The SK test identified two clusters for CV dataset. A total of 70 variants out of 120 belong to the best SK cluster. 67% (47 classifiers) of the best cluster's classifiers were trained using GS optimization while 33% (23 classifiers) were built using Weka-UC. Moreover, except for G-DTORG, U-DTORG, and U-KNNORG, all classifiers based on the original feature set were present in the best cluster. Except for G-DTSN5, U-DTSN5, and G-MLPSN5, none of the classifiers trained on subsets selected with SN4 and SN5 appeared in the best cluster. Moreover, all classifiers based on subsets selected with MR4, MR5, TT4, or TT5 and optimized with GS appear in the best cluster. All classifiers trained

on subsets selected with RF4 and RF5 belong to the best cluster except for U-DTRF5 and U-KNNRF4. Furthermore, all SVM-based classifiers are among the variants of the best cluster, except for those trained with subsets selected with SN4 and SN5. Additionally, with exception of U-KNNRF5, all KNN-based variants belonging to the best cluster are optimized using GS.

The SK test for ORGCV dataset identified six clusters. The best cluster contained a total of 80 classifiers. It is noteworthy that all DT, SVM and MLP-based classifiers belong to the best cluster except for those based on subsets selected using SN4 or SN5. Moreover, only two KNN-based classifiers (G-KNNCON and U-KNNCON) appear in the best cluster.

For ARR dataset, the SK test identified four clusters. The best SK cluster included a total of 55 variants out of 120 classifiers. It is to be noted that 62% (34 classifiers) of the best cluster's classifiers were trained using GS optimization while 38% (21 classifiers) were trained using Weka-UC. The best cluster for this dataset is mainly composed of DT and SVM-based classifiers, in addition to seven MLP-based variants. SVM, MLP and DT trained with subsets selected with RF4 or TT4 and optimized with GS appear in the best cluster, in addition to U-SVMRF4, U-DTRF4, and U-GDTT4. Furthermore, except for U-SVMCFS and U-SVMCON, all SVM, DT and MLP classifiers trained with subsets selected with CFS or CON are present in the best cluster. G-SVMORG, G-DTORG, and U-DTORG are the only classifiers based on the original feature set that belong to the best SK cluster. Moreover, only one variant based on subset selected with SN4 and four variants on SN4 belong to the best cluster.

The SK test for HF dataset identified four clusters. Of 120 variants, a total of 103 variants belong to the best SK cluster. The best cluster contains 53 and 50 classifiers trained with GS and Weka-UC respectively. Moreover, all classifiers trained with subsets obtained with SN4 belong to the second, third or fourth clusters while all those based on SN5 belong to the best cluster. Except for U-MLPORG, U-KNNORG, and G-KNNORG, all classifiers based on the original feature set appear in the best SK cluster.

In order to answer RQ1, RQ2, and RQ3 from Section 1, the classification techniques present in the best SK cluster for each dataset are summarized in Tables 2-4. Moreover, to answer RQ4 the results of Borda count based on kappa, accuracy, and AUC are given in Tables 5-7. Classifiers with the same ranks are marked with the same letter (*e.g.* [a]).

4. DISCUSSION

In this section we firstly discuss the results according to the aforementioned RQs (see Section 1). Thereafter, we present a comparison of our best results with those from the literature.

4.1 RQ1: Comparison of GS Optimization and Weka-UC

Table 2 shows that the total number of classifiers optimized with GS is higher (or slightly higher) than the number of classifiers trained with Weka-UC for ST, CV, ARR, and HF datasets. For the ORGCV dataset the number of classifiers based on Weka-UC and those based on GS is the same. Accordingly, we can conclude that HP optimization can generally improve the performance of classifiers over different datasets.

Moreover, some observations can be made for each classifier separately.
For KNN, we notice that:

1. From Table 2, KNN classifiers did not appear in the best SK clusters of ORGCV and ARR datasets with exception of two that appeared in the best SK cluster of ORGCV. This shows that the smaller is the number of features the more successful KNN can be.

Table 2. Number of occurrences for each OT and classifier of the best cluster regardless of the FS technique used for all datasets.

