A novel filter-wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets[☆]

Tushaar Gangavarapu[a,*], Nagamma Patil[a]

[a] *Department of Information Technology,*
National Institute of Technology Karnataka, Surathkal, Mangaluru, India

Abstract

The predictive accuracy of high-dimensional biomedical datasets is often dwindled by many irrelevant and redundant molecular disease diagnosis features. Dimensionality reduction aims at finding a feature subspace that preserves the predictive accuracy while eliminating noise and curtailing the high computational cost of training. The applicability of a particular feature selection technique is heavily reliant on the ability of that technique to match the problem structure and to capture the inherent patterns in the data. In this paper, we propose a novel filter-wrapper hybrid ensemble feature selection approach based on the weighted occurrence frequency and the penalty scheme, to obtain the most discriminative and instructive feature subspace. The proposed approach engenders an optimal feature subspace by greedily combining the feature subspaces obtained from various predetermined base feature selection techniques. Furthermore, the base feature subspaces are penalized based on specific performance dependent penalty parameters. We leverage effective heuristic search strategies including the greedy parameter-wise optimization and the Genetic Algorithm (GA) to optimize the subspace ensembling process. The effectiveness, robustness, and flexibility of the proposed hybrid greedy ensemble approach in comparison with the base feature selection techniques, and prolific filter and state-of-the-art wrapper methods are justified by empirical analysis on three distinct high-dimensional biomedical datasets. Experimental validation revealed that the proposed greedy approach, when optimized using GA, outperformed the selected base feature selection techniques by 4.17%–15.14% in terms of the

[*] Corresponding author

Email addresses: `tushaargvsg45@gmail.com` (Tushaar Gangavarapu), `nagammapatil@nitk.edu.in` (Nagamma Patil)

URL: `http://infotech.nitk.ac.in/faculty/nagamma-patil` (Nagamma Patil)

prediction accuracy.

Keywords: Biomedical data, Genetic algorithm, Greedy ensemble, High-dimensional data, Hybrid feature selection, Parameter optimization.

1. Introduction

The need for efficient analytical methodologies in healthcare applications has led to an unparalleled development in the field of biomedicine and bioinformatics over the past decade [41, 62]. Research in these fields frequently encounters supervised classification of disease data (e.g., microarray gene data, lung cancer data, and others) [41, 14, 2]. The advances in wet-technology are increasing the volume of data with a large number of dimensions [33]. For example, the profiling of microarray gene [33, 10, 34] aims at measuring the expression levels of tens of thousands of genes over tens of thousands of features. Over the last decade, owing to the availability of high dimensional biomedical data, numerous feature selection methods have become viable processes that provide robust data in low-dimensional spaces [55, 25]. In the sense of high dimensional data, standard statistical methods suffer from the curse of dimensionality [8, 30] signifying a drastic rise in the classification error and computational complexity. This makes it mandatory to use a feature subspace before the classification is undertaken [50, 54, 28]. Therefore, feature selection does not represent the very aim of data analysis but is instead a preliminary step to finding the most informative and discriminative feature subset that optimally represents the given data.

Dimensionality reduction can aid in the provision of better insights to understanding causal relationships, reduce computational complexities, and engender more reliable estimates [61, 12]. There are numerous methods to achieve dimensionality reduction including feature selection based on information gain and minimum Redundancy Maximum Relevance (mRMR). Real-world datasets vary, implying that no single feature selection technique is best suited for all the datasets [18]. The effectiveness of a feature selection technique depends on its ability to match the problem structure and maintain only those features that describe the inherent patterns within the data. The selection of such a technique is usually heuristic and intuition based. The challenge to the machine learner is the selection of a feature selection technique that works best for a given dataset. A naive approach to achieve the same would be to select a technique from the set of predetermined techniques that results in the best performance. This approach is computationally very expensive and infeasible. An alternative approach would be to perform a heuristic selection which is further explored using evolutionary computational algorithms [29]. This approach requires an investment of an arbitrary amount of computation time, and the actual optimal solution and the obtained solution might not converge for a limited number of iterations [1, 22].

Early works [15, 17, 69] aimed at using filter approaches to determine the most optimal feature subspace. These approaches are heavily reliant on the

Table 1: Comparison with state-of-the-art works in feature selection.

	Masood *et al.* 2017	Dong *et al.* 2018	Tu *et al.* 2019	**This work**
Feature selection type	Wrapper and filter-wrapper hybrid	Heuristic search	Heuristic search	Filter-wrapper hybrid with heuristic search
Approaches used	Wrapper and hybrid approaches	Hybrid GA with granularity	Multi-strategy ensemble grey wolf optimizer	Hybrid greedy ensemble selection approach
Ensembling	–	–	3 search strategies	5 filter-wrapper hybrid methods
Search strategy	–	Bottom-up search of ordered feature list	Grey wolf optimizer	Correlation-guided greedy feature search
Parameter optimization	–	GA	Disperse foraging strategy	Greedy-parameter wise optimization and GA
Max. #features	21 (×4 sensors)	12,582	60	2,352
Corresponding #samples	28 (occupants)	72	208	10,015
Corresponding #classes	4	10	3	7
Algorithms used	RIG[a] and ELM	–	–	RF[b], BDT[c], and KNN[d]

[a] *Relative Information Gain*; [b] *Random Forest*; [c] *Bagged Decision Tree*; [d] *K-Nearest Neighbors.*

correlation between the features and are independent of the classifier which limits their accuracy. Min *et al.* [45] developed a backtracking and heuristic search algorithm to search for optimal feature subspaces. The authors showed that the performance of the evolutionary computing algorithm was similar to backtracking but with lower computational time. More recently, Masood *et al.* [42] proposed wrapper and hybrid algorithms which used an incremental search on an ordered set of features and Extreme Learning Machine (ELM) classifier to select the best feature subspace. A hybrid genetic algorithm with feature granulation was developed by Dong *et al.* [16] for feature selection. Tu *et al.* [64] proposed a multi-strategy ensemble grey wolf optimizer with three search strategies and demonstrated its effectiveness in selecting optimal features. From the existing literature, it is evident that hybrid and wrapper feature selection methods overcome the limitations of filter methods. Moreover, evolutionary computing algorithms are widely used in feature selection because of their population-based mechanism and domain adaptability.

Although most state-of-the-art methods aim at effectively determining an optimal feature subspace, they are either extremely data specific or utilize heuristic-based approaches requiring an arbitrary amount of time with no guarantee on their convergence. Furthermore, heuristic search methods using swarm intelligence seldom use correlation measures to guide the search process. To address these problems, we propose a novel ensemble selection approach that uses a set of (five) predetermined feature selection techniques on a representative sample of the dataset to generate multiple feature subspaces. These subspaces are then evaluated using (three) different supervised classification algorithms. The features in the subspaces obtained from the set of chosen feature selection techniques are then penalized based on the evaluation scores, to form an optimal subset of features selected greedily. The penalty factors that affect the choice of features in the hybrid subset are optimized using the greedy parameter-wise optimization and the Genetic Algorithm (GA). Moreover, the penalty factors are modeled in a way that is aimed at selecting smaller and most instructive feature subspace. Since the feature selection is performed on a sample of the dataset as opposed to the entire dataset, the computational cost is relatively low. Furthermore, the values of the penalty factors that affect the choice of the features in the final feature subspace are heuristically determined, limiting the problem of algorithmic convergence occurring when the features themselves are heuristically selected. Table 1 shows the comparison of this work with the existing state-of-the-art methods in effective feature selection. The key contributions of this work are summarized below:

- Design of a filter-wrapper hybrid ensemble selection approach that kindles an optimal feature subspace by greedily combining the subspaces generated by various predetermined feature selection techniques based on specific performance dependent penalty parameters.

- Leveraging heuristic search strategies such as greedy parameter-wise optimization and GA to determine the optimal values of the penalty factors which affect how different feature subspaces are ensembled to engender an optimal feature subspace.

- We present detailed benchmarking results of our hybrid greedy ensemble feature selection approach on three distinct high-dimensional biomedical datasets. Our experimental results indicate the efficiency and robustness of the proposed approach over the base feature selection methods, and other prolific filter and wrapper methods.

The remainder of the paper is structured as follows: Section 2 provides an overview of the existing works and reviews their evaluation approaches, advantages, and limitations. Section 3 presents the statistics of the datasets used and addresses the fundamentals of the utilized feature selection algorithms, classification algorithms, and GA. The proposed greedy methodology is presented in Section 4 and the same is evaluated empirically in Section 5. In Section 6, a sensitivity analysis is presented to assess the performance of the results. Finally, Section 7 concludes this paper with highlights on future research possibilities.

2. Related work

An extensive body of research on the effective determination of most descriptive feature subspace is available in the literature [60, 3]. This section provides an extensive review of a few significant dimensionality reduction approaches to provide an overview of the existing state-of-the-art methods built on large biomedical datasets.

Feature selection approaches can be categorized into four categories including filter, wrapper, embedded, and hybrid models. In the field of biomedicine, feature selection is widely used in sequence analysis (signal analysis and content analysis) [31] and microarray analysis. Sequence analysis aims at the determination of the sequence (e.g., carbohydrates, proteins, and others), its fragmentation, and its interpretation. Apart from the features that represent amino acid or nucleotide, many other features resulting from the combinations of these building blocks can be derived. Since most of these features are redundant or irrelevant, feature selection techniques are mandatory to derive a subset of relevant features [55]. Moreover, most features are extracted from a sequence where adjacent positions in the sequence hold most dependencies. Early works [56, 4] developed and used interpolated Markov model which used the interpolation between various Markov model's orders to deal with the limited number of samples of small sizes. The model was further extended to deal with non-adjacent dependencies by using feature subset sampling with undersampling of majority class. These previous works showed significant performance improvement using Support Vector Machines (SVM) with full undersampling and feature selection.

A more trending area of research is the microarray analysis, where structural elements such as splice sites, Translation Initiation Sites (TIS) are modeled as classification problems [55]. Microarray analysis uses gene expression profiling of tissues or cell samples to determine which combination of genes are turned on. Microarray datasets pose challenges to modeling due to their low samples-to-dimensions ratio [2]. Li and Yen [35] proposed an optimization based on multiobjective binary biogeography (filter approach), with SVM classifier, and evaluated the computational complexity of their approach on multiple datasets. Liao *et al.* [37] used a filter method of selecting genes based on locality-sensitive Laplacian scoring scheme, with SVM classifier. The authors evaluated their approach using a variety of datasets including Leukemia and Lung Cancer datasets. From the criterion of accuracy, it can be inferred that the early works which used filter-based feature selection techniques suffered from the limitation that the correlation measure used to assess the importance of features is classifier independent.

