International Journal of Data Warehousing and Mining
Volume 20 • Issue 1 • January-December 2024

A Machine Learning-Based Wrapper Method for Feature Selection

Damodar Patel
Guru Ghasidas Vishwavidyalaya, India

Amit Saxena
Guru Ghasidas Vishwavidyalaya, India

John Wang
https://orcid.org/0009-0007-0296-3264
Montclair State University, USA

ABSTRACT

This paper presents a two-stage feature selection scheme using machine learning techniques. In the first stage a wrapper method is adopted to select various combinations of subsets of features from the original dataset. The performance of the model is evaluated by three classifiers: K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Random Forest (RF). In the second and final stage, a sequential backward feature selection Method is applied. The proposed method is demonstrated on eighteen datasets and the average classification accuracy of eighteen datasets achieved is 89.81%, 87.55%, and 89.82% using the KNN, SVM, and RF classifiers, respectively with a maximum reduced size of the subset being ten only. Comparing the proposed method to eight other feature selection methods, the former achieves better classification accuracy in terms of selecting the most useful but a smaller number of features.

KEYWORDS

Feature Selection, Wrapper Method, Dimensionality Reduction, Machine Learning, Data Mining

INTRODUCTION

As so many everyday actions in our society are automated using various digital platforms, we need large storage media as backups for those actions. These days, well-formatted datasets with a well-organized structure (such as relational databases) often include a large number of patterns and a small number of classes for use in computer-based applications(Patel et al., 2022).

Every pattern has a set of features that help to represent it. Every pattern belongs to one of the classes. The extraction of relevant details from a huge database is known as data mining(Kamber et al., 2006; Zhu et al., 2023). Classification, an important aspect of data mining, is the process of separating data into different groups or classes. A key component of classification is feature analysis. In feature analysis, two essential parts are feature extraction and feature selection (FS). Finding a subset of features in a dataset is called feature selection (FS) (Chakraborty & Pal, 2008; Iguyon & Elisseeff, 2003). On the other hand, feature extraction may combine or recalculate existing features to produce new ones. A dataset could contain features that are noisy or duplicated. Along with the possibility of increasing confusion, these additional features also increase the complexity and, subsequently, the cost of the classifier. With the proper but less amount of information (or features), a classifier may

DOI: 10.4018/IJDWM.352041

sometimes provide results that are more accurate than those with more features. If the FS makes use of data (such as the class or label of a pattern) that was supplied before the classification process, the approach is known as supervised. An algorithm that classifies patterns without providing any prior information is known as an unsupervised algorithm(Chugh et al., 2023; Saxena et al., 2017). Several popular methods like Artificial Neural networks (Setiono & Liu, 1997), fuzzy logic(Zadeh, 1988), random-forest algorithm(Zhu et al., 2024)and K-NN (Yang et al., 2007) are used in supervised FS techniques.

There are primarily two fundamental types of FS methods, namely filter and wrapper-based methods(Saxena & Dubey, 2015), although these two types can lead to another type known as embedded methods. Filter based techniques use a score for each feature of the dataset based on various evaluation factors. Thus, the relevant characteristics are selected consecutively until either the unique threshold (for example, Classification Accuracy, CA) is not achieved or the required number of features is not obtained. A feature's significance is determined based on each of its unique properties. Furthermore, filter approaches may not take into consideration the relationship between the selected features and the learning process, which might produce poor results on regression tasks.

The wrapper method generates a subset of features (taken randomly usually) that are subjected to some learning algorithm acting as a classifier. The selected features are then frequently updated based on certain search methods until the performance or iteration criteria are satisfied. In the embedded methods, feature selection is a part of the model training process. Irrelevant features are eliminated or penalized during training, and feature relevance is decided as the model learns. Tree-based techniques (like Random Forest) and LASSO (L1 regularization) are examples of popular embedded approaches.

Three benchmarked supervised classifiers namely K-NN (K-Nearest Neighbor) (Yang et al., 2007), SVM (Support Vector Machines) (Wang, 2022), and RF (Random Forest) (Genuer et al., 2010) are used in the experiments to evaluate the CA of prediction by the algorithm.

In this study, we propose a novel Machine Learning-based Wrapper Feature selection method (MLWFS) for FS in large dimensional datasets by reducing the number of features in two stages. The proposed method is compared with the one published by Zhao et al. (Zhao et al., 2022), in which eight FS methods have been used, with a maximum number of thirty features selected. These methods contain FSRRW (relevant-redundant weight-based feature criterion) (Zhao et al., 2022), MIFS (mutual information feature selector) (Battiti, 1994), JMI (join mutual information) (H. Yang & John, 1999), mRMR (minimum-redundancy maximum-relevance) (Peng et al., 2005), CIFE (class-relevant redundancy)(Lin & Tang, 2006), MRI (max-relevance and max-independence) (J. Wang et al., 2017), DCSF (Dynamic Change of Selected Feature with the class) (Gao et al., 2018), & CWJR (conditional weight-based joint relevance) (Zhang et al., 2021). Better CA with fewer selected features is the objective, compared to using eight methods.

The structure of the paper is as follows: In Section 2, some feature selection techniques are provided. Section 3 defines the proposed MLWFS methodology. The experiments are presented in Section 4. Results and Discussions of experiments are covered in Section 5. The last part of the paper presents conclusions and potential areas for further study.

SOME FEATURE SELECTION TECHNIQUES

This section presents some of the feature selection techniques over the last few years. Modern life focuses heavily on machine learning, yet it has difficulties with high-dimensional data that contains extraneous elements. In order to deal with this issue, Rani et al. (2021) presented GARFE, a hybrid feature selection technique that combined recursive and evolutionary feature elimination methods. In tests using a support vector machine classifier, GARFE outperforms eight other techniques by removing features that are not important to the classification system.

The work by Ghosh et al. (2024)uses a two-stage hybrid feature selection strategy to identify breast, lung, and cervical cancer. It evaluates several machine learning models with the goal of reducing mortality through rapid and accurate diagnosis.

In Nesamaniet al. (2023), the features are selected using mutual information and chi-squared approaches, with mutual information working better for precise predictions on breast cancer datasets.

Wang et al. (2022) proposed a novel approach for diagnosing faults in permanent magnet DC motors (PMDCMs) called Gaussian vote feature selection (GVFS). GVFS enhances diagnostic accuracy over other techniques by using the Gaussian probability density function (GPDF) to reduce redundant data. According to experimental data, there is a significant improvement in the accuracy of diagnosing faults, demonstrating that GVFS is a useful tool for PMDCM fault identification.

In Liao et al. research (2014), local geometric structure is prioritized over global information when selecting genes from microarray data for tumor classification. They provided the locality-sensitive Laplacian score (LSLS), a technique that combines variance considerations with discriminative information. Using wrapper techniques, LSLS is included in a two-stage feature selection process and outperforms the most advanced algorithms for classifying tumors on six different datasets.

Chang et al. (Chang et al., 2013) presented combining genetic and clinicopathologic markers to examine the use of machine learning in the prognosis of oral cancer. Due to insufficient data, four classifiers are evaluated using k-fold cross-validation after five feature selection approaches are tried. With its maximum accuracy of 93.81%, the hybrid ReliefF-GA-ANFIS model highlights the need for combined markers for better prediction.

Due to processing limitations, heart disease identification is still difficult even with modern Electrocardiography (ECG) technologies. A novel technique called the Gradient Squirrel Search Algorithm-Deep Maxout Network (GSSA-DMN) is suggested. It uses a DMN trained by GSSA, preprocesses the data, then selects features using ReliefF. GSSA-DMN outperforms current techniques and achieves excellent accuracy (93.2%), predicting increased effectiveness in the identification of heart disease(Balasubramaniam et al., 2024).

Sine-cosine hybrid optimization with a modified whale optimization approach (SCMWOA) is a method for handling limited optimization problems that are presented by El-Sayed et al. (2022). The SCMWOA method was tested on nineteen datasets.

Zeng et al. (2011) aimed to improve the efficiency of the clustering method by improving the data representation in the input or kernel space. It optimizes relevance for high-dimensional data on manifolds by including weights for features or kernels inside the Local Learning-Based Clustering (LLC) framework.

The optimization of feature selection for classification algorithms such as neural networks is the main topic of Kwak et al. (2002). Two techniques are proposed to overcome the constraints of the mutual information feature selector (MIFS). By using mutual information more efficiently than MIFS, the first method performs better and exhibits optimal performance under uniform information distribution. Another algorithm efficiently finds useful features with a few trials by using the Taguchi method. When combined, these algorithms perform well in classification tasks and provide a useful feature selection technique.

A relevant-redundant weight-based feature criterion (FSRRW) is presented in the study by Zhao et al. 2022) to improve classification in machine learning and data mining. In order to extract significant information, the criteria construct a feature known as relevant-redundant weight (RRW). Based on 20 benchmark datasets, the results indicate that FSRRW performs better than seven related approaches in terms of robustness, feature screening, and classification

Yang et al.'s (1999) present feature selection techniques based on independent component analysis (ICA) and joint mutual information, in addition to data visualization. The visualization approaches exceed the current methods for effectively identifying 2-D projections for high-dimensional data analysis. When it comes to minimizing data redundancy, a novel variable selection method outperforms simple mutual information-based techniques. These techniques are shown to be effective in the

analysis of radar signals, which includes the display of 2-D viewing coordinates and the choice of input for a neural network classifier.

Peng et al. (2005) present a two-stage technique that combines improved feature selectors with the minimal-redundancy-maximal-relevance criteria (mRMR) to manage feature selection in pattern classification. This method produces an efficient feature set that has been verified by extensive evaluation of various datasets and classifiers, confirming mRMR's significant improvement in feature selection and CA.

Lin et al. (2006) present an information-theoretic feature learning approach for classification. It analyzes information structure methodically, focusing on redundancy and class relevance. The conditional informative feature extraction method is based on a unique information decomposition model that includes class-relevant redundancy. The local active region technique and the parent window estimate are used to tackle computational problems. Laplacian Sparse Prior and Multivariate Logistic Regression are used in a Bayesian MAP framework for feature fusion. Comparative trials show considerable increases in learning efficiency due to the framework's better architecture, which coordinates the extraction and fusion processes.

In order to balance redundancy and new information, Wang et al. (2017) proposes a novel term, independent classification information, to overcome the imbalance in current mutual information-based feature selection algorithms. Extensive experiments demonstrate that the term successfully detects predictive features with large amounts of new information and low redundancy, improving global discriminative performance.

In order to merge criteria from two categories, minimizing redundancy and maximizing new classification information, Gao et al. (2018) presents a hybrid technique, called MR-MNCI, for feature selection in expert and intelligent systems.

The research work by Zhang et al. (2021) applies information theory to feature selection in data mining and machine learning. It presents a novel method called CWJR-FS, which uses conditional-weight joint relevance to extract important classification information.

Saxena et al. (2010) proposed a two-phase approach to evaluating CA and selecting features using a genetic algorithm (GA). In the first phase, an unsupervised GA with Sammon Error is used, and a better CA is revealed by extensive simulations in the second phase, which examines CA using k-nearest neighbor (k-nn) using various Minkowski metrics.

In high-dimensional datasets, hybrid approaches are essential for feature selection. This work by Saxena et al. (2017) presents two hybrid approaches that combine sequential random search, genetic algorithms, and filter-based feature selection. While the second approach combines information gain with sequential random k-nearest neighbor (SRKNN), the first approach uses information gain when combined with genetic methods. Tested on 21 high-dimensional datasets, results show varying performance, with each method excelling on different datasets compared to other methods.

The research work by Saxena et al. (2021) addresses the problem of initial centroid selection in K-means clustering, which affects the accuracy of the clustering. It explores an initialization method for centroids based on density and distance, demonstrating enhanced accuracy along with faster convergence. Through comprehensive studies on 15 datasets, the paper examines this technique and shows that it works better than random initialization and achieves excellent clustering accuracy with less features which is beneficial for large-scale data mining.