Dataset	Weka-UC					Grid Search				
	KNN	SVM	MLP	DT	Total	KNN	SVM	MLP	DT	Total
ST	8	13	13	10	**44**	14	14	14	13	**55**
CV	1	13	6	3	**23**	13	13	14	7	**47**
ORGCV	1	13	13	13	**40**	1	13	13	13	**40**
ARR	0	9	2	10	**21**	0	14	5	15	**34**
HF	10	14	12	14	**50**	11	14	14	14	**53**

2. The number of KNN classifiers optimized with GS exceeds that of KNNs built using Weka-UC in ST and CV datasets. However, all KNN classifiers ranked in the first ten Borda ranks for ST, CV and HF datasets, were optimized using GS. This shows that KNN is sensitive to the tuning of its K hyperparameter.

For SVM, Table 2 shows that, for ST, CV, ORGCV and HF datasets, there is no (or no significant) difference in the numbers of SVM classifiers (belonging to the best clusters) built with Weka-UC or optimized with GS. Therefore, we conclude that in general, HP tuning does not significantly improve the performance of SVMs in small datasets. Nonetheless, for ARR dataset, Table 2 shows that 14 SVM classifiers were optimized using GS while 9 were built using Weka-UC, therefore, it is difficult to draw conclusions for medium datasets and more investigations might be needed. Moreover, the presence of almost all SVM classifiers in the best SK clusters of all datasets shows the robustness of this classifier.

For MLP, Table 2 shows that, except for CV dataset where the majority (14 out of 20) of the MLP classifiers present in the best SK cluster were optimized using GS, there is either an equality or no significant difference in the number of MLP classifiers optimized with GS and Weka-UC for the rest of datasets. MLP seem to perform inconsistently therefore it is hard to conclude whether optimized MLP classifiers outperform those built using Weka-UC or not. In fact, neural networks are generally known to be hard to tune and a brute-force grid search may not be the best choice to find optimal HP values for MLPs. Other optimization techniques such as Random search or Bayesian optimization can therefore be more efficient to examine the impact of HP tuning on MLPs.

From Table 2, it appears that in ORGCV and HF datasets the numbers of DTs optimized using GS and those built using Weka-UC are the same, while for the rest of datasets the number of DT classifiers optimized with GS exceeds by a minimum of three and maximum of five occurrences that of those built using Weka-UC. According to these results it is difficult to draw conclusions. In fact, the Weka-UC values of DT are known to be adequate to simple classification tasks. However, the Borda ranking results shows that HP

optimization can still improve the performance of DT classifiers in some cases since all the DT classifiers appearing in the best ten ranks for ARR dataset are optimized with GS.

4.2 RQ2: Comparison of Multivariate and Univariate Filters

According to Table 3, and based on the initial number of univariate and multivariate techniques used, 73% of multivariate and 67% of univariate FS techniques were present in the best SK clusters over all datasets. Moreover, taking into account each dataset separately: (1) for ST dataset, 82% of FS techniques present in the best SK cluster are univariate while 87% are multivariate; (2) for CV dataset, 59% of FS techniques present in the best SK cluster are univariate while 50% are multivariate; (3) 62% of FS techniques present in the best SK cluster of ORGCV dataset are multivariate while 87% are univariate; (4) 62% of FS techniques present in the best SK cluster of ARR dataset are multivariate while 43% are univariate; and (5) for HF dataset, 88% of FS techniques present in the best SK cluster are univariate while 81% are multivariate.

Table 3. Number of occurrences for each FS technique present in the best clusters regardless of the OT and classification techniques used for all datasets.

	Multivariate			Univariate												
Dataset	CFS	CON	Total	RF4	RF5	LC4	LC5	IG4	IG5	SN4	SN5	MR4	MR5	TT4	TT5	Total
ST	7	7	**14**	8	8	6	8	8	8	**0**	**3**	**8**	**8**	**6**	**8**	**79**
CV	4	4	**8**	7	7	5	4	5	4	**0**	**3**	**7**	**5**	**5**	**5**	**57**
OR-	6	8	**14**	6	6	6	6	6	6	**0**	**0**	**6**	**6**	**6**	**6**	**60**
ARR	5	5	**10**	5	4	4	4	3	3	**1**	**4**	**3**	**4**	**4**	**3**	**42**
HF	8	5	**13**	7	6	8	8	8	8	**0**	**8**	**8**	**8**	**8**	**8**	**85**
Total	**30**	**29**	**59**	**33**	**31**	**29**	**30**	**30**	**29**	**1**	**18**	**32**	**31**	**29**	**30**	**32**

With the exception of ST dataset the percentage of occurrence of univariate techniques exceeds that of multivariate ones in the small datasets. As for ARR dataset multivariate techniques seem to be more successful than univariate ones. This might be due to the fact that the larger a dataset is the higher is the degree of redundant features it contains and that univariate filters do not take into consideration the relationship between features. Nonetheless, univariate techniques can still be beneficial for large datasets by applying them first to reduce the size of the data and select the most informative features, then multivariate techniques can be applied to handle redundancy.