Wrapper, hybrid, and embedded approaches address the limitations of filter-based approaches. Sharma *et al.* [58] proposed a wrapper-based approach to select features based on null space linear discriminant analysis, with K-Nearest Neighbors (KNN), evaluated the approach using sensitivity analysis. Yu *et al.* [67] used sample weighting to select stable genes from microarray data using recursive feature elimination with SVM (wrapper approach). Liu *et al.* [38] proposed a hybrid feature selection approach that involved the usage of Bhattacharyya distance as the filter and fuzzy interactive self-organizing algorithm as the wrapper. Hajiloo *et al.* [23] proposed a hybrid method of rule-based classification using fuzzy SVM as the wrapper and signal-to-noise ratio as the filter. Masood *et al.* [42] presented wrapper and hybrid algorithms which used bottom-up incremental search on an ordered set of features. The authors used

Table 2: Summary of some key existing works.

Work	Feature selection approach	Classifier	Evaluation method
Liu *et al.* [39]	Wrapper approach based on the fuzzy interactive self-organizing data algorithm for sample selection	KNN, Linear SVM	Recognition rate, Area Under Curve (AUC)
Chang *et al.* [13]	Hybrid feature selection method using GA, ReliefF and adaptive neuro-fuzzy inference system	Neural net, SVM, Logistic regression, Fuzzy system	AUC, K-fold cross-validation
Liang *et al.* [36]	An embedded method with regularized multinomial sparse logistic regression with $L_{1/2}$ penalty	KNN	Leave one-out cross-validation
Song *et al.* [59]	Fast ensemble method that selects feature subsets using graph-theoretic clustering techniques	Naive Bayes, C4.5, IB1, Rule-based RIPPER[e]	Sensitivity, K-fold cross-validation, Runtime
Maulik and Chakraborty [43]	Filter approach that uses rough set based on prediction scheme using fuzzy preference for Cancer datasets	Transductive SVM	K-fold cross-validation
Yu *et al.* [68]	An ensemble semi-supervised clustering approach based on modified double selection for tumor clustering	K-means clustering	SD[f] and Mean of normalized MI[g]

[e] *Repeated Incremental Pruning to Produce Error Reduction*;
[f] *Standard Deviation*; [g] *Mutual Information.*

ELM for the incremental search and relative information gain for feature ranking. Gaafar *et al.* [19] proposed an ensemble selection method based on mRMR and GA, with KNN classifier for cancer diagnosis using microarray data. Table 2 reviews other related key existing works in the field of feature selection in biomedicine and bioinformatics. Although wrapper, hybrid, and embedded approaches overcome the limitations of filter-based approaches by ensuring lower error of the model, they are highly dataset and classifier specific. The challenge of the selection of a dimensionality reduction technique that effectively matches the problem structure is quite difficult and is often heuristic or intuition based.

More recently, metaheuristic search optimizations such as GA and particle swarm optimization have been applied to search for the optimal feature subspace. In comparison with the traditional methods, metaheuristic search approaches do not make assumptions about the search space (e.g., differentiable and linearly separable). Furthermore, the success of these swarm intelligence algorithms can be attributed to their versatility and flexibility, in the sense that they mimic the best features in nature. Dong *et al.* [16] proposed a hybrid genetic algorithm with feature granulation to select significant features. To improve the quality of the feature subset, the authors developed an improved neighborhood rough set approach with sample granulation. Tu *et al.* [64] proposed a multi-strategy ensemble grey wolf optimizer to select the feature subspace effectively. Furthermore, the authors used a parameter self-adjusting strategy to balance between exploitation and exploration of the feature space. Even though evolutionary computing algorithms overcome the limitations of the wrapper and hybrid methods, they are reliant on heuristic search requiring an arbitrary amount of time with no guarantee on the convergence of the obtained solution within the given number of iterations. Furthermore, these swarm intelligence algorithms seldom use correlation measure to guide the search process.

Our work advances the efforts of these previous state-of-the-art methods by using a novel filter-wrapper hybrid ensemble feature selection approach that engenders an optimal feature subspace by greedily combining the subspaces generated from various predetermined feature selection techniques. Furthermore, the feature subspaces are penalized based on their evaluation scores with respect to the predetermined classifier(s). Since the feature selection is performed on a sample of the dataset as opposed to the entire dataset, the computational cost is relatively low. Moreover, the values of the penalty parameters are determined heuristically, limiting the convergence problem occurring when the features themselves are heuristically determined.

3. Materials and methods

The experimental data consists of three biomedical datasets which are first described. All the datasets used are split into three mutually and collectively independent homogeneous samples using stratified random sampling [49]. Stratified random sampling guarantees the adequate representation of all the classes in the data, maintaining homogeneity within stratum and heterogeneity between strata[1]. The feature selection methods used in greedily deriving the hybrid features are discussed, followed by the discussion of the classification algorithms used in the evaluation of the feature selection techniques. Finally, the genetic algorithm used in the optimization of the penalty parameters that are used in deriving the hybrid feature subspace is detailed.

3.1. Biomedical datasets

The main characteristics of the datasets used in this paper are tabulated in Table 3. The datasets chosen have a sufficient number of samples to aid in the creation of three stratified samples. Both balanced and imbalanced datasets are chosen for an unbiased evaluation of the proposed technique. Depending on the size of the dataset, further sampling of the strata can be performed.

[1] Proportionate allocation variant of the stratified random sampling is used in this paper.

Table 3: Overview of the datasets used.

Dataset	**Size**	**#Dim**	**#Classes (#samples per class)**
TIS [51]	13,375	927	2 (3,312/10,063)
Skin Cancer [63]	10,015	2,352	7 (327/514/1,099/115/6,705/142/1,113)
Seizure [5]	11,500	179	5 (2,300/2,300/2,300/2,300/2,300)

Translation Initiation Sites (TIS) dataset [51] is extracted from the genome sequences of a selected set of vertebrates that were extracted from the GenBank [9]. The process involves finding the site at which the translation of mRNA to proteins initiates. The sequences are annotated with TIS (true or false). Since the dataset is comprised of processed DNA sequences, the TIS site is essentially an 'ATG[2]' sequence. The sequences are extracted to build a feature space by matching three nucleotides to one amino acid, counting the frequency of every amino acid and frequency of a pair of amino acids [32].

Skin Cancer dataset is extracted from the pixel information of `28×28 RGB` images of the Skin Cancer MNIST: HAM10000 (Human Against Machine with 10,000 training images) dataset [63]. The dataset comprises of a large collection of dermatoscopic images of the pigmented skin lesions. The dataset consists of all the important diagnostic categories of pigmented lesions including basal cell carcinoma, actinic keratosis and Bowen's disease, benign keratosis-like lesions, melanoma, dermatofibroma, vascular lesions, and melanocytic nevi.

Epileptic Seizure Recognition dataset [5] consists of five sets (A–E), each containing 100 single channel 23.6 seconds long electroencephalogram (EEG) segments. Each EEG segment is weakly stationary and is selected after a visual inspection for artifacts [20]. Surface EEG recordings of five healthy individuals form sets A (with eyes closed) and B (with eyes open). Segments measured from five patients in seizure-free intervals from opposite hemisphere's hippocampal formation and in the epileptogenic zone form sets C and D respectively. Seizure activity corresponding to all the recording sites showing the ictal activity forms set E.

3.2. Feature selection methods

The feature selection methods used to generate feature subspaces which are in turn used in the generation of the hybrid feature subspace are discussed in this section. Five feature selection techniques are used in this paper (four with feature ranking, one without feature ranking). The implementations available in Weka 3.8.3 [27] were used to implement all the predetermined feature selection methods.

[2] Adenine(A), Thymine (T), and Guanine (G).

3.2.1. Information gain-based feature selection

Information gain-based feature selection (*igFeatureEval*) [6] evaluates the goodness (worth) of a feature by computing the Information Gain (IG) of a feature with respect to the target class. Concisely, IG measures the amount of information (in bits/Shannons) obtained to predict the target class by knowing the presence or absence of a feature. IG between a feature (f) and the target class is given by Equation 1, where $H(\cdot)$ represents the marginal entropy, and $H(\text{class}|f)$ measures the conditional entropy of f after observing the target class.

$$\text{IG}(f, \text{class}) = H(\text{class}) - H(\text{class}|f) \tag{1}$$

The *igFeatureEval* is a fast filter-based feature selection method. The selected features (based on the threshold) are ranked in the order of decreasing IG scores.

3.2.2. Correlation-based feature selection

Correlation-based feature selection (*corrFeatureEval*) [24] evaluates the goodness (worth) of a feature by computing the Pearson's (bi-variate) correlation (PCC) between the feature and the target class. Equation 2 gives the Pearson's correlation measure between a feature (f) and the target class, where $E[\cdot]$ represents the expected value, μ_x represents the mean of x, and σ_x represents the standard deviation of x.

$$\text{PCC}(f, \text{class}) = \frac{E[(f - \mu_f)(\text{class} - \mu_{\text{class}})]}{\sigma_f \sigma_{\text{class}}} \tag{2}$$

The *corrFeatureEval* is also a fast filter-based feature selection technique. The features selected (based on the threshold) are ranked in the decreasing order of PCC scores.

3.2.3. Correlation-based feature subset selection

Correlation-based feature subset selection (*cfsSubsetEval*) [24] considers the redundancy between the features and the individual predictive ability of features, to evaluate the goodness (worth) of a feature subset. Subsets with lower inter-correlation and high correlation with the target class are chosen. The worth of a feature subset S with k features is given by Equation 3, where $\mathcal{C}$ measures the relatedness of two variables (correlation, not necessarily Pearson's correlation or Spearman's ρ).

$$\text{Worth}(S_k) = \frac{\sum\limits_{f_i \in S_k} \mathcal{C}(f_i, \text{class})}{\sqrt{\sum\limits_{f_i \in S_k} \sum\limits_{f_j \in S_k - \{f_i\}} \mathcal{C}(f_i, f_j)}} \tag{3}$$

Symmetric uncertainty [65], an entropy-based measure of relatedness is used in this paper. Symmetric uncertainty between two variables X_i, X_j is given by

Equation 4, where $\mathrm{MI}(X_i, X_j)$ measures the mutual information between X_i, X_j and $H(\cdot)$ represents the marginal entropy.

$$\text{Uncertainty}(X_i, X_j) = 2 \cdot \frac{\mathrm{MI}(X_i, X_j)}{H(X_i) + H(X_j)} \tag{4}$$

The subspace of feature subsets is searched forward, starting with an empty feature subspace, by greedy hillclimbing with backtracking. Note that this search approach provides no feature ranking.