PROPOSED MLWFS METHODOLOGY

The MLWFS begins with a small number of features selected at random from the dataset's original feature set. This is an overview of the whole algorithm for any one of the three classifiers named KNN, SVM, RF. The overall proposed method is divided into two stages, as follows:

Stage 1: The objective of this stage is to reduce the size of the original dataset to a reasonably small subset with effective features. The first step is to divide the original dataset into a user defined

number of folds such that each of the folds has an equal number of randomly selected features from the original dataset, in it. In exception case, the last fold may be left with fewer number of features compared to its fellow other folds if number of folds cannot be equally divided into total number of features. Now each of these folds is exposed to a classifier (KNN/SVM/RF) and the fold of features with highest CA is saved. Stage 2: The main objective of this stage is to reduce further the number of features obtained in the Stage 1. The feature set of the fold having highest CA obtained in Stage 1, is used in this stage. The sequential backward selection is applied here. Each time one feature from the feature set is eliminated from the rear end and the CA is computed using the same classifier, which was used in Stage 1, with the remaining number of features. The process is repeated unless we are left with only one feature in the reduced feature set. The feature set with the highest CA is saved. This will ensure the selection of the best feature set with the least dimensionality (cardinality) but having the highest CA.

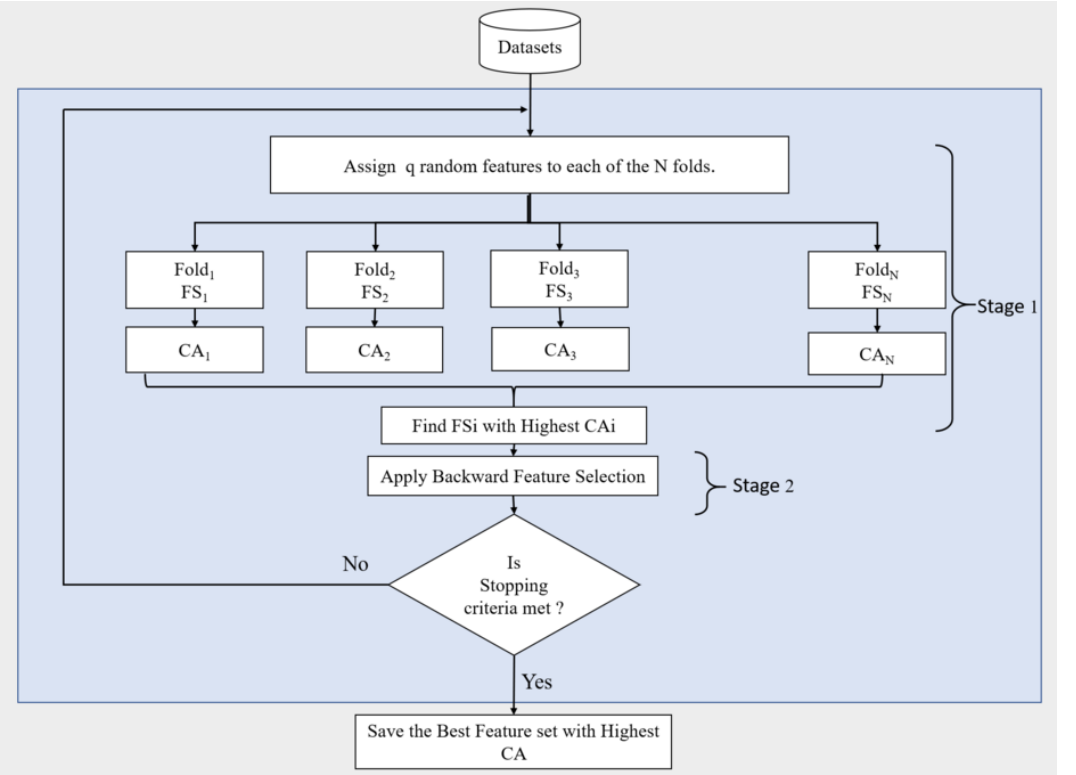
Stage 1 is again invoked to apply the method with another classifiers viz. KNN/SVM/RF. Thus, out of the three classifiers, we get the best feature sets belonging to each classifier.

The expected outcome of State 1 is the achievement of the feature set with highest CA but with minimum possible dimensionality which is the main objective of the present work. The expected outcome of Stage 2 is getting the reduced feature set by eliminating features one-by-one and computing CA each time.

Stage 1 is expressed in Algorithm 1 whereas the Stage 2, a part of Stage 1, is presented separately in Algorithm 2. The overall method comprising of Stage 1 and 2 is depicted in Figure 1. As apparent from Figure 1, the original datasets first split into several folds (decided by user) having equal number of features randomly selected. The CA of each fold is evaluated using one of the classifiers. The feature set with highest CA is selected for further processing. This feature set is used for sequential backward selection one by one at a time. The stopping criteria shown in Figure 1 is taken as the minimum number of features finally left (typically 1 or as decided by user) in the fold while eliminating features one by one using sequential backward feature selection.

```
   Algorithm 1: Proposed method:
   1.  Let the cardinality of the set of all features in the
original dataset be F.
   2.  Divide the F features into N (or N+1) folds as follows. Let
each fold have q features (here q = 10). If mod(F/q) ~= 0 then
(N+1) th fold will contain the (F-N*q) features (i.e. excess of
N*q). For ease of understanding we assume that mod(F/q) = 0 and
there are N folds.
   3.  Select q features randomly from F for each of the N folds.
   4.  Determine the CAi  (i=1,2,…N) for each of the N folds.
   5.  Note the features of the fold with the highest CA for the
current trial.
   6.  Apply Sequential Backward Selection (as given in algorithm
2) on the features of the fold obtained to get the   reduced
feature set for the current trial.
   7.  Repeat steps 2 to 6 and save BFS(t) is the highest CA of
all trials, t=1,...,T for trials T (T=100).
   8.  Find BFj= BFSi(tj) for the highest CAi that the classifier
Cj achieved to obtain for the ith trial.
   9.  Repeat steps 2 to 9 and find the BFi for classifiers with
j=1, 2, and 3.
   Algorithm 2: Sequential Backward Selection:
   1.  Let the cardinality of the subset of features obtained
after Step 5 of Algorithm 1, be q (for example F = 10).
```

Figure 1. Process of proposed MLWFS for any classifier



2. Iteration 1. Generate every possible feature subsets of size q-1(for example 10-1 = 9).
3. Remove the feature that is absent from the subset with the highest CA.
4. Iteration 2. Generate every possible feature subsets of size q-2 (for example 10-2 = 8)
5. Repeat step 3 and 4 until the feature subset contains only one feature. Give an initial feature set of size F, there are total F-1 iteration. (for example, total 9 iteration)
6. Considering all iteration 1……F-1, the subset with highest CA is selected as the final feature subset and in case of a tie, the smallest feature subset is selected.

As an illustration, suppose the original dataset has 200 features and we decide to keep 10 features in each fold (20 folds in all) and these 10 features will be taken randomly from the original dataset to ensure participation of all types (classes) of features in each fold. Calculate the CA of each of these 20 folds using KNN first. Say a feature set of a particular fold with 10 features (30, 130, 45, 56, 23, 147, 198, 3, 20, 115) produces highest CA. Now apply sequential backward feature selection to this set. First find CA with all features using KNN. The drop last feature 115 from the fold and compute CA with remaining 9 features of this fold. We decide to keep atleast 3 features in the fold thus we get CAs for the feature sets (30, 130, 45, 56, 23, 147, 198, 3, 20, 115), (30, 130, 45, 56, 23, 147, 198, 3, 20), (30, 130, 45, 56, 23, 147, 198, 3), (30, 130, 45, 56, 23, 147, 198), (30, 130, 45, 56, 23, 147), (30, 130, 45, 56, 23), (30, 130, 45, 56), (30, 130, 45) consecutively obtained using eliminating last

Table 1. Datasets description

Datasets	Records	Features	class
ALLAML	72	7129	2
Lymphoma	96	4026	9
Lung	203	3312	5
orlraws10P	100	10304	10
lung-discrete	73	325	7
ProsteGE	102	5966	2
Ionosphere	351	34	2
warpPIE10P	210	2420	10
FTM (ForestTypeMapping)	326	27	4
UrbanLandCover	168	148	9
GLIOMA	50	4434	4
CLL_SUB_1111	111	11340	3
QSAR_B (QSAR_Biodegradation)	1055	41	2
ThoracisSurgery	470	17	2
SCADI	70	206	7
CBSMR(ConnectionistBench_SonarMines_Rocks)	208	60	2
warpAR10P	130	2400	10
ORL	400	1024	40

feature every time. Note the feature set which produced highest CA from amongst these feature sets. Repeat the exercise for SVM and then for RF and find the best feature set for each of these classifiers.

EXPERIMENTS

Experiment Setting

The MLWFS method was performed on a Windows 10 machine equipped with an Intel (R) Core i3, 4 GB of RAM, a 2.30 GHz CPU, and a 500 GB SSD card using Python 3.12.1.

Three classification models, namely K-NN (K = 3), SVM (kernel = rbf), and RF, are applied to evaluate the CA obtained by the proposed MLWFS method using the existing labels available at the data source. In the experiment, a standard scaler was used for normalizing the data. A ten-stratified k-fold(Browne, 2000) was used in this experiment. Table 2 provides the performance metrics (CA, F1 score, Precision, Recall) values throughout a ten-stratified k-fold cross-validation period for the study.

Dataset Description

The effectiveness of the proposed MLWFS method is evaluated using eighteen benchmark datasets from ASU (*Arizona State University Library*, n.d.) and the UCI machine learning repository (Dua, D. and Graff, 2019). The essential information about the datasets is shown in Table 1. The majority of these datasets are high-dimensional datasets, with sizes ranging from 17 to 11,340 features. Out of the eighteen datasets, six datasets have binary classes whereas the twelve datasets have multiclass.

Models Used

K-NN

The K-Nearest Neighbor classifier (K-NN)(S. Yang et al., 2007) uses a distance measure like Euclidean distance to classify test data based on the known labels of the k closest neighbors. The majority of the class labels predicted by training patterns determine the class of the test patterns(A. Saxena et al., 2022).

SVM

Support Vector Machines (SVM)(Q. Wang, 2022) are operated by locating the maximal margin hyperplane, or the linear separator, that is as far away from the positive and negative training data as possible. To make the linear separator very non-linear in the input space, kernel functions may be implemented to project the data into a high-dimensional space.

RF

The supervised learning method includes the well-known machine learning algorithm Random Forest (RF) (Genuer et al., 2010). It is used for ML problems involving both classification and regression. The idea of ensemble learning provides its basis. The outcomes of many decision trees are merged to get a single conclusion. Its popularity is increasing as a result of its adaptability and simplicity (Allheeib et al., 2023). The ultimate result is predicted by the random forest using predictions from each tree, and the majority votes for those predictions(Patel & Saxena, 2022).

Performance Evaluation and Correlation Heatmap

We evaluated the performance across multiple experiments in this research. The confusion matrix (CM) was computed to identify the true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The following is a definition of TP, TN, FP, and FN:

Given below the definitions and formulas to evaluate the performance of the model: accuracy, Precision, Recall, and F1 score.