4.3 RQ3: Is There a Best Performing FS Technique?

According to Tables 3 and 4 some observations can be made:

- From Table 3, we observe that the total number of occurrences of RF (64 occurrences over all datasets) and mRMR (63 occurrences over all datasets) techniques exceeds those of other univariate techniques. The efficiency of ReliefF can be explained by the fact that it uses the concept of nearest neighbors to derive feature statistics that indirectly consider feature interactions without evaluating pair-wise feature combinations. Moreover, the mRMR technique, as its name suggests, also considers interactions

between features which can explain its performance. However, both RF and mRMR are mainly influenced by the number of features and this can be confirmed by the fact that the number of occurrences of these two techniques in the ARR dataset is smaller than those in other datasets.

- LC, IG, and TT also seem to provide satisfactory results over different datasets but IG4, IG5 and TT5 appear to be the most affected by the number of features. The classifiers present in the best cluster of ARR dataset and based on these FS techniques are, G-DTIG4, G-DTIG5, G-DTTT5, G-SVMIG4, G-SVMIG5, G-SVMTT5, U-SVMIG4, U-SVMIG5, and U-SVMTT5. Comparing these techniques to baseline classifiers (*i.e.* G-DTORG, G-SVMORG and U-SVMRORG), an improvement was scored for SVMs only. In fact, the SK test compares the kappa means with a significant level of 5%; however small improvements or decreases in the accuracy can still be observed. For instance, we observe that G-DTORG (with a kappa score of 0.64) appears before G-DTIG4 (with a kappa score of 0.62) and G-DTIG5 (a kappa score of 0.60) and thus slightly outperforms them in the ARR dataset.
- The SN technique seem to fail at selecting the relevant features, except in some cases when a threshold of 50% was used.
- There is no difference between the total number of occurrences of CON and CFS techniques present in the best SK clusters of ST, CV, and ARR datasets. For the ORGCV dataset the number of occurrences of CON exceeds that of CFS by two occurrences this is because G-KNNCON and U-KNNCON appeared in the best cluster for this dataset since CON selected the smallest feature subset for this dataset (6 features) compared to CFS (15 features) which may explain this difference. On the contrary, CON selected more features (10 features) than CFS (4 features) for the Heart failure dataset which may explain the difference between the number of occurrences of CFS and CON in this dataset.

Table 4. Number of occurrences of each FS technique present in the best clusters by classifier over all datasets regardless of the OT used.

		Multivariate			Univariate												
	ORG	CFS	CON	Total	RF4	RF5	LC4	LC5	IG4	IG5	SN4	SN5	MR4	MR5	TT4	TT5	Total
KNN	2	4	4	**8**	4	4	4	5	5	5	**0**	**3**	**5**	**5**	**4**	**5**	**49**
SVM	9	9	9	**18**	10	10	10	10	10	10	**0**	**5**	**10**	**10**	**9**	**9**	**103**
MLP	7	9	8	**17**	9	8	8	7	8	7	**0**	**5**	**8**	**7**	**8**	**7**	**82**
DT	7	8	8	**16**	10	9	7	8	7	7	**1**	**5**	**9**	**9**	**8**	**9**	**89**

- Based on the results of Table 4, we notice that for KNN classifiers only 2 occurrences based on the original feature set appeared in the best clusters over all datasets. These occurrences appeared in the best clusters of ST and CV datasets and were optimized using GS. This shows that in addition to HP optimization, feature selection can significantly improve the performance of KNN. In fact, in the case of ORGCV dataset which contains a set of features higher than the other small datasets, only two occurrences of KNN (G-KNNCON and U-KNNCON) appeared in the best cluster and this is because CON selected the smallest subset (6 features) which improved the performance of KNN

remarkably. Therefore, we conclude that feature selection is essential for KNN. Moreover, almost all feature selection techniques of Table 4 improved the performance of KNN to a greater or lesser extent. Nonetheless, we noticed that CFS consistently resulted in slight improvement of KNN in ST, CV, and HF datasets when GS was used since all G-KNNCFS classifiers appeared before G-KNNORG in the best clusters. Moreover, from the Borda count results, we observe that RF4 gave the best results for KNN in ST and CV datasets.