3.2.4. Minimum redundancy maximum relevance

Minimum Redundancy Maximum Relevance (*mRMR*) [52] is an incremental search method which integrates relevance and redundancy into a single objective function that aims at maximizing relevance and minimizing redundancy. The scoring function can combine redundancy and relevance as: 1) relevance−redundancy, which is Mutual Information Difference (MID) or 2) relevance/redundancy, which is Mutual Information Quotient (MIQ). The MID objective function (Φ) used to achieve *mRMR* is given by Equation 5, where $\mathrm{MI}(X_i, X_j)$ measures the mutual information between X_i, X_j.

$$\Phi = \frac{1}{|S_k|} \sum_{f_i \in S_k} \mathrm{MI}(f_i, \text{class}) - \frac{1}{|S_k|^2} \sum_{f_i, f_j \in S_k} \mathrm{MI}(f_i, f_j) \tag{5}$$

The *mRMR* approach is used with C4.5 decision trees (information gain). This feature selection approach ranks in decreasing order, the selected features (based on the threshold) based on *mRMR* scores.

3.2.5. OneR-based feature selection

OneR-based feature selection (*oneRFeatureEval*) [48] evaluates the worth of a feature by using OneR as the filter to select features, by recursive elimination. The OneR algorithm aims at deducing a rule that predicts the target class based on the given values of the features. The algorithm chooses the feature with more information and forms an entire rule based on that feature [7].

The *oneRFeatureEval* technique uses a rule to evaluate the usefulness of features. The selected features (based on the threshold) are ranked in the order of decreasing OneR rule scores.

3.3. Classification algorithms

Three classification algorithms from the existing literature including Random Forest (RF) [26], Bootstrap Aggregating with C4.5 Decision Trees (BDT) [11], and K-Nearest Neighbors (KNN) [53] are used in the evaluation of the predictive capabilities (in the form of accuracy scores) of the selected informative features. The implementations available in the Python Scikit-learn package were used to implement all the classifiers used in this paper.

Random Forest [26] is an ensemble learning technique that operates by constructing a number of decision trees while training. RF predicts the target class as the mode of the classes of individual trees. Bootstrap Aggregating (Bagging)

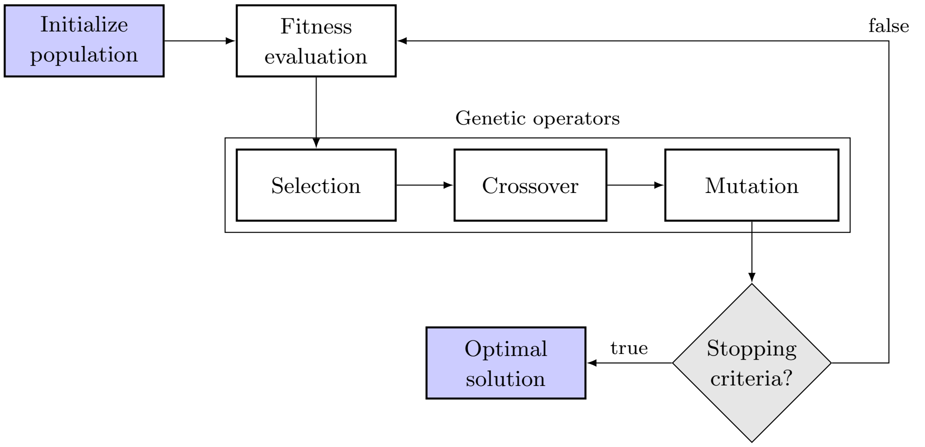

Figure 1: The flow of the genetic algorithm used in the optimization of the penalty parameters.

[11] is a machine learning ensemble meta-algorithm that improves the stability and accuracy of machine learning algorithms (here, decision trees). Bagging is a special case of the averaging technique. The method reduces variance and avoids overfitting. K-Nearest Neighbors [53] is an instance-based (lazy) learner that uses the majority vote of its k closest neighbors (distance between the data points gives a measure of their closeness) to determine the target class.

In this paper, RF classifier was used with 100 classification and regression trees of maximum depth 2. Furthermore, BDT classifier was used with an ensemble of 100 C4.5 decision trees as base estimators to obtain diversity among the base trees. Finally, 15 closest neighbors were considered (empirically determined using grid search) in this analysis, where closeness was weighted as the inverse of the distance between instances.

3.4. Genetic algorithm

The Genetic Algorithm (GA) [46] is a bio-inspired metaheuristic belonging to the class of evolutionary algorithms. Evolutionary algorithms are essentially swarm intelligence based heuristic search methods. The GA was implemented in Python 2.7.

In solving optimization problems, the idea of GA is that they start with a randomly generated population of individual solutions. The fitness function measures the quality of an individual in the population. Genetic operators aid in the conversion of one generation into the next one. The first operator is the selection operation which aims at selecting a portion of the existing population that breeds into the next generation. Individuals are selected based on their fitness scores, and higher fitness scores imply higher reproductive capability. Thus the fittest individuals are more likely to be selected while individuals with lower fitness scores may not be selected for reproduction [66]. The next step is to generate a new population using crossover (recombination) and mutation.

Table 4: Summary of the stratified samples used in hybrid feature selection.

Sample	Feature space	Summary
$\mathcal{S}_1$	#Features(dataset)	Feature selection using the chosen methods
$\mathcal{S}_2$	#Features($\mathcal{S}_1$)	Evaluation of the selected features and deriving the hybrid feature subspace
$\mathcal{S}_3$	Hybrid	Evaluation of the hybrid feature subspace

Crossover and mutation aim at replicating the randomness in any evolutionary process. For every new population produced, a pair of parent individuals are chosen for breeding and thus the child produced as a result of crossover and mutation shares many characteristics of the parents. The overall flow of GA is shown in Figure 1.

The genetic operators ensure that the subsequent generation population of chromosomes is different from the previous one. More often than not, the average fitness of the new generation will have increased, as only the best individuals from the previous generation are chosen for breeding, together with a small proportion of less fit individuals which ensures the genetic diversity within the pool of parents and thus ensures the genetic diversity within the children of the next generation.

In this paper, GA is used to determine the optimal values of the penalty factors that determine how different feature subspaces can be effectively combined. Thus the size of each chromosome is equal to the number of penalty parameters, and the population size is set to 50 to achieve optimal intensification and diversification within the given number of iterations. Furthermore, GA is implemented with roulette-wheel selection (fitness-proportionate selection) [21], a crossover factor (P_c) of 0.6, and a mutation factor (P_m) of 0.1 (for a maximum of 25 iterations).

4. Proposed novel filter-wrapper hybrid greedy ensemble approach for optimal feature selection

The proposed filter-wrapper hybrid feature selection approach uses three samples that are derived from the dataset using stratified random sampling [49]. Division of population into strata reduces the computational complexity and the sampling error. The first sample ($\mathcal{S}_1$) is used in selecting features from the predetermined feature selection technique(s) (five here). The feature space of the second sample ($\mathcal{S}_2$) is then reduced to the set of features selected using $\mathcal{S}_1$. Then, $\mathcal{S}_2$ is evaluated using the selected classifier(s) (three here). Based on the features selected in $\mathcal{S}_1$ and the accuracies obtained from the evaluation of $\mathcal{S}_2$, the feature subspace for the third sample ($\mathcal{S}_3$) is determined greedily using penalty parameters. Table 4 summarizes the use of stratified samples in hybrid feature selection. Figure 2 presents an overview of the proposed hybrid greedy ensemble approach and additional details of the same are presented below.

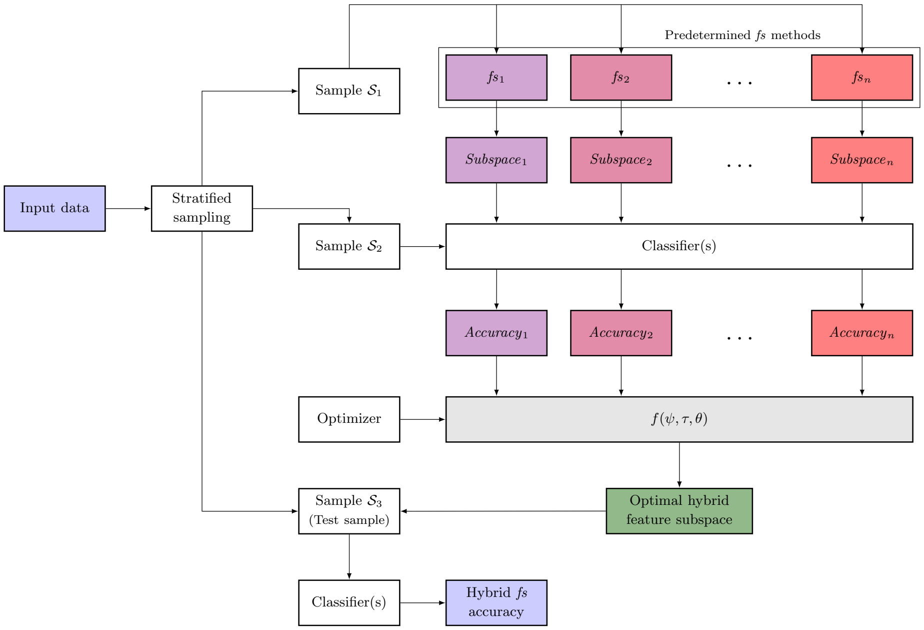

Figure 2: An overview of the proposed greedy hybrid ensemble feature selection modeled from a set of n (five here) predetermined feature selection methods (fs_is).