Classification Accuracy (CA): In terms of CA, it is a ratio of accurately predicted occurrences to all observations in the dataset. CA was calculated using formula (1). The total number of TP and TN divided by the total number of TP, TN, FP, and FN provides the CA.

$$Accuracy\left(CA\right) = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision: The ratio of accurately predicted positive observations to the total number predicted positives is known as precision. Precision was calculated using equation (2). The total number of TP is divided by the total number of TP, and FP provides the precision.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall: The ratio of accurately predicted positive observations to all of the dataset's actual positive observations is known as recall. Recall was calculated using formula (3). The total number of TP divided by the total number of TP, and FN provides the precision.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 score: The harmonic mean of recall and precision is the F1 score. It provides a balance between recall and precision. F1 score was calculated using equation (4).

$$F1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

Correlation Heatmap (CH): A table that shows the correlation coefficients for multiple features is called a correlation matrix. The matrix displays the correlation between each group of numbers in a table. A CH provides a 2D correlation coefficient between two discrete dimensions and displays the data on a usually monochrome scale using colored pixels. In the table, the values of the first dimension are shown as rows, while the values of the second dimension are represented as columns. The color of the cell indicates the percentage of measurements that match the dimensional value. The correlation ranges from 1 to -1, where

- A value of 1 denotes highly redundant features in a positive direction.
- A value of -1 denotes highly redundant features in a negative direction.
- A value of 0 denotes that there is no redundancy in the features.

CH is used to check the redundancy and data analysis.

OBJECTIVES AND HYPOTHESES OF EXPERIMENTS

The objective of carrying experiments in the proposed method is to investigate the performance of the proposed model by empirical study on the eighteen datasets. Having eighteen datasets will ensure that the model can be used for other more datasets as well. The hypothesis is to see the size of the feature sets as minimum as possible but still producing reasonably good CA. This will reduce the time and computational efforts by any machine as only a limited (small) number of features will be able to address the whole dataset as far as prediction accuracy is concerned. The comparison with other reported methods on the similar datasets justifies the performance of the proposed or any particular method over other methods. A single method whether proposed or any other may not be able to predict highest CA in all datasets as the internal structure of dataset may vary at individual level. Thus, it is needed to know the performance of other methods and let the user decide which method is suitable for a particular dataset. Keeping this in mind, comparison with other reported methods is also presented in the paper.

RESULTS AND DISCUSSION

The experimental results of the proposed MLWFS method are shown in Table 2. In Table 2, the first column presents the name of the datasets, the second column presents the classifier name, and the third to sixth column present the values of different performance metrics namely classification accuracy (CA), F1 score, precision, recall respectively and the last column of Table 2 presents the number of selected features (SF).

When we notice in Table 2, the number of features required to achieve the evaluation metrics (CA, F1, Precision, Recall), it is noted that with a minute reduction/ variation in the values of these parameters, the number of features are increased significantly. E.g. The highest CA is 99.76%(SVM), F1 score is 99.78%(SVM),, precision is 99.81% (SVM),, and recall is 99.78% (SVM), for lymphoma but number of SFis 10. If we consider ALLAML, then CA, F1 score, precision, and recall is obtained as 99%, 98.99%, 99.17%,99% respectively in KNN with just 5 features: a reduction of 50%. In terms of average performance evaluation, when compared to all eighteen datasets in KNN, we achieve average CA, average F1 score, average precision, and average recall of 89.81%, 88.54%, 88.38%, and 89.38%, respectively; in the SVM, we achieve average CA, average F1 score, average precision, and average recall of 87.55%, 85.95%, 86.39%, and 87.5%, respectively; and in RF, we achieve average CA, average F1 score, average precision, and average recall of 89.82%, 86.32%, 87.21%, and 88.05%, respectively.

Table 2. Result of eighteen datasets using KNN, SVM, and RF classifiers in the MLWFS method (in %)

Datasets	Classifier	CA	F1 Score	Precision	Recall	SF
ALLAML	KNN	99	98.99	99.17	99	5
	SVM	97	96.97	97.5	97	4
	RF	99	97.88	99.17	99	7
lymphoma	KNN	97.83	97.59	98.39	97.83	10
	SVM	**99.76**	99.78	99.81	99.78	10
	RF	98.57	97.98	98.58	98	8
Lung	KNN	98.27	98.24	98.45	98.27	7
	SVM	99.14	99.13	99.2	99.13	10
	RF	98.56	98.24	98	98.26	10
orlraws10P	KNN	95	93.33	92.5	95	7
	SVM	96	94.67	94	96	10
	RF	96	90.67	94	97	7
lung-discrete	KNN	94.62	94.37	96.31	94.76	9
	SVM	95.33	95.12	96.81	95.24	9
	RF	94.67	92.45	92.52	92.62	9
Prostate_GE	KNN	94.27	94.18	95.24	94.17	6
	SVM	92.18	92.16	92.67	92.33	6
	RF	93.27	91.22	90.65	91.33	8
Ionosphere	KNN	94.22	94.43	92.21	96.92	7
	SVM	93.78	93.97	92.09	96.03	7
	RF	93.11	92.87	93.13	92.52	7
warpPIE10P	KNN	92.86	92.17	92.5	93	9
	SVM	90.48	90.18	92.69	90.83	8
	RF	93.33	92.1	92.5	92.5	9
FTM	KNN	90.08	89.98	90.67	90.04	6
	SVM	86.23	86.03	87.27	86.14	6
	RF	91.01	90.31	91.42	90.41	10
UrbanLandCover	KNN	90.01	89.26	91.81	89.81	7
	SVM	86.61	85.41	87.89	86.3	7
	RF	89.63	84.62	88.35	86.67	10
GLIOMA	KNN	90	87.33	88.33	90	8
	SVM	93.33	92.42	93.75	93.75	8
	RF	90	81	83.75	85	6
CLL_SUB_1111	KNN	88.83	88.3	91.29	88.67	8
	SVM	83	81.74	86.59	82.89	8
	RF	87.04	82.29	83.76	83.56	10

continued on following page

Table 2. Continued

Datasets	Classifier	CA	F1 Score	Precision	Recall	SF
QSAR_B	KNN	87.63	87.55	88.44	87.63	9
	SVM	84.7	84.68	84.88	84.7	9
	RF	88.06	88.08	88.02	87.28	10
ThoracisSurgery	KNN	85.12	83.85	81.32	82.75	7
	SVM	74.5	72.71	69.33	77	7
	RF	86.88	84.43	87.09	85	10
SCADI	KNN	84.29	79.59	72.92	75.67	8
	SVM	84.29	74.82	73.67	78.17	6
	RF	84.29	74.15	74.58	77.5	3
CBSMR	KNN	82.49	82.42	83.02	82.5	6
	SVM	78.12	77.77	80.25	78.18	6
	RF	78.95	77.62	80.59	77.08	6
warpAR10P	KNN	79.23	74.6	73.03	80	4
	SVM	68.46	61.67	61.2	68.5	4
	RF	72.31	63.87	61.75	71	4
ORL	KNN	72.75	67.6	65.23	72.75	10
	SVM	73	67.92	65.5	73	10
	RF	82	73.95	71.88	80.25	10
Average	KNN	89.81	88.54	88.38	89.38	7.39
	SVM	87.55	85.95	86.39	87.5	7.5
	RF	89.82	86.32	87.21	88.05	8

Overall, the K-NN classifier performs better in terms of performance evaluation compared to SVM and RF classifiers. In our proposed MLWFS method, the number of SF is minimum, average, and maximum, which are three, eight, and ten, respectively.

Tables 3 to 5 present a comparison of other methods. The dataset names are presented in the first column. The second to ninth columns present CA obtained using FS methods, namely MLWFS, FSRRW, MIFS, JMI, mRMR, CIFE, MRI, DCSF, and CWJR, and the tenth column presents the Number of SF using the MLWFS method. Eight well-known FS methods (FSRRW, MIFS, JMI, mRMR, CIFE, MRI, DCSF, and CWJR) are used as a comparison for the MLWFS method, these eight FS methods are referred to as "other methods.". The W/L in the last row of each table displays the win/loss scores. A win denotes that the proposed method MLWFS method outperforms other methods "better or equally," while a loss indicates that the proposed method MLWFS method is better than the other methods "lesser than." Each row's highest CA is shown in bold font.

Table 3 and Figure 2 show that MLWFS method outperforms other FS methods on thirteen datasets with respect to CA. The MLWFS method produces slightly lower CA for the Prostate_GE, warpPIE10P, CBSMR, warpAR10P, and ORL datasets. The FSRRW method outperforms compared to other methods in terms of CA on the Prostate_GE, warpPIE10P, and ORL datasets. When compared to other techniques, the MRI approach produces better CA on the CBSMR dataset. The CWJR method outperforms compared to other methods in terms of CA on the CBSMR datasets. When compared against eighteen datasets, the ALLAML dataset obtains a maximum CA of 99% in

Table 3. Comparison of averages CA (in %) evaluated by the K-NN of MLWFS method and other methods (in %) (the highest values are shown in bold)

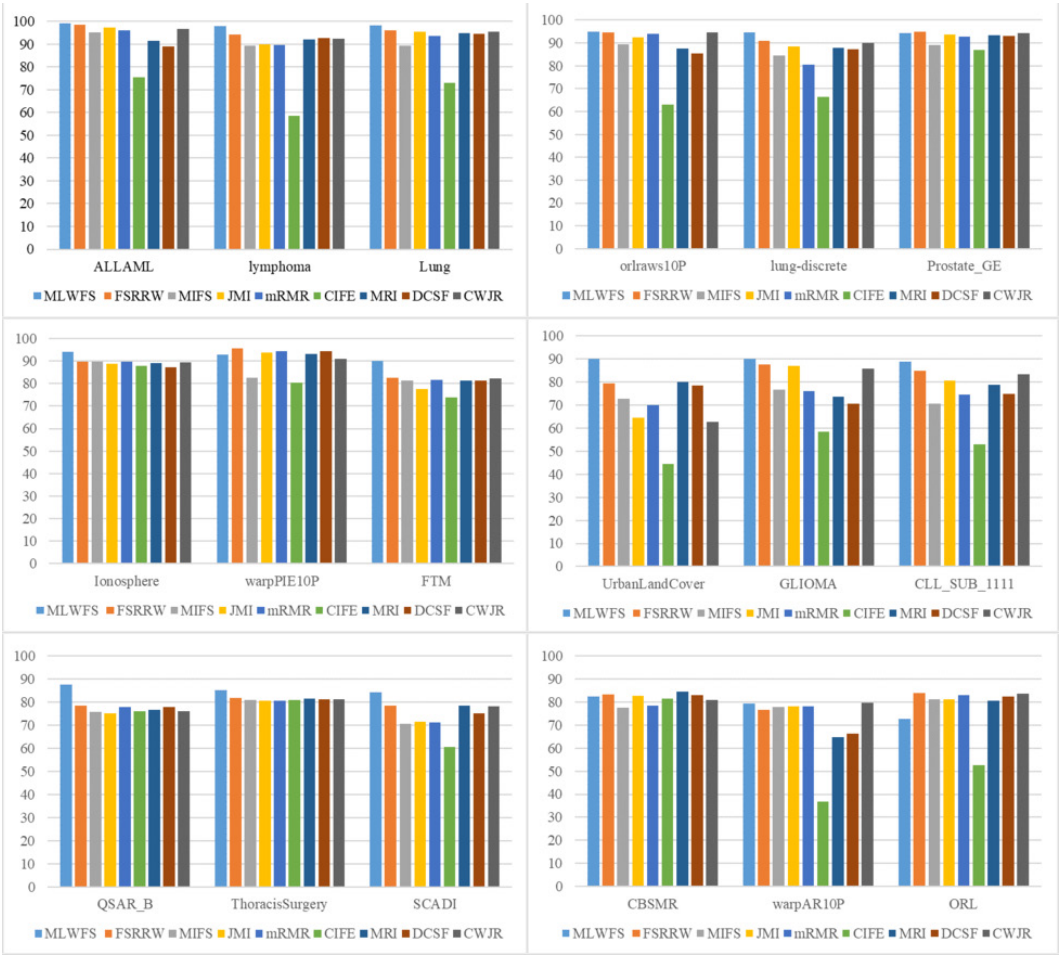
Datasets	MLWFS	FSRRW	MIFS	JMI	mRMR	CIFE	MRI	DCSF	CWJR	SF
ALLAML	**99**	98.39	95.03	97.42	95.95	75.53	91.58	89.02	96.78	5
lymphoma	**97.83**	94.33	89.21	89.9	89.68	58.56	92.01	92.62	92.3	10
Lung	**98.27**	96.04	89.17	95.56	93.74	73.13	94.93	94.37	95.36	7
orlraws10P	**95**	94.6	89.5	92.6	93.93	63.17	87.6	85.37	94.7	7
lung-discrete	**94.62**	90.84	84.42	88.37	80.46	66.28	87.87	87.3	90.08	9
Prostate_GE	94.27	**95.04**	89.01	93.64	92.9	86.99	93.33	93.17	94.31	6
Ionosphere	**94.22**	89.92	89.83	88.92	89.91	87.82	89.29	87.34	89.45	7
warpPIE10P	92.86	**95.58**	82.63	93.85	94.38	80.41	93.14	94.59	91.12	9
FTM	**90.08**	82.62	81.45	77.76	81.54	73.76	81.36	81.42	82.15	6
UrbanLandCover	**90.01**	79.27	72.74	64.48	70.06	44.51	80.03	78.36	62.74	7
GLIOMA	**90**	87.59	76.75	86.97	76.05	58.34	73.72	70.67	85.7	8
CLL_SUB_1111	**88.83**	84.77	70.7	80.6	74.54	52.99	78.92	74.99	83.41	8
QSAR_B	**87.63**	78.46	75.67	75.14	77.94	75.96	76.86	77.79	76.21	9
ThoracisSurgery	**85.12**	81.84	81.01	80.75	80.57	80.94	81.49	81.35	81.16	7
SCADI	**84.29**	78.69	70.61	71.43	71.12	60.67	78.67	75.1	78.29	8
CBSMR	82.49	83.15	77.42	82.67	78.38	81.39	**84.48**	82.9	80.73	6
warpAR10P	79.23	76.52	77.72	78.2	78.06	36.87	64.62	66.2	**79.77**	4
ORL	72.75	**83.78**	81.03	81.12	82.92	52.69	80.63	82.43	83.54	10
Average	**89.81**	87.30	81.88	84.41	83.45	67.22	83.92	83.06	85.43	7.39
W/L		14/4	17/1	15/3	16/2	18/0	15/3	15/3	15/3	