- From Table 4, it can be observed that except SN5, there is no significant difference between the presence frequencies of the different FS techniques for SVM. Moreover, 9 out of 10 SVMs trained with the original feature set over the five datasets performed comparably to FS-based SVMs which shows that this classifier has a low sensitivity to FS in general. However, as previously mentioned small performance improvements can be underlined in each best cluster, and CFS seem to consistently improve the performance of SVM when GS is used. In fact, G-SVMCFS classifiers for ST, CV, HF, and ARR datasets (with a kappa score of 0.68, 0.68, 0.64, and 0.61 respectively) appeared before G-SVMORG classifiers (with a kappa score of 0.66, 0.67, 0.61, and 0.54 respectively).
- Similar to SVM, there is no significant difference between the presence frequencies of the different FS techniques for MLP, except for SN5. A total of 7 out of 8 MLPs trained with the original feature set over the four small datasets performed comparably to FS-based MLPs, while in the medium datasets no MLP based on the original feature set appeared in the best cluster. Moreover, in the ORGCV dataset which contains a set of features higher than the other small datasets G-MLPORG and U-MLPORG appeared last in the best cluster. This shows that MLP is affected by the number of features and FS can help to improve its performance. In fact, MLPs, are not known to be adopted to high dimensional spaces because of the use of classical concepts such as Euclidean distance which scales poorly in high dimensions. For MLP, the most stable FS technique seem to be RF4 when GS is used. In fact, although it was outperformed by other FS techniques in some cases, RF4 improved the performance of MLP in CV (with a kappa score 0.67 vs 0.64), Heart failure dataset (with a kappa score of 0.57 vs 0.53), ORGCV (with a kappa score 1.0 vs 0.98), and ARR dataset (with a kappa score of 0.54 vs 0.50) compared to G-MLPORG, and conserved the same accuracy in ST dataset (a kappa score of 0.69). Also, CON FS gave good results for MLP for ST, OTGCV, and ARR dataset based on the Borda results.
- Compared to SVM and MLP, DT classifiers seem to be very competitive. No specific FS performed the best for DT in all datasets, however, RF resulted in the best classification performance for DT in ST and CV datasets when GS is used, while CFS produced the best results for DT in ARR and HF dataset when GS was used. For ORGCV, with exception of those based on SN4 and SN5, all DT-based classifiers achieved an accuracy of 100% (*i.e.* a kappa score of 1.0).

4.4 RQ4: Best OT, FS, and Classifier Combination

According to the Borda count results (Tables 5-7), there are several combinations that have performed well in different datasets. For instance, G-MLRF4 appear in the top ranks for ST, CV and ORGCV datasets. This confirms the fact that ReliefF and HP optimization

can improve the performance of MLP. Moreover, the combination G-SVMCFS also appears among the best ten ranks of ST, CV, ORGCV and ARR datasets. This shows that although SVM proved to be a robust technique, HP optimization and feature selection can still be beneficial in terms of accuracy. Moreover, as long as feature selection do not worsen its performance it can help in terms of reducing training time. As for white-box classifiers, we can notice that different KNN-based combinations appear in the top ranks for small datasets while different DT-based combinations appear in the top ranks for the medium dataset. For instance, the combinations G-KNNIG4, G-KNNMR4 appear in ST and HF datasets, G-KNNRF4 appears in ST and CV datasets, while G-KNNCON appears in ORGCV. Moreover, G-DTCFS and other DT-based combinations appear in the top ranks for the medium dataset (ARR) and ORGCV which contains a set of features higher than the other small datasets. This demonstrates that using HP optimization and FS can make simple classifiers such as KNN and DT very competitive with more robust black-box classifiers in terms of accuracy, not to mention their interpretability aspect.

Table 5. Borda count results for ST and CV datasets.

ST dataset				CV dataset			
Rank	Classifiers	Rank	Classifiers	Rank	Classifiers	Rank	Classifiers
1	G-MLPORG	6	[b]G-MLPLC4	1	G-KNNRF4	6	G-MLPRF4
2	[a]G-SVMTT5	7	G-SVMCFS	2	G-SVMCON	7	G-MLPTT5
3	[a]G-MLPRF4	8	[c]G-KNNMR4	3	U-SVMCON	8	U-SVMCFS
4	[b]G-SVMMR5	9	[c]G-KNNIG4	4	G-KNNRF5	9	G-KNNTT5
5	[b]G-KNNRF4	10	[c]G-MLPCON	5	G-SVMCFS	10	G-SVMORG

Table 6. Borda count results for ORGCV and ARR dataset.