4.1. Scoring of features and feature selection methods

The feature subspaces obtained (one for every feature selection technique, five here) from $\mathcal{S}_1$ are used to derive the feature scores (*featScore*). The feature score of a feature f with respect to a feature selection method (with the feature subspace $\mathcal{FS}$ of length $|\mathcal{FS}|$) with rank ρ_f $(= index(f) + 1)$[3] is derived using the Equation 6.

$$featScore(f, \mathcal{FS}, \rho_f) = \begin{cases} \frac{|\mathcal{FS}| - \rho_f + 1}{|\mathcal{FS}|}, & f \in \text{ranked } \mathcal{FS} \\ \frac{1}{|\mathcal{FS}|}, & f \in \text{unranked } \mathcal{FS} \\ \frac{-1}{|\mathcal{FS}|}, & f \notin \mathcal{FS} \end{cases} \tag{6}$$

Feature scores can be positive or negative depending on the presence or absence of a feature in the given feature subspace. Also, it can be noted that *featScore* gives importance to selecting a lesser number of features, thus achieving the very aim of dimensionality reduction.

The accuracy scores obtained (one for every feature selection method, five here[4]) from $\mathcal{S}_2$ are used to derive the scores of the chosen base feature selection techniques (*accScore*). The *accScore* of a feature selection method m from the

[3] The rank ρ_f is only calculated when features in $\mathcal{FS}$ are ranked.

[4] The average accuracy of RF, BDT, and KNN is considered for simplicity.

set of chosen base feature selection methods $\mathcal{M}$ ($|\mathcal{M}| = 5$ here) with rank ρ_m (= $index(m) + 1$, $m \in \mathcal{M}$ ranked in the decreasing order of accuracies) is derived using the Equation 7.

$$accScore(m, \mathcal{M}, \rho_m) = \frac{|\mathcal{M}| - \rho_m + 1}{|\mathcal{M}|} \tag{7}$$

The accuracy scores are positive scores that ensure the selection of many features from those feature selection methods with higher accuracy. Furthermore, the accuracy scores are only positive to account for the possibility of feature selection from a base method with reduced dimensions and comparable but lower performance.

4.2. Penalty parameters for greedy ensembling of base feature subspaces

Penalty parameters facilitate performance dependent greedy selection of optimal features from the base selection techniques. They affect the extent of the impact of both the informativeness of the features and the classification accuracy of the base selection methods. The accuracy penalty (ψ) and feature penalty (τ) aim at penalizing the feature scores and accuracy scores respectively. The accuracy penalty aims at penalizing feature subspaces of the feature selection methods with $\mathcal{S}_2$ accuracy less than the $\mathcal{S}_2$ accuracy with the entire feature space. Accuracy penalty reduces the impact of the *accScore*. Concisely, the *accScore* becomes $accScore/\psi$.

Similarly, the feature penalty aims at increasing the negative impact of those features which are not selected by a feature selection technique, only when the $\mathcal{S}_2$ accuracy of the feature selection technique is greater than the $\mathcal{S}_2$ accuracy with the entire feature space. Concisely, the *featScore* becomes $featScore \times \tau$ (only for features with a negative *featScore*).

4.3. Overall feature scoring and hybrid feature selection

Overall scoring aims at combining the feature scores and accuracy scores to obtain the overall score which helps in the determination of the greedily selected most optimal hybrid feature subspace. Overall feature score of a feature f with respect to the given set of base selection methods $\mathcal{M}$ is given by the Equation 8.

$$overallScore(f, \mathcal{M}) = \sum_{m}^{\mathcal{M}} featScore(f) \times accScore(m) \tag{8}$$

By setting the decision parameter (threshold (θ)), we can filter the features based on their overall feature scores. The decision parameter aims at selection higher-ranked ($|\mathcal{FS}| - \rho_f + 1$) features from better performing base selection methods. The features thus selected form the greedily selected optimal hybrid feature subspace. Table 5 summarizes the scores and parameters used in the proposed greedy ensemble hybrid selection approach. Hereafter, the decision parameter (θ) is referred to as a penalty parameter as it affects the selection

Table 5: Summary of the scores and parameters used in hybrid feature selection.

Parameter	**Inference**	**Summary**
featScore	Positive or negative scores	Ensures that the hybrid feature subspace is formed from the features selected by the base methods
accScore	Positive scores	Ensures the selection of features from high accuracy feature selection methods
ψ	Reduces impact of *accScore*	Penalizes the selection of features from base methods (fs_i) with $\mathcal{S}_3$ accuracy < $\mathcal{S}_3$ accuracy with entire feature space (fs_{nil})
τ	Increases negative impact of *featScore*	Penalizes the selection of features not selected in a feature selection technique (only when method's $\mathcal{S}_3$ accuracy > $\mathcal{S}_3$ accuracy with entire feature space (fs_{nil}))
θ	Selection criteria	Determines the number of features to be selected based on the *overallScore*

process through overall scores which are penalized by both accuracy and feature penalties.

Algorithm 1 depicts the procedure to obtain the ensembled optimal hybrid feature subspace greedily from a given list of feature subsets ($\mathcal{FS}$_Lists), $\mathcal{S}_2$_Accuracies, $\mathcal{S}_2$ accuracy with the entire feature space ($\mathcal{S}_2$_All_Features_Acc), total number of features (totalFeat), accuracy rank list (ρ_m_List) and penalty parameters (ψ, τ, θ). Note that Algorithm 1 assumes that the penalty parameters are optimized prior to the greedy feature search.

4.4. Optimization of the penalty parameters

Optimization of the penalty parameters (ψ, τ, θ) used in the deduction of the optimal hybrid feature subspace is mandatory as these parameters determine the greedy selection of features from the base feature subspaces. We leverage heuristic search strategies such as greedy parameter-wise optimization and GA to obtain the best selection results. Compared to the traditional search strategies, heuristic approaches do not need any domain knowledge and do not make any assumptions about the search space. Furthermore, heuristic search strategies can reveal multiple optimal solutions in a single run. In greedy parameter-wise optimization, the penalty parameters are varied greedily starting with the accuracy penalty (ψ), followed by the feature penalty (τ), and finally the threshold (θ) to obtain the optimal values of these parameters. In GA, the initial generation of population solutions are generated by selecting random values in the predetermined range(s) (dataset dependent). The predetermined ranges were set with higher feature penalty range and comparably lower accuracy penalty range. Higher feature penalty range was set to heavily penalize those less discriminative features that were not selected by better performing base methods but were selected by methods with lower performance. Lower accuracy penalty

Algorithm 1: Proposed hybrid greedy ensemble feature selection

```
Input: $\mathcal{S}_2$_All_Features_Acc: Average accuracy with all features of $\mathcal{S}_2$,
       $\mathcal{S}_2$_Accuracies: List of average accuracies from predetermined methods,
       $\mathcal{FS}$_Lists: List of all selected feature subsets,
       totalFeat: Total number of features in the given dataset,
       $\rho_m$_List: List of ranks of predetermined selection methods,
       $\psi$: Accuracy penalty parameter,
       $\tau$: Feature penalty parameter,
       $\theta$: Selection threshold.
Output: Hybrid $\mathcal{FS}$: Greedily selected optimal feature subset.
accScores $\leftarrow$ [0] * $|\mathcal{FS}$_Lists$|$
overallScores $\leftarrow$ [0] * totalFeat
for $idx \leftarrow 0$ to $|\mathcal{FS}$_Lists$|$ do
    accScores[idx] $\leftarrow$ accScore(method, $|\mathcal{FS}$_Lists$|$, $\rho_m$_List[idx])
    if $\mathcal{S}_2$_Accuracies[idx] $<$ $\mathcal{S}_2$_All_Features_Acc then
        accScores[idx] $\leftarrow$ accScores[idx]/$\psi$
    end
    for $featIdx \leftarrow 0$ to totalFeat do
        featScore $\leftarrow$ featScore(feat, $\mathcal{FS}$_Lists[idx], featIdx + 1)
        if $\mathcal{S}_2$_Accuracies[idx] $>$ $\mathcal{S}_2$_All_Features_Acc and feat $\notin$ $\mathcal{FS}$_Lists[idx] then
            featScore $\leftarrow$ featScore * $\tau$
        end
        overallScore $\leftarrow$ featScore * accScore[idx]
        overallScores[featIdx] $\leftarrow$ overallScores[featIdx] + overallScore
    end
end
hybridFeatures $\leftarrow$ [ ]
for score $\in$ overallScores do
    if score $> \theta$ then
        hybridFeatures.append(feat)
    end
end
return hybridFeatures
```

range accounts for the possibility of selection from a base method with reduced dimensions and comparable but lower performance. Furthermore, the minimum number of features to be selected was set to 0.1× the total number of features to reject extremely low dimensional feature subspaces resulting in near-zero performance. The fitness function used to evaluate individuals is the average of $\mathcal{S}_3$ (with features of the hybrid subspace) accuracies obtained using RF, BDT, and KNN classifiers with 10-fold cross-validation. The stopping criteria for GA was achieved when either the optimal solution convergence or a limit on the maximum of iterations was reached. The flow of the proposed hybrid greedy ensemble feature selection optimized using GA is illustrated in Figure 3.