the MLWFS method, while on ORL datasets, it achieves a minimum CA of 72.75%. The MLWFS method outperforms other methods, achieving an average CA of 89.81% across all datasets. The MLWFS method determines the highest eighteen wins when compared to the CIFE method and the least fourteen wins when compared to the FSRRW method, with respect to the W/L ratio. This refers to the hardest competitor of MLWFS method is FSRRW in the K-NN classifier.

According to Table 4 and Figure 3, we achieved the better CA on fourteen datasets using the proposed MLWFS method when compared to other methods. The MLWFS method produces slightly less CA for the Thoracis Surgery, SCADI, CBSMR, and warpAR10P datasets. In the ThoracisSurgery and SCADI datasets, the CWJR method outperforms other methods in terms of CA. In comparison to other methods, the FSRRW method produces better CA in the CBSMR and warpAR10P datasets. When compared to eighteen datasets, the lymphoma dataset in the MLWFS method achieves a maximum CA of 99.76%, while the warpAR10P dataset obtains a minimum CA of 68.46%. The MLWFS method achieves an average CA of 87.55% across all datasets with only an average of eight features, and the average CA of the MLWFS method is better compared to other methods. In W/L ratio, MLWFS method directs maximum win, i.e., sixteen in comparison with the MIFS, mRMR CIFE, and DCSF methods, and minimum win, i.e., fourteen in comparison with the FSRRW, and CWJR methods.

Table 5 and Figure 4 show that MLWFS method outperforms other FS methods on twelve datasets with respect to CA. The MLWFS method produces slightly lower CA for the ALLAML, warpPIE10P, SCADI, and ORL datasets. The FSRRW method outperforms MLWFS method in terms of CA on the ALLAML, warpPIE10P, SCADI, warpAR10P, and ORL datasets. The CWJR method outperforms

Figure 2. Comparison of average CA (in %) evaluated by the K-NN of MLWFS method and other methods



MLWFS method in terms of CA on the orlraws10P datasets. When compared against eighteen datasets, the ALLAML dataset obtains a maximum CA of 99% in the MLWFS method, while on warpAR10P datasets, it achieves a minimum CA of 72.31%. The MLWFS method outperforms other methods, achieving an average CA of 89.82% across all datasets. The MLWFS method determines the highest eighteen wins when compared to the CIFE method and the least thirteen wins when compared to the FSRRW method concerning the W/L ratio.

The statistical win/loss statistics of MLWFS method compared to other methods are effectively shown in Figures 5-7. The MLWFS method mainly contains the win frequencies of the K-NN, SVM, and RF classifiers. This study shows that the MLWFS method is more effective than other methods and that certain features provide more relevant information, which could significantly boost the classification efficacy.

Table 6 summarizes the comparative analysis between MLWFS method and other methods. The best CA among all classifiers in other methods is shown in the first block; the best CA among all classifiers in the MLWFS method is shown in the second column; and the number of SF in the MLWFS method is shown in the third block. Figure 8 summarizes the comparative analysis between MLWFS method and other methods.

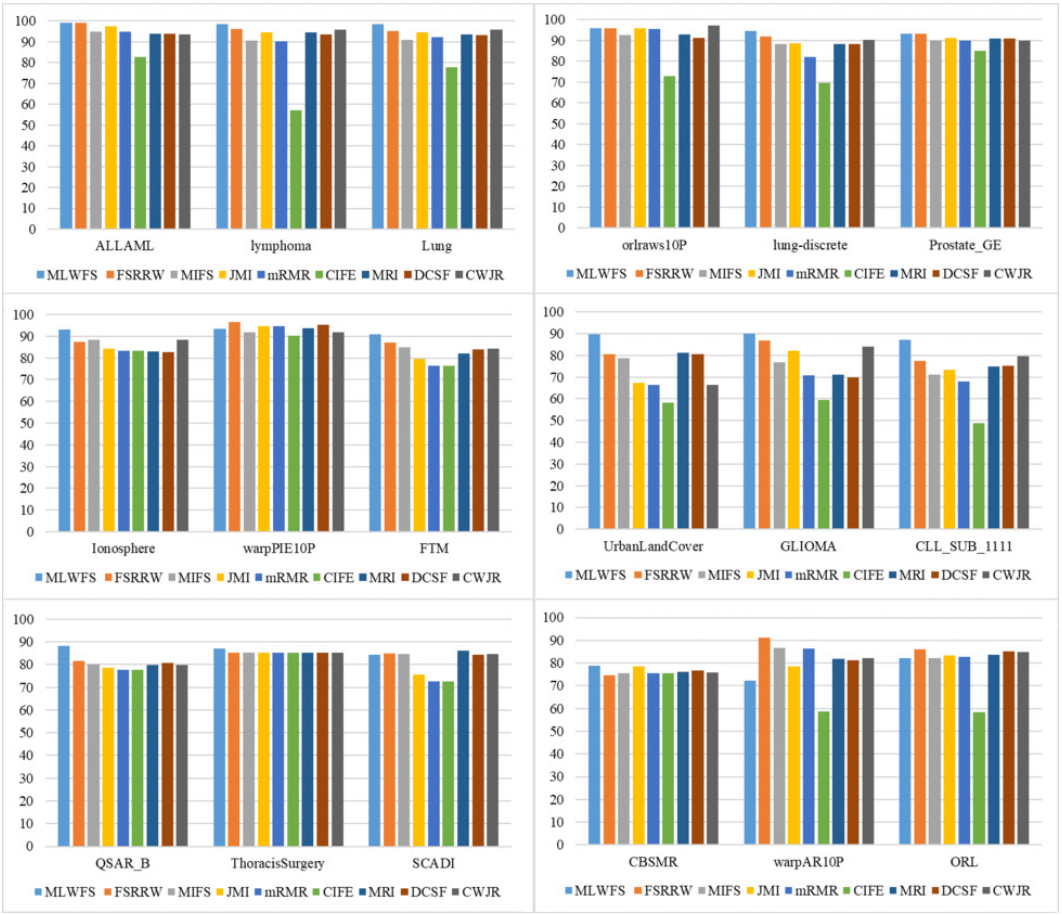
Table 4. Comparison of average CA (in %) evaluated by the SVM of MLWFS method and other methods

Datasets	MLWFS	FSRRW	MIFS	JMI	mRMR	CIFE	MRI	DCSF	CWJR	SF
ALLAML	**97**	96.58	94.25	95.39	94.37	82.04	92.47	91.47	95.31	4
lymphoma	**99.76**	86.73	77.69	84.26	78.1	55.06	84.47	84.09	86.44	10
Lung	**99.14**	92.5	88.11	91.91	89.84	77.79	90.83	89.37	93.03	10
orlraws10P	**96**	93.67	82.2	92.83	87.1	53.9	84.03	78.57	95.27	10
lung-discrete	**95.33**	78.5	69.5	76.43	66.89	58.45	75.73	76.12	77.85	9
Prostate_GE	**92.18**	91.76	88.5	91.23	89.64	85.48	92.62	90.53	91.49	6
Ionosphere	**93.78**	92.61	91.29	92.51	91.54	91.03	91.56	92.16	92	7
warpPIE10P	**90.48**	88.66	86.56	86.8	86.78	74.22	88.31	88.54	87.32	8
FTM	**86.23**	83.26	81.9	78.23	82.48	74.89	81.82	82.3	83.12	6
UrbanLandCover	**86.61**	79.68	75.96	68.37	70.79	51.5	79.05	78.19	66.04	7
GLIOMA	**93.33**	82.44	80.2	79.96	65.19	48.72	62.57	62.66	79.88	8
CLL_SUB_1111	**83.0**	81.9	70.3	76.53	71.23	53.14	76.43	73.68	80.06	8
QSAR_B	**84.7**	81.22	80.02	78.58	81.24	77.9	79.08	80.86	80.77	9
ThoracisSurgery	74.5	82.09	83.54	81.71	82.63	81.49	81.57	82.07	**83.47**	7
SCADI	84.29	84.35	77.86	76.24	80.35	68.68	84.01	83.56	**85.28**	6
CBSMR	78.12	**79.45**	74.43	79.03	77.1	78.33	80.12	80.87	78.38	6
warpAR10P	68.46	**79.02**	71.86	76.63	77.08	38.8	68.95	64.97	78.37	4
ORL	**73**	69.79	61.95	67.46	67.19	39.18	62.64	63.41	69.78	10
Average	**87.55**	84.68	79.78	81.89	79.97	66.14	80.9	80.19	83.55	7.5
W/L		14/4	16/2	15/3	16/2	16/2	15/3	16/2	14/4	

Table 6 and Figure 8 clearly show that, in comparison to other methods, the feature set selected by the MLWFS method is more significant. Using only three SF results in a remarkable CA of 84.29%, in the SCADI dataset; demonstrating the usefulness of this method. This CA result is deemed satisfactory, showcasing the robust performance achieved with a minimal set of features. While other methods select up to thirty features for each dataset, the MLWFS method selects a maximum of ten features. In the MLWFS method, an average CA of90.95% is better compared to other methods, 89.97% of the average CA.