ORGCV dataset				ARR dataset			
Rank	Classifiers	Rank	Classifiers	Rank	Classifiers	Rank	Classifiers
1	[a]G-DTCFS	6	[a]G-KNNCON	1	G-DTCFS	6	G-MLPCFS
2	[a]G-DTTT4	7	[a]G-SVMRF4	2	G-MLPCON	7	G-DTLC4
3	[a] G-MLPRF4	8	[a]G-MLPCON	3	G-DTTT4	8	G-DTCON
4	[a]G-SVMCFS	9	[a]G-DTRF5	4	[a]G-DTMR5	9	[b]G-SVMCFS
5	[a]G-DTCON	10	[a]G-SVMCON	5	[a]G-DTTT5	10	

Table 7. Borda count results for HF dataset.

HF dataset			
Rank	Classifiers	Rank	Classifiers
1	G-MLPMR5	6	[a]G-KNNIG4
2	G-MLPMR4	7	[a]G-MLPLC4
3	G-MLPTT4	8	[a]G-KNNTT4
4	[a]G-KNNCFS	9	G-MLPTT5
5	[a]G-KNNMR4	10	G-MLPSN5

4.5 Accuracy Results Comparison

The accuracy results of the best performing classifiers of the present study are compared with those from previous studies in Table 8. In fact, the majority of the studies presented in Table 8 used black-box classifiers such as neural networks, support vector machine, and Random forest. Moreover, some studies used feature extraction techniques such

as principal component analysis which transforms the original features into new ones. Nonetheless, critical decisions such as the diagnosis of heart disease require white-box classifiers that are more comprehensible and easily interpreted [3]. Hence, in case the best performing classifier is a black-box, we also added the first best performing white-box (*i.e.* DT or KNN) classifier from the Borda results for each dataset to conduct the comparisons. As shown in Table 8, the results achieved are very promising compared to those of the previously published studies.

Table 8. Accuracy results comparison with previous studies.

Dataset	Study	Technique	No. of features	Accuracy
Statlog	**Our study**	**G-MLPORG**	**13**	**85.18%**
		G-KNNRF4	**5**	**84.81%**
	Polat, *et al.* [10]	RBF kernel F-score FS + LS-SVM	117	83.70%
	Jaganathan, *et al.* [11]	Fuzzyentropy-NNTS FS + RBF	4	85.19%
Processed Cleveland	**Our study**	**G-KNNRF4**	**5**	**86.13%**
	Jaganathan, *et al.* [11]	Fuzzyentropy-NNTS FS + RBF	3	84.46%
	Vivekanandan, *et al.* [12]	Modified differential evolution FS + fuzzy AHP + feed-forward neural network	9	83%
Unprocessed Cleveland	**Our study**	**G-DTCON and U-DTCON**	**6**	**100%**
	Garate-Escamila, *et al.* [13]	CHI-PCA FS + Random Forest	13	98.7%
	Miao, *et al.* [14]	Adaptive boosting	29	80.14%
	Patra, *et al.* [15]	IG FS + DT	51	87.12%
Arrhythmia	**Our study**	**G-DTCFS**	38	**82.52%**
	Mustaqeem, *et al.* [16]	Random Forest wrapper FS + MLP	–	78.26%
	Sasikala, *et al.* [17]	PCA + ReliefF-Shapley FS + SVM	–	65.78%
	Niazi, *et al.* [18]	Improved F-score FS + KNN	60	73.8%
Heart failure	**Our study**	**G-MLPMR5**	**6**	**85.61%**
		G-KNNCFS	**4**	**84.93%**
	Oladimeji, *et al.* [19]	Filter FS + Random Forest	4	83.17%

5. CONCLUSIONS AND FUTURE WORK

This paper studied the effect of feature selection and HP optimization on HD classification. To this end, the relevant features of five datasets related to heart disease were selected using CFS and CON multivariate filters, and RF, LC, IG, SN, MR, and TT univariate filters with 40% and 50% thresholds. KNN, SVM, MLP and DT classifiers were trained with the entire and reduced feature sets, and evaluated using the 10-fold cross validation method. The accuracy results of the best performing white-box classifiers of the present study were compared with those from previous studies. In addition to the interpretability advantage, the constructed techniques showed very promising results in terms of accuracy as well.

Ongoing works aim to investigate the application of other preprocessing tasks such

as normalization and missing data handling in attempt to obtain better results. Constructing ensemble feature selection techniques will also be investigated to combine the robustness and eliminate the drawbacks of the individual ones.