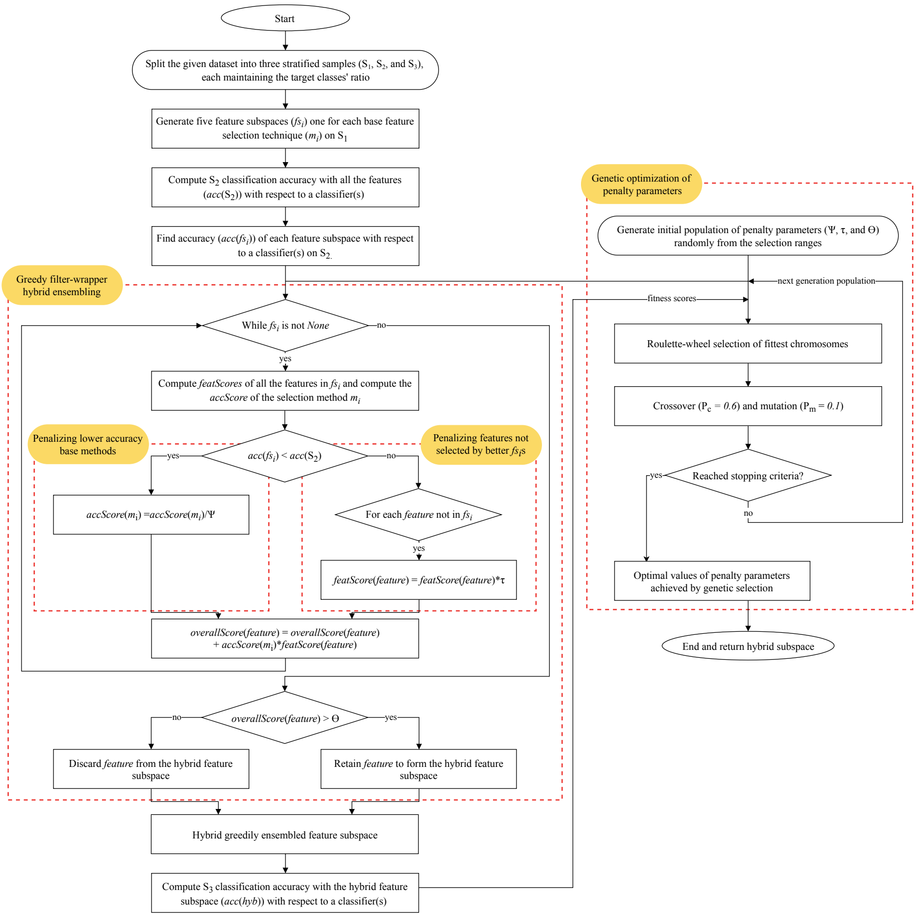
Start
Split the given dataset into three stratified samples (S1, S2, and S3), each maintaining the target classes' ratio
Generate five feature subspaces (fsi) one for each base feature selection technique (mi) on S1
Compute S2 classification accuracy with all the features (acc(S2)) with respect to a classifier(s)
Find accuracy (acc(fsi)) of each feature subspace with respect to a classifier(s) on S2.
Greedy filter-wrapper hybrid ensembling
While fsi is not None
no
yes
Compute featScores of all the features in fsi and compute the accScore of the selection method mi
Penalizing lower accuracy base methods
acc(fsi) < acc(S2)
yes
no
Penalizing features not selected by better fsis
accScore(mi) =accScore(mi)/Ψ
For each feature not in fsi
yes
featScore(feature) = featScore(feature)*τ
overallScore(feature) = overallScore(feature) + accScore(mi)*featScore(feature)
overallScore(feature) > Θ
no
yes
Discard feature from the hybrid feature subspace
Retain feature to form the hybrid feature subspace
Hybrid greedily ensembled feature subspace
Compute S3 classification accuracy with the hybrid feature subspace (acc(hyb)) with respect to a classifier(s)
Genetic optimization of penalty parameters
Generate initial population of penalty parameters (Ψ, τ, and Θ) randomly from the selection ranges
next generation population
fitness scores
Roulette-wheel selection of fittest chromosomes
Crossover (Pc = 0.6) and mutation (Pm = 0.1)
Reached stopping criteria?
yes
no
Optimal values of penalty parameters achieved by genetic selection
End and return hybrid subspace

Figure 3: The main process of the proposed hybrid greedy ensemble feature selection optimized using GA.

5. Experimental results and discussion

In this section, we report a detailed benchmarking of our filter-wrapper hybrid greedy ensemble approach on three high-dimensional biomedical datasets. We first describe the implementation setup, the working environment, and the validation procedure used. Then we discuss the parameter setup, their affect on the proposed system, and the performance of the proposed model, followed by its complexity analysis and training details. Finally, we elucidate on the implications of using our proposed hybrid ensemble in real-world biomedical applications.

5.1. Experimental setup and validation

To investigate the effectiveness of the proposed filter-wrapper hybrid greedy ensemble feature selection approach, we carried out a detailed benchmarking on

Table 6: Parameters used in the proposed hybrid greedy ensemble approach.

	Greedy parameter-wise optimization	**Genetic selection of optimal parameters**
ψ	$1 - 1.5$	$1 - 10$
τ	$1 - 10$	$1 - 25$
θ	$0 - 1$	Dataset-specific
Scaling factor	ψ: 0.1, τ: 2, and θ: 0.2	$-$
P_c and P_m	$-$	0.6 and 0.1

three high-dimensional biomedical datasets (see Table 3). Experiments related to hybrid feature selection were performed on a PC with Intel Core i5 4×1.8 GHz CPU with 8 GB RAM in the MAC 10.14 OS and the experiments related to parameter optimization were performed on a server with Intel Xeon 2×2.40 GHz processor with 8 GB RAM and 1×TESLA C-2050 (3 GB memory). All the experiments were coded in Python 2.7 and Weka 3.8.3. All the experiments were carried out by 10-fold cross-validation, and the overall performance was estimated as the average across all folds. The biomedical datasets have adequate samples to aid in the creation of three stratified samples. Furthermore, two balanced (TIS [51] and Skin Cancer [63]) and one imbalanced (Seizure [5]) high-dimensional datasets were chosen for an unbiased evaluation of the proposed technique.

Accuracy was used as the standard performance evaluation metric in this paper. Accuracy computes the average number of correct predictions over the given samples. Accuracy with $\mathcal{Y}_{\text{true}}$ ground truth labels, $\mathcal{Y}_{\text{pred}}$ predicted class labels, and $I(x, y)$ indicator function that returns 1 only when $x = y$, can be defined as in Equation 9.

$$\text{Accuracy}(\mathcal{Y}_{\text{true}}, \mathcal{Y}_{\text{pred}}) = \frac{1}{\#\text{samples}} \sum_{i=1}^{\#\text{samples}} I(\mathcal{Y}_{\text{true}_i}, \mathcal{Y}_{\text{pred}_i}) \tag{9}$$

Furthermore, to simplify the evaluation, the accuracy computed for three classifiers used in this paper (RF, BDT, and KNN) were aggregated by averaging the individual accuracy scores.

5.2. Parameter setup and performance benchmarking

The ranges of the penalty parameters must be preset to facilitate the ensembling of the base feature selection approaches in the most optimal way. The predetermined ranges were set with a higher τ range and comparably lower ψ range. A higher τ range was set to heavily penalize those less informative features that were not selected by the better performing base feature selection methods but were selected by methods with lower performance. Lower ψ range accounts for the possibility of selection from a base method with reduced dimensions and comparable but lower performance. While the ranges of τ and ψ

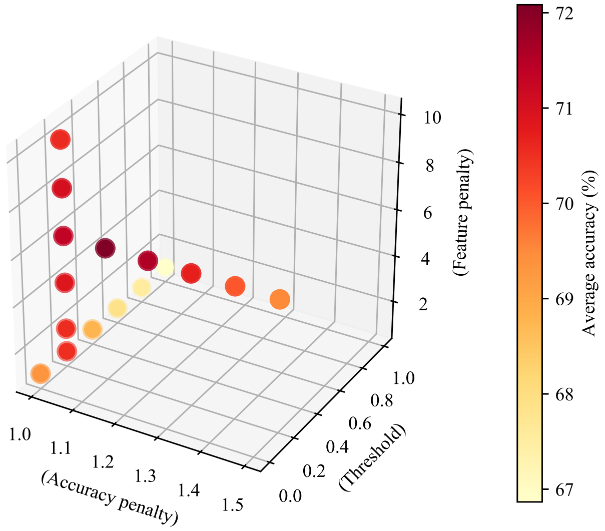

Figure 4: The effect of ψ, τ, and θ on the proposed hybrid feature selection technique using the Skin Cancer dataset [63].

can be set greedily by hillclimbing for an optimal range, θ is highly reliant on the overall scores of the features. The range of θ is set from the minimum of all *overallScore* values to the maximum of all *overallScore* values. In the case of greedy parameter-wise optimization, θ was set from 0.0 to 1.0 since this range was common to all the datasets used in this paper.

In the case of greedy parameter-wise optimization, an empirical analysis was conducted to evaluate the variations in the accuracy with the change in the penalty parameters (ψ, τ, and θ). Figure 4 shows the variations in the hybrid feature selection accuracy on the Skin Cancer dataset [63] with respect to penalty parameters ψ ranging from 1 to 1.5 (increments of 0.1), τ ranging from 1 to 10 (increments of 2) and θ ranging from 0.0 to 1.0 (increments of 0.2) as a heat map.

Table 7[5] and Table 8[5] present detailed insights into the empirical analysis performed on the Skin Cancer dataset [63] using greedy parameter-wise optimization. In Table 7 and Table 8 the parameters were greedily selected, starting with θ (varied from 0.0 to 1.0 (increments of 0.2)), followed by τ (varied from 1 to 10 (increments of 2)), and ψ (varied from 1 to 1.5 (increments of 0.1)). The threshold was varied initially, to find the best possible value which was then set throughout the analysis. A maximum average accuracy of 70.535% was obtained using $\psi = 1$, $\tau = 1$, and $\theta = 0.2$ (fix θ). Then, the feature penalty was varied to find the best possible value, which was then set. A maximum average accuracy of 71.354% was obtained using $\psi = 1$, $\tau = 6$, and $\theta = 0.2$. Finally, the accuracy penalty was varied to find the best possible value. Thus, by changing the values of the penalty parameters within the preset range, 72.085% average

[5] fs_{hyb} denotes the proposed hybrid feature selection, fs_{nil} denotes no feature selection, and fs_1 to fs_5 denote the base feature selection methods in the order of *cfsSubsetEval*, *mRMR*, *oneRFeatureEval*, *corrFeatureEval*, and *igFeatureEval*.

Table 7: The effect of ψ, τ, and θ on the proposed hybrid feature selection technique using the Skin Cancer dataset [63].

ψ	τ	θ	**Number of selected features**							**Classifier**	$\mathcal{S}_3$ **accuracy (%)**						
			fs_{hyb}	fs_{nil}	fs_1	fs_2	fs_3	fs_4	fs_5		fs_{hyb}	fs_{nil}	fs_1	fs_2	fs_3	fs_4	fs_5
1	1	0	275	2352	66	32	100	121	100	RF	71.5698	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	70.3116	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	66.1474	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	1	0.2	99	2352	66	32	100	121	100	RF	**73.2217**	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	**71.2630**	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	**67.1216**	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	1	0.4	64	2352	66	32	100	121	100	RF	71.2217	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	69.5630	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	65.5216	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	1	0.6	50	2352	66	32	100	121	100	RF	70.2217	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	69.2730	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	64.1216	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	1	0.8	34	2352	66	32	100	121	100	RF	69.9290	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	68.8472	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	63.8890	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	1	1	21	2352	66	32	100	121	100	RF	69.3829	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	68.5740	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	62.6423	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	2	0.2	98	2352	66	32	100	121	100	RF	73.2441	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	71.2700	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	67.1311	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	4	0.2	90	2352	66	32	100	121	100	RF	73.6821	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	71.4173	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	67.5230	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820

Table 8: The effect of ψ, τ, and θ on the proposed hybrid feature selection technique using the Skin Cancer dataset [63] (contd.).