Correlation heatmap (CH) illustrating correlation matrices for specific features from eighteen datasets are shown in Figures 9 to 26, i.e. one figure for one dataset respectively. In each figure, a feature is checked for its relationship with other features. If two features are highly correlated then their correlation value tends towards 1 otherwise it decreases towards 0. Lower the correlation value, lower will be the redundancy. It means that two features are redundant if their correlation is close to 1. The diagonal values in each of the figures 9-26 is 1 as the feature is related to itself so it will be obviously highly or perfectly correlated. Other than the diagonal values, the remaining values are considered for checking the redundancy among features. The details of CHs for each of the eighteen datasets are as follows where we mention only figure numbers, the names of concerned datasets are provided in the figures their selves. The CH for SF is shown in Figure 9 of the ALLAML dataset. Figure 9 clearly shows that the highest correlation value among the SFs is 0.5 (except for the diagonal values). This indicates that the SF is non-redundant. Similarly, Figures 13, 15, 20, and 22 show the CH for SF. In these figures, it is noted that in each dataset, feature correlation values are less than 0.50, indicating that the SF consists mostly of non-redundant features. Figures 10, 11, 19, 23, 24, and

Figure 3. Comparison of average CA (in %) evaluated by the RF of MLWFS method and other methods



26 show the CH for SF. In these figures, and it is noted that in each dataset taken in these figures, feature correlation values are less than 0.70, and in some of these figures, the maximum correlation values are less than 0.5, indicating that the SF consists mostly of non-redundant features. Figures 12, 14, and 16 show the CH for SF. In these figures, the maximum feature correlation values are less than 0.80, indicating that the SF consists mostly of non-redundant features. Figures 17 and 21 show the CH for SF. In these figures, the maximum feature correlation values are less than 0.90 indicating slightly redundancy among features. For the warpAR10P dataset, Figure 25 shows that the highest correlation value is 0.96, and for the UrbanLandCover dataset, Figure 18 shows that the highest correlation value is 0.99, indicating that some redundant features are still present and can be removed by applying some other technique as in future work. Overall, Figures 9 to 26 demonstrate that the SFs mostly consist of non-redundant features. The overall findings highlight the superiority of the MLWFS method over the other eight methods, and the MLWFS method has the best ability to select relevant and non-redundant features that significantly improves the classifier's efficiency.

There is always a tradeoff between the evaluation parameters (like CA) and the number of selected features (dimensionality reduction). It is quite challenging to optimize both objectives. Multi-objective Evolutionary techniques(Deb, 2011) can be a good solution for such scenario.

An important point to remind here is that the present work aims to find the feature set with the highest classification accuracy but with least cardinality (dimensionality) being of wrapper approach.

Table 5. Comparison of average CA (in %) evaluated by the RF of MLWFS method and other methods

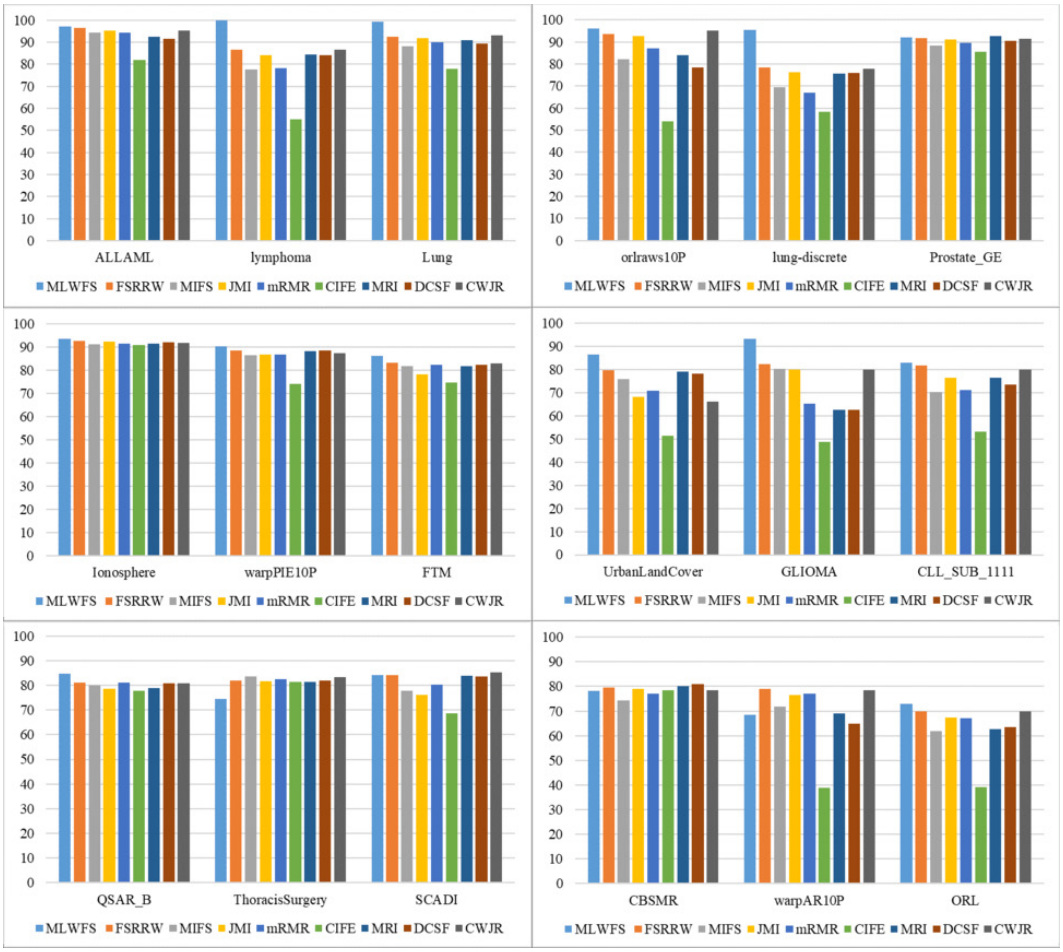
Datasets	MLWFS	FSRRW	MIFS	JMI	mRMR	CIFE	MRI	DCSF	CWJR	SF
ALLAML	99	**99.02**	94.69	97.44	94.92	82.73	93.89	93.81	93.64	7
lymphoma	**98.57**	96.28	90.58	94.45	90.09	56.98	94.37	93.56	95.73	8
Lung	**98.56**	95.14	90.79	94.36	92.04	77.61	93.46	93.34	95.75	10
orlraws10P	96	95.93	92.5	95.73	95.43	72.97	92.9	91.23	**97.07**	7
lung-discrete	**94.67**	91.96	88.29	88.77	81.98	69.5	88.22	88.29	90.17	9
Prostate_GE	**93.27**	93.2	89.83	91.38	89.81	85.02	90.88	90.76	90.08	8
Ionosphere	**93.11**	87.54	88.26	84.37	83.19	83.19	83	82.75	88.24	7
warpPIE10P	93.33	**96.52**	91.88	94.71	94.71	90.14	93.86	95.36	91.89	9
FTM	**91.01**	87.19	84.9	79.62	76.5	76.5	81.96	83.81	84.16	10
UrbanLandCover	**89.63**	80.56	78.82	67.5	66.57	58.24	81.2	80.46	66.57	10
GLIOMA	**90**	86.72	76.82	82.24	70.7	59.53	71	70.01	84.01	6
CLL_SUB_1111	**87.04**	77.36	70.98	73.31	67.93	48.77	74.83	75.16	79.7	10
QSAR_B	**88.06**	81.48	80.1	78.48	77.76	77.76	79.73	80.83	79.72	10
ThoracisSurgery	**86.88**	85.11	85.11	85.11	85.11	85.11	85.11	85.11	85.11	10
SCADI	84.29	**84.79**	84.7	75.59	72.67	72.67	86.03	84.33	84.52	3
CBSMR	**78.95**	74.65	75.44	78.61	75.54	75.54	76.15	76.64	75.69	6
warpAR10P	72.31	**91.03**	86.6	78.6	86.41	58.8	81.72	81.11	82.22	4
ORL	82	**86.02**	82.03	83.38	82.88	58.55	83.64	85.22	84.86	10
Average	**89.82**	88.36	85.13	84.65	82.46	71.65	85.11	85.1	86.06	8
W/L		13/5	15/3	15/3	15/3	18/0	14/4	14/4	14/4	

In the case of filter approach, the strength of each individual feature is determined using some of the methods like Laplacian Score, Variance, Correlation (A. Saxena et al., 2017) etc. In the present paper, the wrapper approach is adopted to select a random group of features and calculate the CA collectively of this group instead of doing it for each individual feature of this group. Thus, finding CA is very important to select a group and simultaneously picking the group having minimum number of features with highest CA from amongst all other groups (wrappers).

CONCLUSION AND FUTURE WORK

Feature selection is an important component of machine learning. In order to develop a robust model for prediction of class/labels of a dataset with high accuracy, it is expected to deal with only relevant features required to decide the class of the dataset for any unseen pattern. Irrelevant, noisy and redundant features can not only increase the time and space complexity of the classification process but also affect adversely the classification results. This paper aims to remove the harmful features and select only essential features to predict classification labels correctly. The method named MLWFS is divided into two stages. The first stage ensures assembling useful features in the reduced feature subset. The Second stage further eliminates the features one by one until a sufficient and acceptable accuracy of prediction is achieved by the model. The wrapper method based MLWFS model first divides the original dataset into sub folds of equal and small size of features. The sub fold yielding highest CA is separated; the same fold is reduced using sequential backward selection by eliminating one feature at a time. All the reduced feature sets in decreasing size are tested for

Figure 4. Comparison of average CA (in %) evaluated by the RF of MLWFS method and other methods



CAs and the feature set with the highest CA is saved. Three classifiers KNN, SVM, and RF are used one by one for validation purposes. It is observed that compared to the larger-dimensional original dataset, the size of the features finally selected is much smaller. The proposed model produced CAs while testing on unknown patterns, that are either superior to or very similar to eight comparable techniques on various types of datasets as reported in the literature. Eighteen datasets were used to evaluate the performance of the proposed MLWFS method for a fair and broader acceptance of the model. Thus, the objectives of obtaining a small but effective feature set to be used for predicting classes at a high accuracy is met up with satisfaction.

One of the challenges in adopting the model is to test the model's performance on datasets with thousands or millions of features. The internal structure of the dataset also needs to be preserved while eliminating features. Further, the model computes CAs using many kinds of deep learning and machine learning tools with a well-chosen set of parameter values, the computational costs can be an issue especially on an average machine with moderate configuration. Some multi-objective optimization methods can be applied to satisfy more than one conflicting requirements (objectives) and that can be a future scope of the proposed model for researchers.