REFERENCES

1. H. Benhar, A. Idri, and J. L. Fernández-Alemán, "A systematic mapping study of data preparation in heart disease knowledge discovery," *Journal of Medical Systems*, Vol. 43, 2019, p. 17.
2. I. Kadi, A. Idri, and J. L. Fernandez-Aleman, "Knowledge discovery in cardiology: A systematic literature review," *International Journal of Medical Informatics*, Vol. 97, 2017, pp. 12-32.
3. H. Benhar, A. Idri, and J. L Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," *Computer Methods and Programs in Biomedicine*, Vol. 195, 2020, pp. 1-30.
4. A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics*, 2015, pp. 1200-1205.
5. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, Vol. 11, 2009, pp. 10-18.
6. D. Dua and C. Graff, "UCI machine learning repository," School of Information and Computer Sciences, University of California, Irvine, 2019, http://archive.ics.uci.edu/ml.
7. S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, Vol. 4, 2010, pp. 40-79.
8. H. Benhar, A. Idri, and M. Hosni, "Impact of threshold values for filter-based univariate feature selection in heart disease classification," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies*, 2020, pp. 391-398.
9. N. Pilnenskiy and I. Smetannikov, "Feature selection algorithms as one of the python data analytical tools," *Future Internet*, Vol. 12, 2020, No. 54.
10. K. Polat and S. Gunes, "A new feature selection method on classification of medical datasets: Kernel F-score feature selection," *Expert Systems with Applications*, Vol. 36, 2009, pp. 10367-10373.
11. P. Jaganathan and R. Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification," *Computers in Biology and Medicine*, Vol. 43, 2013, pp. 2222-2229.
12. T. Vivekanandan and N. I. N. C. Sriman, "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease," *Computers in Biology and Medicine*, Vol. 90, 2017, pp. 125-136.
13. A. K. Gárate-Escamila, A. H. el Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, Vol. 19, 2020, No. 100330.
14. K. H. Miao, J. H. Miao, and G. J. Miao, "Diagnosing coronary heart disease using ensemble machine learning," *International Journal of Advanced Computer Science and Applications*, Vol. 7, 2016, No. 071004.

15. R. Patra and B. Khuntia, "Predictive analysis of rapid spread of heart disease with data mining," in *Proceedings of IEEE International Conference on Electrical, Computer and Communication Technologies*, 2019, pp. 1-4,
16. A. Mustaqeem, S. M. Anwar, M. Majid, and A. R. Khan, "Wrapper method for feature selection to classify cardiac arrhythmia," in *Proceedings of IEEE Annual International Conference of Engineering in Medicine and Biology Society*, 2017, pp. 3656-3659.
17. S. Sasikala, A. B. S. Appavu, and S. Geetha, "RF-SEA-based feature selection for data classification in medical domain," *Intelligent Computing, Networking, and Informatics*, Vol. 243, 2014, pp. 599-608.
18. K. A. K. Niazi, S. A. Khan, A. Shaukat, and M. Akhtar, "Identifying best feature subset for cardiac arrhythmia classification," in *Proceedings of Science and Informa-tion Conference*, 2015, pp. 494-499.
19. O. O. Oladimeji and O. Oladimeji, "Predicting survival of heart failure patients using classification algorithms," *Journal of Information Technology and Computer Engineering*, Vol. 4, 2020, No. 02.

Houda Benhar is a Ph.D. student at the Software Project Management Team in the ENSIAS, University Mohammed V, Rabat, Morocco in fall 2016. Her research focuses on the application of data preprocessing techniques, particularly feature selection, in medical data mining.

Mohamed Hosni received the Engineering degree of Computer Science from the ENSA, Oujda, Morocco in 2014, and Ph.D. in Software Engineering in 2018 from ENSIAS, University Mohammed V in Rabat, Morocco. He is currently an Assistant Professor with the College of Art and Crafts (ENSAM), University Moulay Ismail of Meknes.

Ali Idri is a Full Professor at the Computer Science and Systems Analysis School (ENSIAS), University Mohammed V, Rabat, Morocco. He received his Master and Doctorate of 3rd Cycle in Computer Science from the University of Mohamed V in 1994 and 1997 respectively. He received his Ph.D. in Cognitive and Computer Sciences from the University of Quebec at Montreal in 2003. He published more than 250 papers in well-recognized journals and conferences. His research interests include medical informatics, machine learning and software engineering.