ψ	τ	θ	**Number of selected features**							**Classifier**	$\mathcal{S}_3$ **accuracy (%)**						
			fs_{hyb}	fs_{nil}	fs_1	fs_2	fs_3	fs_4	fs_5		fs_{hyb}	fs_{nil}	fs_1	fs_2	fs_3	fs_4	fs_5
1	6	0.2	80	2352	66	32	100	121	100	RF	**74.1121**	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	**72.0360**	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	**67.9131**	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	8	0.2	72	2352	66	32	100	121	100	RF	74.0230	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	71.8850	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	67.2371	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1	10	0.2	70	2352	66	32	100	121	100	RF	73.5461	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	71.4010	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	66.7712	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1.1	6	0.2	73	2352	66	32	100	121	100	RF	**74.6113**	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	**72.3316**	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	**69.3110**	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1.2	6	0.2	70	2352	66	32	100	121	100	RF	74.0120	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	71.9162	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	68.7710	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1.3	6	0.2	68	2352	66	32	100	121	100	RF	73.1211	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	70.9913	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	68.0110	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1.4	6	0.2	64	2352	66	32	100	121	100	RF	72.5810	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	69.7113	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	67.8103	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820
1.5	6	0.2	62	2352	66	32	100	121	100	RF	72.0180	70.4014	70.1618	69.5626	70.9100	69.8322	69.0833
										BDT	69.4391	69.7723	68.9934	69.2630	69.2930	68.9934	68.0348
										KNN	67.0410	65.2187	62.4326	62.5225	65.3984	63.0917	62.8820

Table 9: Comparison of the average accuracies of the proposed hybrid feature selection technique optimized using GA (top ten chromosomes) over the base methods.

Dataset	**Chromosome**			**$\mathcal{S}_3$ average accuracy (%)**	
	ψ	τ	θ	fs_{hyb}	*Base selection methods*
TIS [51]	6.08	20.63	0.78	**87.449**	
	9.04	18.46	1.54	86.795	
	9.04	18.46	1.57	86.264	fs_{nil}: 81.787
	9.04	9.34	1.29	86.137	fs_1: 83.506
	1.05	24.06	0.58	85.592	fs_2: 73.391
	1.05	18.46	0.99	85.562	fs_3: 80.153
	8.12	24.06	0.58	85.434	fs_4: 83.395
	9.04	18.46	0.65	85.405	fs_5: **83.948**
	9.04	9.34	0.80	85.337	
	9.04	24.06	0.25	85.068	
Skin Cancer [63]	1.07	7.72	0.14	**78.912**	
	1.65	6.77	0.04	78.783	
	1.56	6.17	0.09	78.374	fs_{nil}: 68.464
	1.11	6.07	0.14	78.343	fs_1: **68.534**
	1.28	6.77	0.04	78.323	fs_2: 67.306
	1.55	6.46	0.12	78.292	fs_3: 66.667
	1.24	6.67	0.03	78.253	fs_4: 67.116
	2.28	5.97	0.17	68.910	fs_5: 67.196
	1.07	11.87	0.24	67.860	
	1.07	20.72	0.24	67.312	
Seizure [5]	1.39	2.58	0.01	**51.811**	
	1.31	3.09	0.43	50.822	
	1.62	5.96	0.19	49.422	fs_{nil}: **47.131**
	1.16	2.30	0.33	48.517	fs_1: 45.412
	1.92	7.58	-0.05	47.663	fs_2: 46.723
	1.81	1.21	0.11	47.063	fs_3: 45.988
	1.38	8.78	0.20	46.818	fs_4: 44.988
	1.14	9.53	0.31	46.421	fs_5: 45.858
	1.40	9.05	0.41	46.322	
	1.45	9.91	0.41	46.158	

accuracy (3.6% more than that for any predetermined feature selection method) was obtained with the penalty parameters $\psi = 1.1$, $\tau = 6$, and $\theta = 0.2$ selected using greedy parameter-wise optimization. Also, it was observed that our proposed method took an average running time[6] of 1.283 seconds while classification using all the features took 18.71 seconds.

[6] Average of the runtime obtained using RF, BDT, and KNN classifiers.

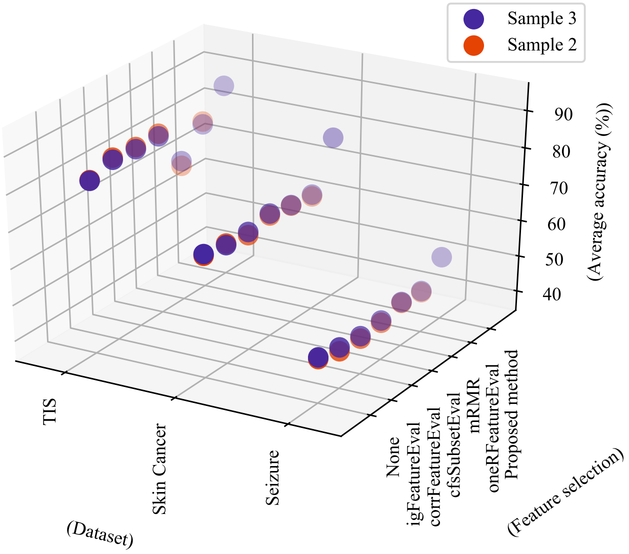

Figure 5: Comparison of the average accuracies of various feature selection methods across different datasets.

Figure 5 shows the superiority of the proposed hybrid greedy ensemble method optimized using GA, in terms of average accuracy over the base feature selection methods, via empirical analysis. Also, it is interesting to note from Figure 5 that the samples ($\mathcal{S}_2$, $\mathcal{S}_3$) obtained from stratified sampling were very similar. Table 9[5] compares the accuracies obtained using various penalty parameters optimized by GA with the selected base selection techniques. The penalty parameters (chromosomes in GA) were varied in the range of 1 to 10 for ψ, 1 to 25 for τ, and minimum *overallScore* to maximum *overallScore* for θ. The minimum number of features to be selected was set to 0.1× the total number of features, i.e., the penalty parameters resulting in a hybrid feature subspace with less than the set minimum threshold of features were rejected. Note that all the top ten chromosomes for the TIS dataset [51], top eight (of ten) chromosomes for the Skin Cancer dataset [63], and top five (of ten) chromosomes for the Seizure dataset [5] outperform the base selection methods. It can be noted that using GA to optimize the penalty parameters produces higher accuracies in comparison to the results obtained using greedy parameter-wise optimization. This can be attributed to the fact that the swarm-intelligence based heuristic search is flexible and versatile, in the sense that they mimic the best features in the nature. It can be observed that the value of ψ must be kept relatively low while the value of τ must be moderately adjusted to obtain an optimal feature subset. This can be attributed to the fact that ψ aims at penalizing those base selection methods with lower $\mathcal{S}_2$ accuracy than that obtained using the entire feature space, while τ penalizes the features not selected by the better performing feature selection methods ($\mathcal{S}_2$ accuracy greater than that obtained using entire feature space). Higher accuracy penalty indicates rejection of the features from those base selection techniques with lower accuracy than that obtained with no feature selection, implying that that technique was of no use at all in reducing the feature space. It was observed that the optimization of the proposed ensemble using GA took 1.86 hours for TIS dataset [51], 2.31 hours for the Skin Cancer dataset [63], and 1.44 hours for the Seizure dataset [5].

From Table 9 it can be noted that the proposed method, when optimized using GA outperforms the base methods by a maximum of 4.17% (3.5% higher) and at least by 1.34% (1.12% higher) for the TIS dataset [51], by a maximum of 15.14% (10.37% higher) and at least by 0.55% (0.38% higher) for the Skin Cancer dataset [63] and by a maximum of 9.93% (4.68% higher) and at least by 1.13% (0.53% higher) for the Seizure Dataset [5]. From the obtained results, the following two major trends were predominantly observed:

- The genetic selection performs an exhaustive heuristic search leading to better optimization of the values of the penalty parameters as compared to the values obtained using greedy parameter-wise optimization.
- A significantly lower value of the accuracy penalty (ψ) and a higher value of the feature penalty (τ) often leads to the optimal ensembling of base subspaces to produce the most informative feature subspace.

The obtained results indicate the superiority and efficiency and robustness of our proposed hybrid greedy ensemble optimized using GA over the base selection techniques. Furthermore, the proposed greedy ensemble approach was compared with state-of-the-art prolific wrapper methods [57] including Recursive Feature Elimination (RFE) using SVM with linear kernel and RFE using SVM with Radial Basis Function (RBF) kernel [40]. We also compare our results with widely used filter selection approaches including feature importance using RF [70] and chi-square test [47]. Table 10 presents the superiority of the proposed ensemble optimized using GA over prolific filter and wrapper methods. RFE using SVM with linear kernel, feature importance using RF, and chi-square selection approaches were performed on sample $\mathcal{S}_2$ to retain 100 best features, while RFE using SVM with RBF kernel was programmed to retain about top 10% of the features using sample $\mathcal{S}_2$. It can be observed that the proposed method outperforms prolific filter methods by 4% for the TIS dataset, 16.4% for the Skin Cancer dataset, and 11.2% for the Seizure dataset. Additionally, it can be remarked that the proposed method also outperforms state-of-the-art wrapper methods by 5% for the TIS dataset, 17.12% for the Skin Cancer dataset, and 2.7% for the Seizure dataset.

5.3. Computational complexity analysis

Concerning the training of the proposed ensemble approach, the feature subspaces from various predetermined base selection methods must be computed

Table 10: Comparison of the proposed hybrid greedy ensemble approach over state-of-the-art prolific filter and wrapper selection approaches.