Figure 5. Comparison of MLWFS method and other methods W/L ratio with the K-NN classifiers

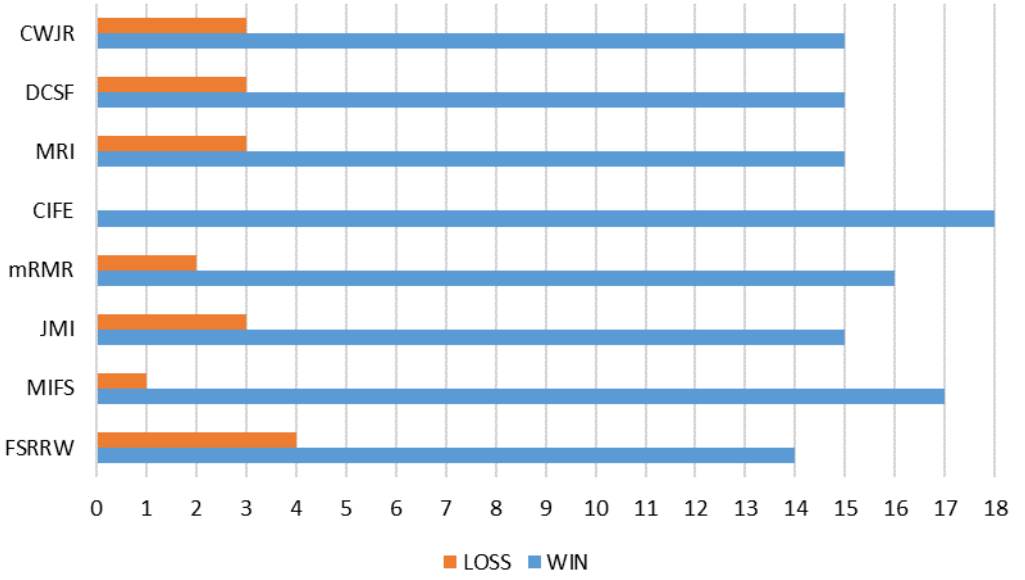
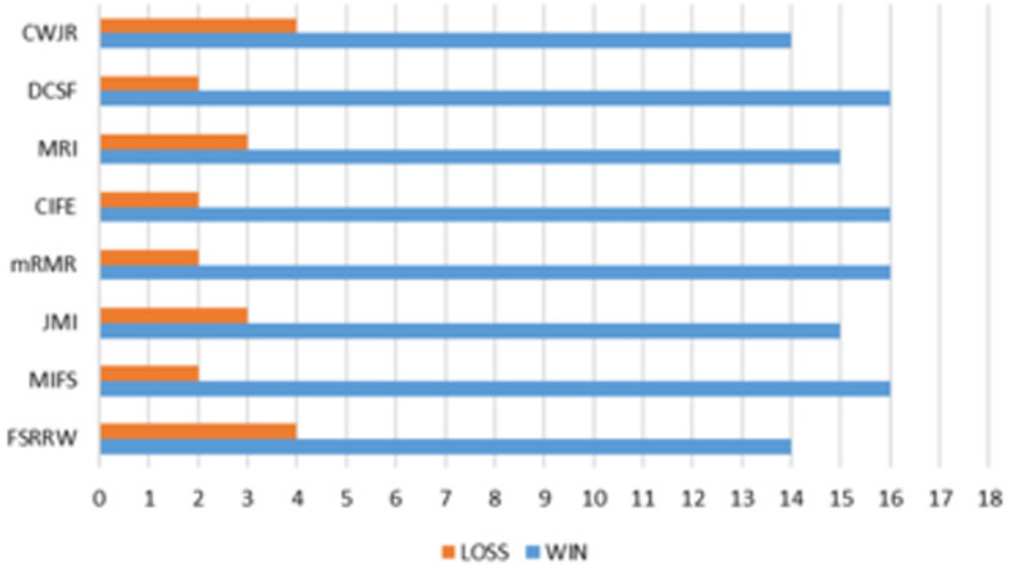


Figure 6. Comparison of MLWFS method and other methods W/L ratio with the SVM classifiers

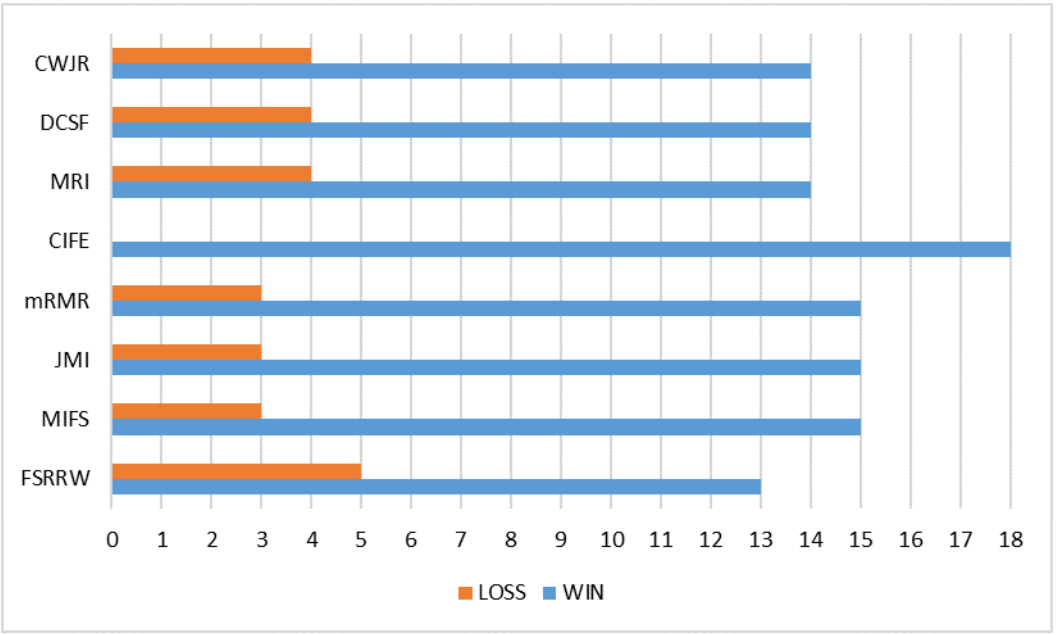


COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Figure 7. Comparison of MLWFS method and other methods W/L ratio with the RF classifiers



Informed consent: Informed consent was obtained from all individual participants included in the study.

Data availability statements: Data is available from the authors upon reasonable request.

Funding: Not Applicable

Process Dates

07, 2024

This manuscript was initially received for consideration for the journal on 03/15/2024, revisions were received for the manuscript following the double-anonymized peer review on 07/12/2024, the manuscript was formally accepted on 07/10/2024, and the manuscript was finalized for publication on 07/24/2024

Corresponding Author

Correspondence should be addressed to Amit Saxena; amitsaxena65@rediffmail.com

Table 6. Summary of comparative analysis between MLWFS method and other methods

Datasets	Best CA among all classifiers in other methods		Best CA among all classifiers in MLWFS method		SF
	CA	Classifier	CA	Classifier	
ALLAML	**99.02**	SVM	99	KNN	5
lymphoma	96.28	SVM	**99.76**	SVM	10
Lung	96.04	KNN	**99.14**	SVM	10
orlraws10P	**97.07**	SVM	96	RF	7
lung-discrete	91.96	SVM	**95.33**	SVM	9
Prostate_GE	**95.04**	KNN	94.27	KNN	6
Ionosphere	92.61	RF	**94.22**	KNN	7
warpPIE10P	**96.52**	SVM	93.33	RF	9
FTM	87.19	KNN	**91.01**	RF	10
UrbanLandCover	81.2	RF	**90.01**	KNN	7
GLIOMA	87.59	KNN	**93.33**	SVM	8
CLL_SUB_1111	84.77	KNN	**88.83**	KNN	8
QSAR_B	81.48	SVM	**88.06**	RF	10
ThoracisSurgery	85.11	SVM	**86.88**	RF	10
SCADI	**86.03**	SVM	84.29	RF	3
CBSMR	**84.48**	KNN	82.49	KNN	6
warpAR10P	**91.03**	SVM	79.23	KNN	4
ORL	**86.02**	SVM	82	RF	10
Average	89.97		**90.95**		7.72

Figure 8. Summarizes the comparative analysis between MLWFS method and other methods

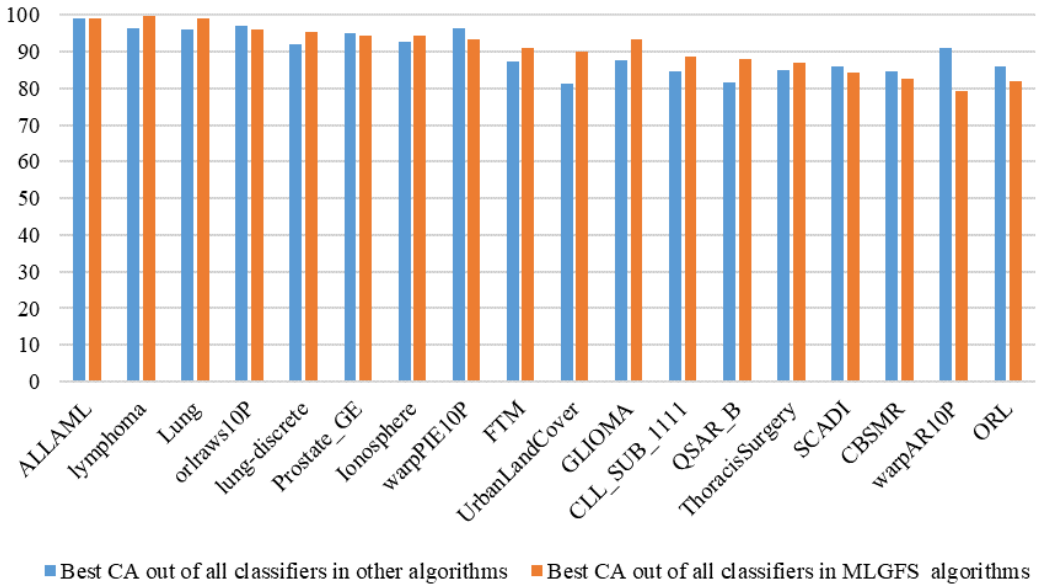


Figure 9. Heat map representation of a CH on a selected feature of ALLAML dataset

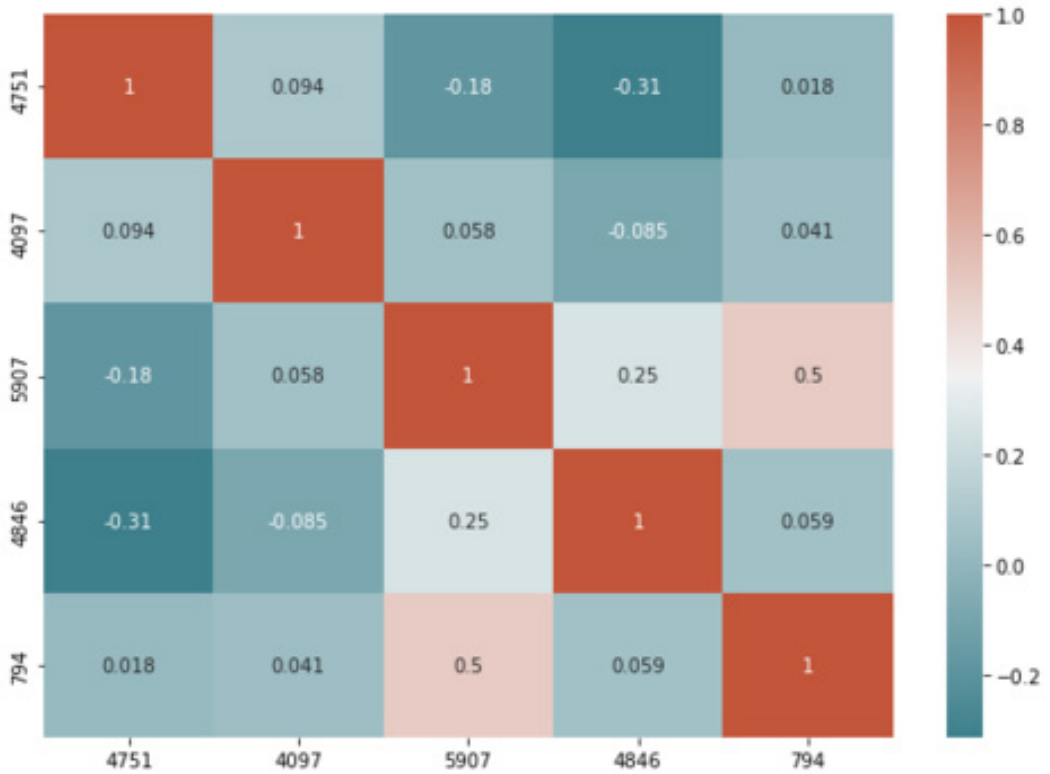


Figure 10. Heat map representation of a CH on a selected feature of lymphoma dataset