Dataset	**$\mathcal{S}_3$ average accuracy (%)**				
	Filter approaches		**Wrapper approaches**		**Proposed greedy ensemble (GA)**
	Feature importance (RF)	Chi-square test	RFE using SVM with linear kernel	RFE using SVM with RBF kernel	
TIS [51]	84.182	84.220	79.968	83.290	**87.449**
Skin Cancer [63]	67.356	67.794	67.307	67.375	**78.912**
Seizure [5]	46.601	46.299	46.818	50.450	**51.811**

apriori. Additionally, the genetic selection of the optimal penalty parameters must also be achieved apriori. With the prescient knowledge of the optimal penalty parameters and feature subspaces, the proposed hybrid approach ensembles the subspaces greedily based on the penalty factors. Thus, the computational complexity of the proposed algorithm is heavily reliant on the complexity of obtaining predetermined feature subspaces ($= O(fs)$) and the genetic selection of the penalty parameters. It is interesting to note that the computational complexity of the proposed method is significantly reduced by using filter selection methods as the base selection approaches. Furthermore, since the selection of features is performed on a sample of the high-dimensional dataset as opposed to the entire dataset, the computational cost is reduced further.

To solve real-life optimization problems with less computational volume, an optimization or heuristic search strategy needs to be computationally feasible. To analyze the computational cost of the optimization strategy used in terms of worst-case computation time, the step-wise complexity analysis is performed. The initialization of a population of size P in GA is $O(n \cdot P) \approx O(P)$ complex, where n (constant, 3 here) is the size of each chromosome. The fitness evaluation used in this study is the average of the accuracies obtained from RF, BDT, and KNN classifiers. Assume that computing the average accuracy using these classifiers takes $O(\text{fitness})$ time which is radically dataset dependent. Roulette-wheel measures the area covered by a chromosome in a given population P using the fitness scores. Every chromosome forms a part of the wheel with its slice size proportionate to its fitness score. Roulette-wheel selection can be achieved in $O(P^2)$. Finally, crossover and mutation genetic operations take $O(\mathrm{P}_c \cdot O(\text{crossover}))$ and $O(\mathrm{P}_m \cdot O(\text{mutation}))$ times respectively. The GA optimization to find optimal penalty parameters is run for G iterations. Since P, G, P_c, and P_m are constant, the worst-case time complexity to optimize penalty parameters simplifies to $O(O(\text{fitness}) \cdot (O(\text{crossover}) + O(\text{mutation})))$ $(= O(\text{GA}))$. As a result, the worst-case time complexity of the proposed greedy ensemble using GA is $O(O(fs) + O(\text{fitness}) + O(\text{GA}))$. Note that $O(fs)$ is on the sample $\mathcal{S}_1$, $O(\text{fitness})$ is on the sample $\mathcal{S}_2$, and $O(\text{GA})$ is on the sample $\mathcal{S}_3$.

Table 11: Descriptive statistics of the proposed hybrid greedy ensemble approach across various datasets.

Dataset	**Minimum value**	**Maximum value**	**Mean**	**Standard deviation**
TIS [51]	85.068	87.449	85.904	0.751
Skin Cancer [63]	67.312	78.912	75.336	5.063
Seizure [5]	46.158	51.811	48.102	1.993

5.4. Effectiveness of the proposed greedy ensemble in real-world biomedical applications

The richness and variety of datasets available in the biomedical field have opened new horizons for researchers and investigators. The generated biomedical big data inherits the curse of dimensionality as one of its characteristics. Effective dimensionality reduction techniques emulate predictive capability while eliminating the noise and curtailing the computational complexity. Time and cost-effective approaches to select the most discriminative and informative features are indispensable, especially in the fields of bioinformatics and healthcare. The applicability of a feature selection technique to given data is heavily reliant on its ability to match the structure of the problem and maintain only those features that reveal the inherent patterns in the data. Thus there is a need to develop efficient techniques that aid in the optimal ensembling of such selection techniques for better performance aiding the decision-making process.

The proposed greedy approach can be used to ensemble a variety of effective selection approaches to generate an optimal feature subspace that captures the inherent nature of the data. The weighted occurrence scheme and the penalty scheme used in the proposed approach aid in the appropriate selection of most informative features. Such an appropriate selection of features needed for clustering, classification, pattern extraction, and prediction of high-dimensional biomedical datasets with hundreds of attributes can be facilitated effectively using the proposed greedy ensembling approach. Furthermore, the proposed approach can prominently aid in the achievement of efficient analysis in various biomedical applications including the analysis of large volumes of genomic data produced daily due to the advancements in the sequencing technology. This is particularly vital as the selection of important genes is essential to discover the knowledge hidden within the genetic code and to identify significant biomarkers. The robustness and flexibility of the proposed approach facilitate the effective feature selection needed for a wide variety of healthcare applications including disease prediction, risk management, and others. The flexibility or the ability to work with any predetermined set of selection methods allows the proposed greedy approach to work effectively to match the problem structure aiding in effective feature selection.

Table 12: A paired samples Wilcoxon signed-rank test (two-tailed, $p < 0.05$) for the proposed greedy ensemble against base selection methods across different datasets.

Dataset	**Selection method**	$p-$**value**	$z-$**value**	**Null hypothesis decision**	**Significant difference**
TIS [51]	None	0.00512	−2.8031	**Reject**	**Yes**
	igFeatureEval	0.00512	−2.8031	**Reject**	**Yes**
	corrFeatureEval	0.00512	−2.8031	**Reject**	**Yes**
	cfsSubsetEval	0.00512	−2.8031	**Reject**	**Yes**
	mRMR	0.00512	−2.8031	**Reject**	**Yes**
	oneRFeatureEval	0.00512	−2.8031	**Reject**	**Yes**
Skin Cancer [63]	None	0.02202	−2.2934	**Reject**	**Yes**
	igFeatureEval	0.02202	−2.2934	**Reject**	**Yes**
	corrFeatureEval	0.00512	−2.8031	**Reject**	**Yes**
	cfsSubsetEval	0.00512	−2.8031	**Reject**	**Yes**
	mRMR	0.00512	−2.8031	**Reject**	**Yes**
	oneRFeatureEval	0.00512	−2.8031	**Reject**	**Yes**
Seizure [5]	None	0.33204	−0.9683	Retain	No
	igFeatureEval	0.00512	−2.8031	**Reject**	**Yes**
	corrFeatureEval	0.09296	−1.6818	Retain	No
	cfsSubsetEval	0.00512	−2.8031	**Reject**	**Yes**
	mRMR	0.00512	−2.8031	**Reject**	**Yes**
	oneRFeatureEval	0.00512	−2.8031	**Reject**	**Yes**

6. Sensitivity analysis

The experimental results highlight the effectiveness and robustness of the proposed approach over the base selection methods. To further analyze the obtained results, a sensitivity analysis was performed. Sensitivity analysis helps in making decisions concerning more than a solution to the given problem [44]. Sensitivity analysis measures the extent to which the optimal solution is sensitive to the change in the input to one or more parameters. The Kolmogorov-Smirnov test of normality revealed that the obtained results were not normally distributed. Thus, a non-parametric paired samples Wilcoxon signed-rank test at 5% significance level was employed to evaluate the significance of the proposed hybrid ensemble over the base selection methods across various datasets. The top ten chromosomes were used in the determination of the significance of the proposed approach over base selection methods as it was assumed that the optimal values would converge to the values of the top ten chromosomes after many finite cycles. Table 11 summarizes the statistical analysis of the proposed approach for top ten chromosomes in terms of accuracy (mean) and robustness (standard deviation).

Table 12 shows the results of the paired samples Wilcoxon signed-rank test for the proposed greedy ensemble against conventional base selection methods. The null hypothesis claims no significant difference between the proposed greedy approach and a base selection approach. When the significance is greater than 5%, the null hypothesis is retained implying no significant improvement using the proposed hybrid approach. From Table 9 it can be observed that the proposed approach is significantly better than the base selection approaches except when the features of the Seizure dataset are all used or when they are selected using *corrFeatureEval*. All in all, the proposed hybrid greedy ensemble approach significantly outperforms the chosen base feature selection approaches.

7. Conclusions, limitations, and future directions

Feature selection in the field of biomedicine and bioinformatics is indispensable. In this study, we proposed a penalty based filter-wrapper hybrid greedy ensemble approach to facilitate optimal feature selection. The proposed approach greedily selects the features from the subspaces obtained from the predetermined base selection methods. Specific performance dependent penalty parameters were used to penalize the base feature subspaces essential to achieve the optimal ensembling of those subspaces. At any point in time, only a stratified sample and not the entire dataset is not used for computation; the computational complexity is significantly reduced. Furthermore, we leverage effective heuristic search strategies including the greedy parameter-wise optimization and the GA to obtain optimal values of the penalty parameters. Various applications in the field of bioinformatics and healthcare were detailed. Experimental validation using three high-dimensional biomedical datasets proves the superiority (in terms of prediction accuracy), efficiency, and robustness of the proposed ensemble approach. The proposed approach is scalable and flexible as it can accommodate (by ensembling) a variety of feature selection approaches. Empirically, we showed that the proposed greedy approach outperformed the chosen base feature selection methods by 4.17% for the TIS dataset, by 15.14% for the Skin Cancer dataset and by 9.93% for the Seizure dataset. The proposed approach also outperformed prolific filter and state-of-the-art wrapper methods by a 5% for the TIS dataset, by 17.12% for the Skin Cancer dataset and by 11.2% for the Seizure dataset.

Although the proposed approach effectively enhances the feature selection, it has some limitations which call for further research on this topic. First, the proposed method requires the existence of a significant number of records in the dataset for precise sampling. Second, the proposed hybrid greedy ensemble approach introduces additional (penalty) parameters. These vital penalty parameters require prior training to obtain the optimal setting in advance. Thus, parameter self-adaptive greedy ensemble or parameter-free greedy ensemble will be a prominent future research direction. Furthermore, we also aim at investigating the computational power of a hybrid of various metaheuristics including cuckoo search, firefly optimization, and GA which establishes an optimal balance between intensification and diversification.

Conflict of interest

The authors confirm that there are no known potential conflicts of interest associated with this publication.

Ethical approval

All the procedures performed by either of the authors in this study do not involve any human participants or animals.

References

[1] Nadia Abd-Alsabour. A review on evolutionary feature selection. In *Modelling Symposium (EMS), 2014 European*, pages 20–26. IEEE, 2014.

[2] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2009.

[3] S Agarwal and P Ranjan. Dimensionality reduction methods classical and recent trends: a survey. *IJCTA*, 9(10):4801–4808, 2016.

[4] Ali Al-Shahib, Rainer Breitling, and David Gilbert. Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics*, 4(3):195–203, 2005.

[5] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6): 061907, 2001.

[6] B Azhagusundari and Antony Selvadoss Thanamani. Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(2):18–21, 2013.

[7] Michael Taynnan Barros, Reinaldo Cezar Gomes, Marcelo Sampaio de Alencar, and Anderson Fabiano Costa. Feature filtering techniques applied ip traffic classification. In *IADIS International Conference WWW/Internet*, pages 227–234, 01 2013.

[8] Richard E Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.

[9] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.

[10] Verónica Bolón-Canedo, Sohan Seth, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, and José C Príncipe. Statistical dependence measure for feature selection in microarray datasets. In *ESANN*, 2011.

[11] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[12] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[13] Siow-Wee Chang, Sameem Abdul-Kareem, Amir Feisal Merican, and Rosnah Binti Zain. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*, 14(1):170, 2013.

[14] Yongqiang Dai, Bin Hu, Yun Su, Chengsheng Mao, Jing Chen, Xiaowei Zhang, Philip Moore, Lixin Xu, and Hanshu Cai. Feature selection of high-dimensional biomedical data using improved sfla for disease diagnosis. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 458–463. IEEE, 2015.

[15] Bing Dong, Burton Andrews, Khee Poh Lam, Michael Höynck, Rui Zhang, Yun-Shang Chiou, and Diego Benitez. An information technology enabled sustainability test-bed (itest) for occupancy detection through an environmental sensing network. *Energy and Buildings*, 42(7):1038–1046, 2010.

[16] Hongbin Dong, Tao Li, Rui Ding, and Sun, Jing. A novel hybrid genetic algorithm with granular information for feature selection and optimization. *Applied Soft Computing*, 65:33–46, 2018.

[17] Tobore Ekwevugbe, Neil Brown, and Vijay Pakka. Realt-time building occupancy sensing for supporting demand driven hvac operations. In *International Conference for Enhanced Building Operations*. Energy Systems Laboratory (http://esl. tamu. edu), 2013. doi: 10.13140/RG.2.1.2983.4326.

[18] Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H Witten. Data mining in bioinformatics using weka. *Bioinformatics*, 20(15):2479–2481, 2004.

[19] Mohammed A Gaafar, Noha A Yousri, and Mohamed A Ismail. A novel ensemble selection method for cancer diagnosis using microarray datasets. In *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, pages 368–373. IEEE, 2012.

[20] Temujin Gautama, Danilo P Mandic, and Marc M Van Hulle. Indications of nonlinear structures in brain electrical activity. *Physical Review E*, 67 (4):046204, 2003.

[21] David E Goldberg and Kalyanmoy Deb. A comparative analysis of selection schemes used in genetic algorithms. In *Foundations of genetic algorithms*, volume 1, pages 69–93. Elsevier, 1991.

[22] Walter J Gutjahr. A generalized convergence result for the graph-based ant system metaheuristic. *Probability in the Engineering and Informational Sciences*, 17(4):545–569, 2003.

[23] Mohsen Hajiloo, Hamid R Rabiee, and Mahdi Anooshahpour. Fuzzy support vector machine: an efficient rule-based classification technique for microarrays. *BMC bioinformatics*, 14(13):S4, 2013.

[24] Mark A Hall. Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*, 1998.

[25] Matthew He and Sergey Petoukhov. *Mathematics of Bioinformatics: Theory, Methods and Applications*, volume 19. John Wiley & Sons, 2011.

[26] Tin Kam Ho. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE, 1995.

[27] Geoffrey Holmes, Andrew Donkin, and Ian H Witten. Weka: A machine learning workbench. In *Proceedings of ANZIIS'94 - Australian New Zealnd Intelligent Information Systems Conference*, pages 357–361, Nov 1994. doi: 10.1109/ANZIIS.1994.396988.

[28] Anil K Jain, Robert PW Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.

[29] Alan Jović, Karla Brkić, and Nikola Bogunović. A review of feature selection methods with applications. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*, pages 1200–1205. IEEE, 2015.

[30] Jan Kalina and Anna Schlenker. Dimensionality reduction methods for biomedical data. *Lékař a technika-Clinician and Technology*, 48(1):29–35, 2018.

[31] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee, and Byoung-Tak Zhang. mitarget: microrna target gene prediction using a support vector machine. *BMC bioinformatics*, 7(1):411, 2006.

[32] Jinyan Li and Huiqing Liu. Kent ridge bio-medical data set repository. *Institute for Infocomm Research. http://sdmc. lit. org. sg/GEDatasets/Datasets. html*, 2002.

[33] Jinyan Li, Huiqing Liu, and Limsoon Wong. Mean-entropy discretized features are effective for classifying high-dimensional bio-medical data. In *Proceedings of the 3rd International Conference on Data Mining in Bioinformatics*, pages 17–24. Springer-Verlag, 2003.

[34] Tao Li, Chengliang Zhang, and Mitsunori Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.

[35] Xiangtao Li and Minghao Yin. Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Transactions on NanoBioscience*, 12(4):343–353, 2013.

[36] Yong Liang, Cheng Liu, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, and Hai Zhang. Sparse logistic regression with a l 1/2 penalty for gene selection in cancer classification. *BMC bioinformatics*, 14(1):198, 2013.

[37] Bo Liao, Yan Jiang, Wei Liang, Wen Zhu, Lijun Cai, and Zhi Cao. Gene selection using locality sensitive laplacian score. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(6):1146–1156, 2014.

[38] Quanjin Liu, Zhimin Zhao, Ying-Xin Li, and Yuanyuan Li. Feature selection based on sensitivity analysis of fuzzy isodata. *Neurocomputing*, 85: 29–37, 2012.

[39] Quanjin Liu, Zhimin Zhao, Ying-xin Li, Xiaolei Yu, and Yong Wang. A novel method of feature selection based on svm. *JCP*, 8(8):2144–2149, 2013.

[40] Quanzhong Liu, Chihau Chen, Yang Zhang, and Zhengguo Hu. Feature selection for support vector machines with rbf kernel. *Artificial Intelligence Review*, 36(2):99–115, 2011.

[41] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

[42] Mustafa K Masood, Yeng Chai Soh, and Jiang, Chaoyang. Occupancy estimation from environmental parameters using wrapper and hybrid feature selection. *Applied Soft Computing*, 60:482–494, 2017.

[43] Ujjwal Maulik and Debasis Chakraborty. Fuzzy preference based feature selection and semisupervised svm for cancer classification. *IEEE transactions on nanobioscience*, 13(2):152–160, 2014.

[44] Asma Meddeb, Nesrine Amor, Mohamed Abbes, and Souad Chebbi. A novel approach based on crow search algorithm for solving reactive power dispatch problem. *Energies*, 11(12):3321, 2018.

[45] Fan Min, Qinghua Hu, and William Zhu. Feature selection with test cost constraint. *International Journal of Approximate Reasoning*, 55(1):167–179, 2014.

[46] Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[47] Michal Moran and Goren Gordon. Curious feature selection. *Information Sciences*, 485:42–54, 2019.

[48] Craig G Nevill-Manning, Geoffrey Holmes, and Ian H Witten. The development of holte's 1r classifier. In *Artificial Neural Networks and Expert Systems, 1995. Proceedings., Second New Zealand International Two-Stream Conference on*, pages 239–242. IEEE, 1995.

[49] Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

[50] Mykola Pechenizkiy, Alexey Tsymbal, and Seppo Puuronen. Local dimensionality reduction and supervised learning within natural clusters for biomedical data analysis. *IEEE Transactions on Information Technology in Biomedicine*, 10(3):533–539, 2006.

[51] Anders Gorm Pedersen and Henrik Nielsen. Neural network prediction of translation initiation sites in eukaryotes: perspectives for est and genome analysis. In *Ismb*, volume 5, pages 226–233. Citeseer, 1997.

[52] Hanchuan Peng, Fuhui Long, and Chris Ding. 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, 2005.

[53] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[54] Thu Zar Phyu and Nyein Nyein Oo. Performance comparison of feature selection methods. In *MATEC Web of Conferences*, volume 42, page 06002. EDP Sciences, 2016.

[55] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[56] Steven L Salzberg, Arthur L Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated markov models. *Nucleic acids research*, 26(2):544–548, 1998.

[57] Hector Sanz, Clarissa Valim, Esteban Vegas, Josep M Oller, and Ferran Reverter. Svm-rfe: selection and visualization of the most relevant features through non-linear kernels. *BMC bioinformatics*, 19(1):432, 2018.

[58] Alok Sharma, Seiya Imoto, Satoru Miyano, and Vandana Sharma. Null space based feature selection method for gene expression data. *International Journal of Machine Learning and Cybernetics*, 3(4):269–276, 2012.

[59] Qinbao Song, Jingjie Ni, and Guangtao Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE transactions on knowledge and data engineering*, 25(1):1–14, 2013.

[60] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[61] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

[62] Divya Tomar and Sonali Agarwal. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013.

[63] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *arXiv preprint arXiv:1803.10417*, 2018.

[64] Qiang Tu, Xuechen Chen, and Liu, Xingcheng. Multi-strategy ensemble grey wolf optimizer and its application to feature selection. *Applied Soft Computing*, 76:16–30, 2019.

[65] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[66] Jihoon Yang and Vasant Honavar. Feature subset selection using a genetic algorithm. In *Feature extraction, construction and selection*, pages 117–136. Springer, 1998.

[67] Lei Yu, Yue Han, and Michael E Berens. Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):262–272, 2012.

[68] Zhiwen Yu, Hongsheng Chen, Jane You, Hau-San Wong, Jiming Liu, Le Li, and Guoqiang Han. Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11 (4):727–740, 2014.

[69] Rui Zhang, Khee Poh Lam, Yun-Shang Chiou, and Bing Dong. Information-theoretic environment features selection for occupancy detection in open office spaces. In *Building Simulation*, volume 5, pages 179–188. Springer, 2012.

[70] Xi Zhu, Xiaofei Du, Mike Kerich, Falk W Lohoff, and Reza Momenan. Random forest based classification of alcohol dependence patients and healthy controls using resting state mri. *Neuroscience letters*, 676:27–33, 2018.