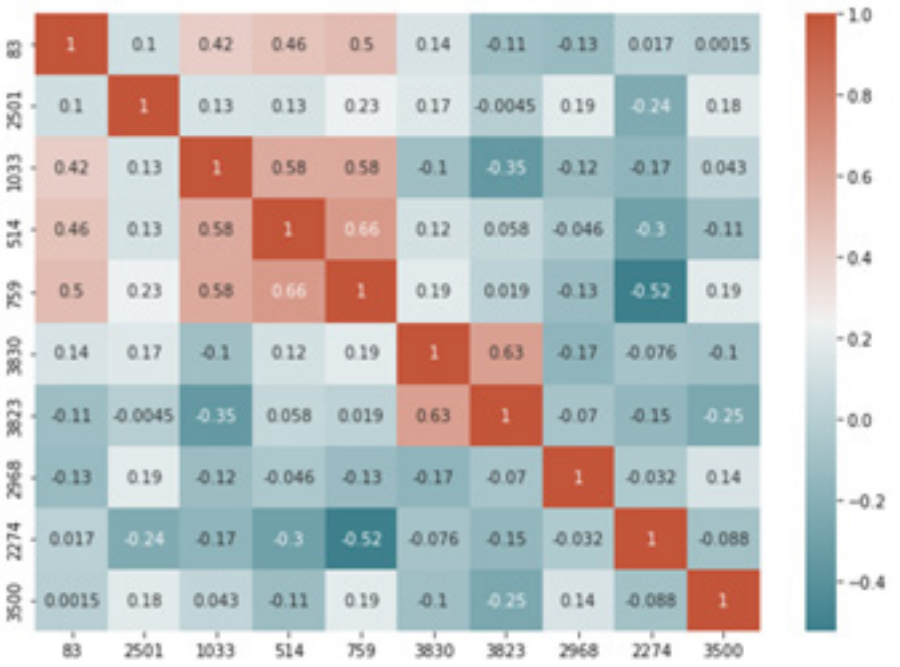


Figure 11. Heat map representation of a CH on a selected feature of lung dataset

Figure 12. Heat map representation of a CH on a selected feature of orlraws10P dataset

Figure 13. Heat map representation of a CH on a selected feature of lung-discrete dataset

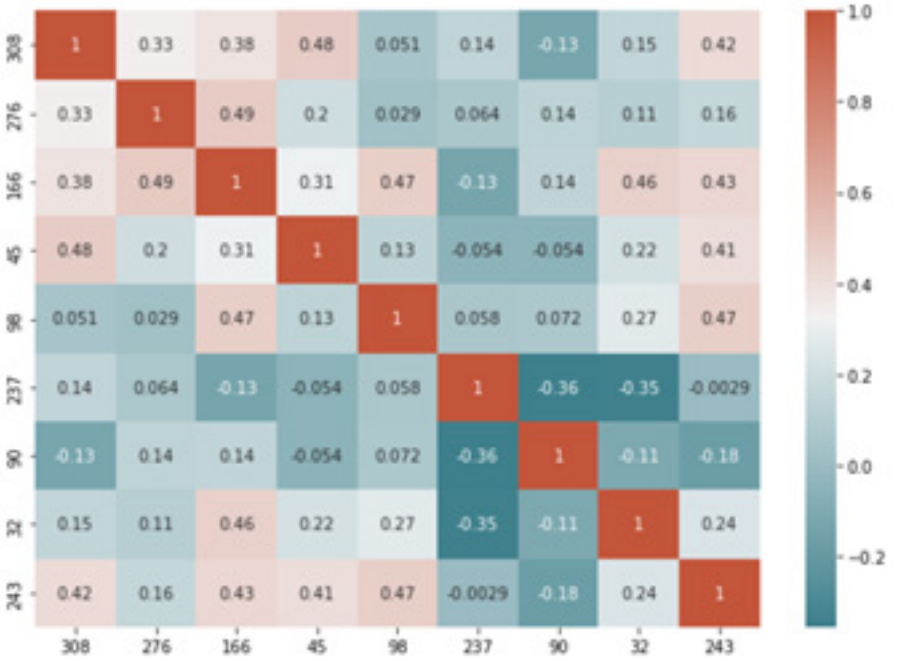


Figure 14. Heat map representation of a CH on a selected feature of prostate_GE dataset

Figure 15. Heat map representation of a CH on a selected feature of Ionosphere dataset

Figure 16. Heat map representation of a CH on a selected feature of warpPIE10P dataset

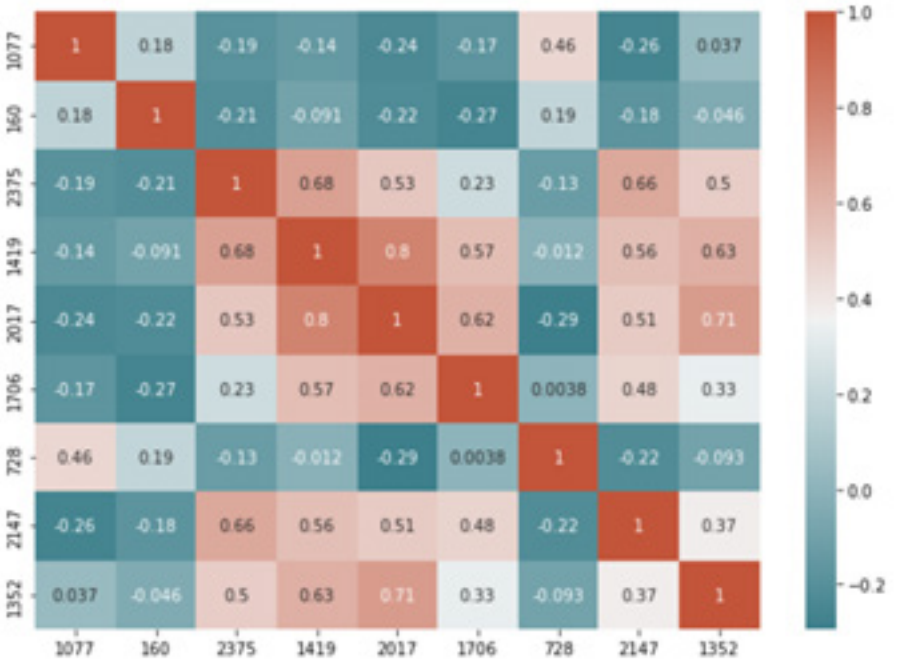


Figure 17. Heat map representation of a CH on a selected feature of FTM dataset

Figure 18. Heat map representation of a CH on a selected feature of UrbanLandCover dataset

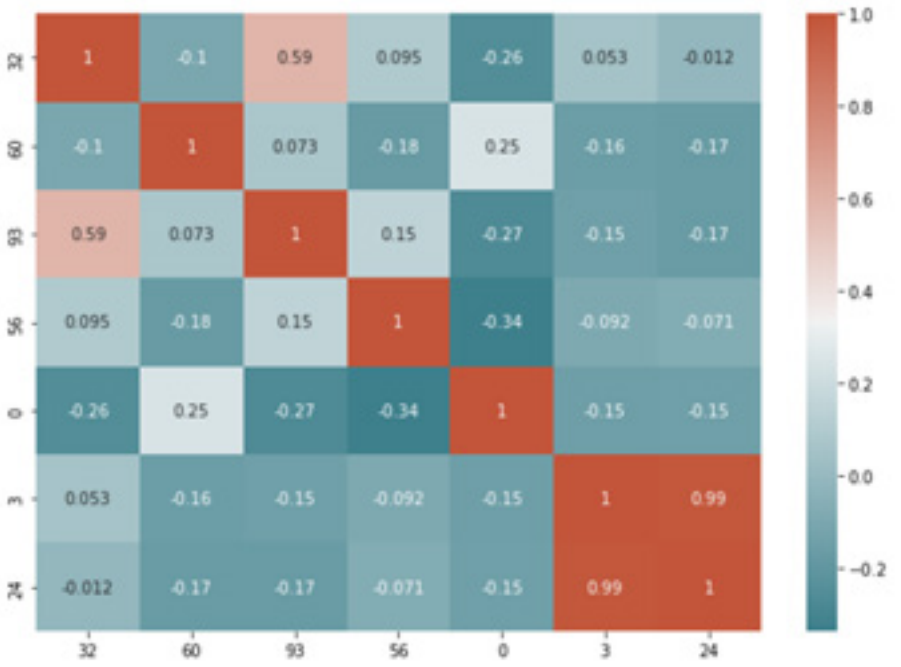


Figure 19. Heat map representation of a CH on a selected feature of GLIOMA dataset

Figure 20. Heat map representation of a CH on a selected feature of CLL_SUB_1111 dataset

Figure 21. Heat map representation of a CH on a selected feature of QSAR_B dataset

Figure 22. Heat map representation of a CH on a selected feature of ThoracisSurgery dataset

Figure 23. Heat map representation of a CH on a selected feature of SCADI dataset

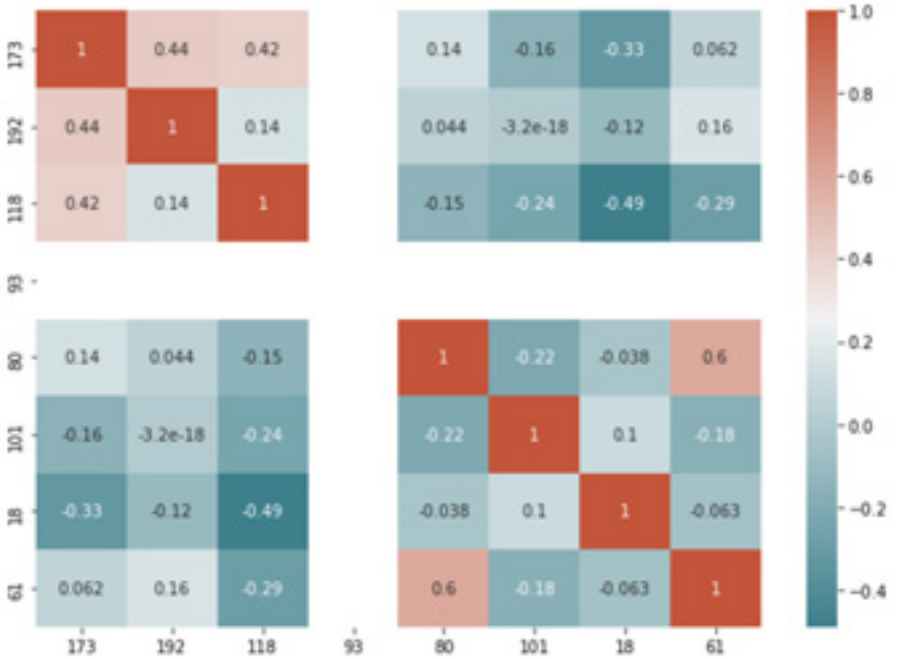


Figure 24. Heat map representation of a CH on a selected feature of CBSMR dataset

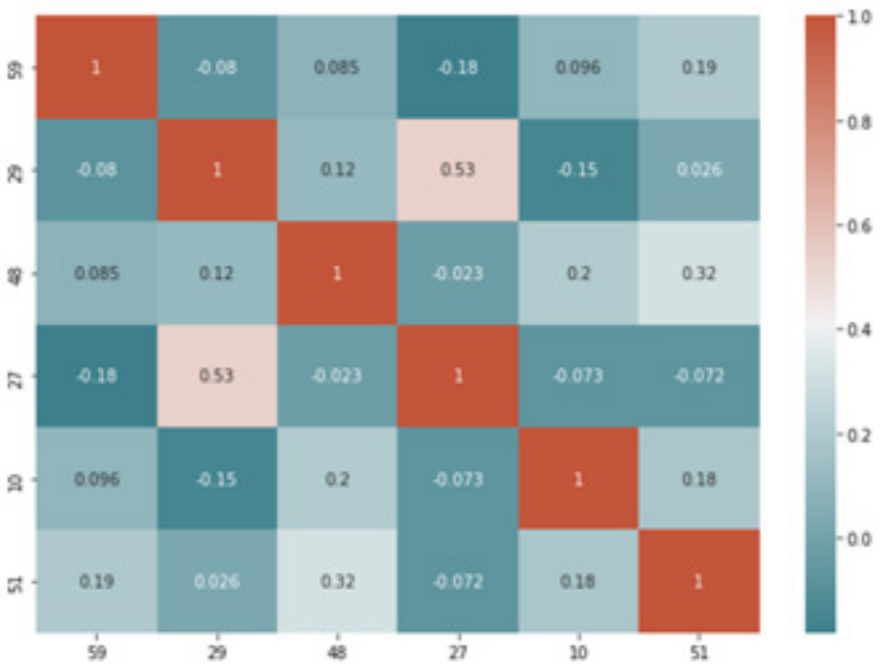


Figure 25. Heat map representation of a CH on a selected feature of warpAR10P

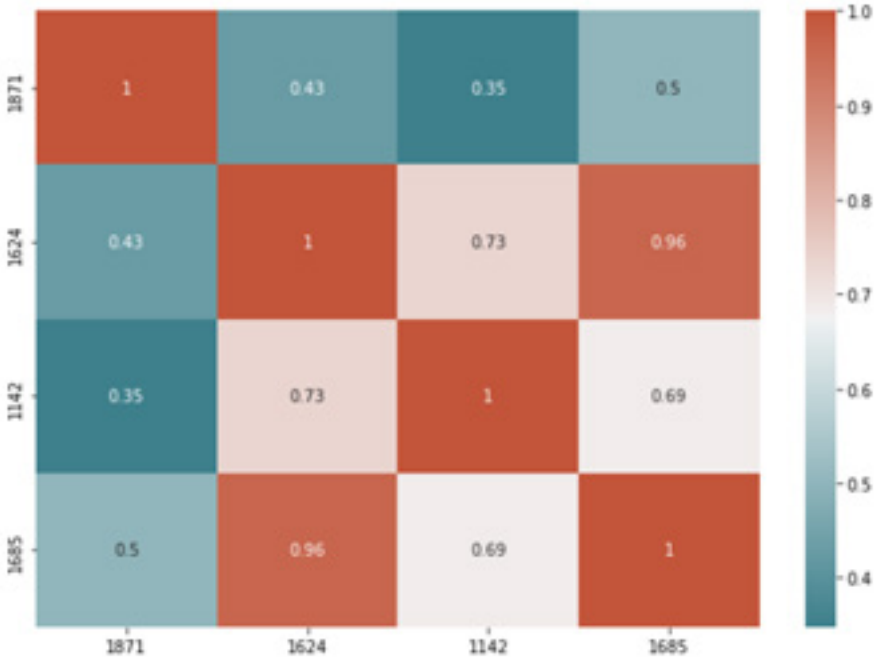


Figure 26. Heat map representation of a CH on a selected feature of ORL dataset

REFERENCES

Allheeib, N., Kanwal, S., & Alamri, S. (2023). An Intelligent Heart Disease Prediction Framework Using Machine Learning and Deep Learning Techniques. [IJDWM]. *International Journal of Data Warehousing and Mining*, *19*(1), 1–24. DOI:10.4018/IJDWM.333862

Arizona State University Library. (n.d.). Retrieved May 20, 2024, from https://lib.asu.edu/

Balasubramaniam, S., Vijesh Joe, C., Manthiramoorthy, C., & Satheesh Kumar, K. (2024). ReliefF based feature selection and Gradient Squirrel search Algorithm enabled Deep Maxout Network for detection of heart disease. *Biomedical Signal Processing and Control*, *87*, 105446. Advance online publication. DOI:10.1016/j.bspc.2023.105446

Battiti, R. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks*, *5*(4), 537–550. DOI:10.1109/72.298224 PMID:18267827

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. DOI:10.1006/jmps.1999.1279 PMID:10733860

Chakraborty, D., & Pal, N. R. (2008). Selecting useful groups of features in a connectionist framework. *IEEE Transactions on Neural Networks*, *19*(3), 381–396. Advance online publication. DOI:10.1109/TNN.2007.910730 PMID:18334359

Chang, S. W., Abdul-Kareem, S., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics*, *14*(1), 170. Advance online publication. DOI:10.1186/1471-2105-14-170 PMID:23725313

Chugh, D., Mittal, H., Saxena, A., Chauhan, R., Yafi, E., & Prasad, M. (2023). Augmentation of Densest Subgraph Finding Unsupervised Feature Selection Using Shared Nearest Neighbor Clustering. *Algorithms 2023, Vol. 16, Page 28, 16*(1), 28. DOI:10.3390/a16010028

Deb, K. (2011). Multi-objective Optimisation Using Evolutionary Algorithms: An Introduction. In *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*. DOI:10.1007/978-0-85729-652-8_1

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. UCI Machine Learning Repository.

El-Kenawy, E. S. M., Mirjalili, S., Alassery, F., Zhang, Y. D., Eid, M. M., El-Mashad, S. Y., Aloyaydi, B. A., Ibrahim, A., & Abdelhamid, A. A. (2022). Novel Meta-Heuristic Algorithm for Feature Selection, Unconstrained Functions and Engineering Problems. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 40536–40555. DOI:10.1109/ACCESS.2022.3166901

Gao, W., Hu, L., Zhang, P., & Wang, F. (2018). Feature selection by integrating two groups of feature evaluation criteria. *Expert Systems with Applications*, *110*, 11–19. DOI:10.1016/j.eswa.2018.05.029

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. DOI:10.1016/j.patrec.2010.03.014

Ghosh, S., Dutta, A., Ghosh, S., & Chaudhuri, A. K. (2024). Enhancing Prediction Precision and Reliability in Cervical, Lung, and Breast Cancer Diagnosis. *Driving Smart Medical Diagnosis Through AI-Powered Technologies and Applications*, 49–80.

Iguyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. In *Journal of Machine Learning Research* (Vol. 3).

Kamber, M., Han, J., & Pei, J. (2006). *Data Mining Concepts and Techniques* (2nd ed.). Morgan Kaufmann.

Kwak, N., & Choi, C. H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, *13*(1), 143–159. DOI:10.1109/72.977291 PMID:18244416

Liao, B., Jiang, Y., Liang, W., Zhu, W., Cai, L., & Cao, Z. (2014). Gene selection using locality sensitive Laplacian score. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(6), 1146–1156. Advance online publication. DOI:10.1109/TCBB.2014.2328334 PMID:26357051

Lin, D., & Tang, X. (2006). Conditional infomax learning: An integrated framework for feature extraction and fusion. *European Conference on Computer Vision*, 68–82. DOI:10.1007/11744023_6

Nesamani, L., & S. Nirmala Sigirtha, R. (2023). Predictive Modeling for Classification of Breast Cancer Dataset Using Feature Selection Techniques. *Research Anthology on Medical Informatics in Breast and Cervical Cancer*, 166–177.

Patel, D., & Saxena, A. K. (2022). A Novel Sequential Feature Selection in Large Dimensional Datasets. *Chhattisgarh Journal of Science and Technology, 19*(4), 137–144. https://new.ggu.ac.in

Patel, D., Saxena, A. K., Laha, S., & Ansari, G. M. (2022). A Novel Scheme for Feature Selection Using Filter Approach. *Proceedings of the 2022 7th International Conference on Computing, Communication and Security, ICCCS 2022 and 2022 4th International Conference on Big Data and Computational Intelligence, ICBDCI 2022*, 1–4. DOI:10.1109/ICCCS55188.2022.10079604

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. DOI:10.1109/TPAMI.2005.159 PMID:16119262

Rani, P., Kumar, R., Jain, A., & Chawla, S. K. (2021). A hybrid approach for feature selection based on genetic algorithm and recursive feature elimination. *International Journal of Information System Modeling and Design*, *12*(2), 17–38. Advance online publication. DOI:10.4018/IJISMD.2021040102

Saxena, A., Chugh, D., Mittal, H., Sajid, M., Chauhan, R., Yafi, E., Cao, J., & Prasad, M. (2022). A Novel Unsupervised Feature Selection Approach Using Genetic Algorithm on Partitioned Data. *Advances in Artificial Intelligence and Machine Learning*, *2*(4), 500–515. Advance online publication. DOI:10.54364/AAIML.2022.1134

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681. DOI:10.1016/j.neucom.2017.06.053

Saxena, A., & Wang, J. (2010). Dimensionality reduction with unsupervised feature selection and applying non-Euclidean norms for classification accuracy. [IJDWM]. *International Journal of Data Warehousing and Mining*, *6*(2), 22–40. DOI:10.4018/jdwm.2010040102

Saxena, A., Wang, J., & Sintunavarat, W. (2021). An empirical study on initializing centroid in k-means clustering for feature selection. [IJSSCI]. *International Journal of Software Science and Computational Intelligence*, *13*(1), 1–16. DOI:10.4018/IJSSCI.2021010101

Saxena, A. K., & Dubey, V. K. (2015). A Survey on feature selection algorithms. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(4), 1895–1899. DOI:10.17762/ijritcc2321-8169.150431

Saxena, A. K., Dubey, V. K., & Wang, J. (2017). Hybrid feature selection methods for high-dimensional multi-class datasets. *International Journal of Data Mining. Modelling and Management*, *9*(4), 315–339.

Setiono, R., & Liu, H. (1997). Neural-network feature selector. *IEEE Transactions on Neural Networks*, *8*(3), 654–662. DOI:10.1109/72.572104 PMID:18255668

Wang, J., Wei, J. M., Yang, Z., & Wang, S. Q. (2017). Feature selection by maximizing independent classification information. *IEEE Transactions on Knowledge and Data Engineering*, *29*(4), 828–841. DOI:10.1109/TKDE.2017.2650906

Wang, Q. (2022). Support Vector Machine Algorithm in Machine Learning. *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 750–756. DOI:10.1109/ICAICA54878.2022.9844516

Wang, W., Lu, L., & Wei, W. (2022). A Novel Supervised Filter Feature Selection Method Based on Gaussian Probability Density for Fault Diagnosis of Permanent Magnet DC Motors. *Sensors (Basel)*, *22*(19), 7121. Advance online publication. DOI:10.3390/s22197121 PMID:36236219

Yang, H., & John, M. (1999). Data visualization and feature selection: New algorithms for nongaussian data. *Advances in Neural Information Processing Systems*, ●●●, 12.

Yang, S., Jian, H., Ding, Z., Hongyuan, Z., & Giles, C. L. (2007). IKNN: Informative K-nearest neighbor pattern classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4702 LNAI*. DOI:10.1007/978-3-540-74976-9_25

Zadeh, L. A. (1988). Fuzzy Logic. *Computer*, *21*(4), 83–93. DOI:10.1109/2.53

Zeng, H., & Cheung, Y. M. (2011). Feature selection and kernel learning for local learning-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(8), 1532–1547. DOI:10.1109/TPAMI.2010.215 PMID:21135434

Zhang, P., Gao, W., Hu, J., & Li, Y. (2021). A conditional-weight joint relevance metric for feature relevancy term. *Engineering Applications of Artificial Intelligence*, *106*, 104481. DOI:10.1016/j.engappai.2021.104481

Zhao, S., Wang, M., Ma, S., & Cui, Q. (2022). A feature selection method via relevant-redundant weight. *Expert Systems with Applications*, *207*, 117923. DOI:10.1016/j.eswa.2022.117923

Zhu, Q., Wang, C., Jin, W., Ren, J., & Yu, X. (2024). Deep Transfer Learning Based on LSTM Model for Reservoir Flood Forecasting. *International Journal of Data Warehousing and Mining*, *20*(1), 1–17. DOI:10.4018/IJDWM.338912

Zhu, X., & Li, L. (2023). Estimating the Number of Clusters in High-Dimensional Large Datasets. [IJDWM]. *International Journal of Data Warehousing and Mining*, *19*(2), 1–14. DOI:10.4018/IJDWM.316142