Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 3984082, 20 pages
https://doi.org/10.1155/2022/3984082

Research Article

A Hybrid Feature Reduction Approach for Medical Decision Support System

Bikram Kar and Bikash Kanti Sarkar

Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Correspondence should be addressed to Bikram Kar; kar.bikram2012@gmail.com

Received 30 May 2022; Revised 8 August 2022; Accepted 9 August 2022; Published 30 September 2022

Academic Editor: Rajesh Kaluri

Feature reduction is essential at the preprocessing stage of designing any reliable and fast disease diagnosis model. Addressing the limitations like disease specificity, information loss, and operating NP problem in polynomial time, this paper introduces a two-step hybrid feature selection approach to identify a subset of most relevant and contributing features of each medical dataset for constructing diagnostic model. The concept of information gain is used in Step I to select the informative features, whereas a correlation coefficient-based approach is employed in Step II to retain the informative features possessing much dependency with class attribute but less dependency among the non-class attributes. In particular, both the approaches are sequentially fused to select approximately optimal features in order to construct better classification model in terms of performance and time. The optimal threshold criteria are decided to choose the most appropriate features from the datasets. The effectiveness of the proposed approach is assessed using six individual competent learners and one ensemble learner over seventeen disease datasets of smaller to larger dimensions. The empirical results indicate that the proposed approach improves the performance over the datasets after feature selection, reducing considerable amount of irrelevant and redundant data.

1. Introduction

Today, the healthcare sector is considered one of the essential industries in information technology (IT). But due to the application of IT to the healthcare industry, a huge amount of health data is constantly being generated. As a result, this industry demands overwork of human professionals like doctors, nurses, and other health workers to make more effective and efficient health services (e.g., diagnosis, nursing care, counselling, therapy, and nutrition). However, almost all these services are primarily associated with the diagnosis of diseases which should be accurate and prompt (on demand). It may be noted that the use of data mining (DM) and machine learning (ML) (including statistical analysis) has been an indispensable part of IT in the healthcare industry to improve the quality of health services. Undoubtedly, it is essential to design disease diagnosis support systems (DDSSs) by applying DM and ML approaches so that healthcare professionals benefit from accurate and fast diagnosis, reducing their time and efforts. Further, DDSSs can fill the gaps of the existing techniques adopted in the health units, and such models avoid information loss and reduce diagnosis costs.

However, clinical datasets are very complex in nature, for example, disease datasets generally possess huge amount of data with high dimensionalities (input variables/features/attribute); data in the dataset are usually collected from different sources in a different format (i.e., heterogeneous); there may exist lots of missing data, outlier, inconsistent data, and so on in the dataset, and characteristics of data are dynamically changing. Among these complexities, the dimensionality curse makes a major issue in designing good DDSS. Especially, clinical data with high dimensions may have many redundant/unnecessary (i.e., highly correlated non-target features) and irrelevant attributes (less relevant features with class feature), and they do not contribute to designing DDSS. Instead, they often degrade the performance of the designed DDSS [1]. For example, machine learning algorithms like C4.5 [2], K-nearest neighbour (KNN) [3], and Naïve Bayes [4] show often adverse effect on their performances due to the presence of such redundant attributes. Also, their presence in the database makes time concern during construction of DDSS and decision making. So, feature reduction is the only solution to overcome these concerns, and any ideal reduction approach assists in developing stable DDSS, even though the characteristics of medical data dynamically change.

Dimensionality reduction is an essential but challenging task in data mining. It helps in *data compression* and hence reduces storage space. It also reduces *computation time*. The reduction techniques are primarily divided into two categories: *feature extraction* (FE) and *feature selection* (FS). FE methods (usually applicable in image processing and natural language processing) aim to reduce the number of features in a dataset by creating new features (combing the existing ones) and then discarding the original (actual) features. On the contrary, FS methods reduce the dataset size by choosing only the relevant and non-redundant features but retaining adequate information for the learning task. Several FS approaches are introduced so far specifically to tackle medical datasets, and research is still going on for further improvement. The systematic review by Kawamoto et al. showed research interests prior to 2005 to improve clinical practice using clinical decision support systems through FS approaches [5]. A list of feature selection-based research works carried out over the last 20 years on clinical datasets is cited here to show the substantial research interest in FS in the medical domain [6–19].

Importantly, FS techniques are being extensively applied to reduce data dimensions in big data analytics [20, 21]. In 2021, Majid and Maryam proposed a distributed ensemble imbalanced FS framework to deal with big imbalanced datasets [22]. López et al. [23] proposed a distributed feature weighting algorithm based on RELIEF technique applied for small problem to estimate features importance of large-scale data. Reddy et al. [24] investigated two well-known dimension reduction approaches, namely, linear discriminant analysis (LDA) and principal component analysis (PCA), in the perspective of big datasets (including cardiotocography (CTG) dataset and diabetic retinopathy (DR) as medical datasets) and concluded that if the dimensionality of datasets is low, ML algorithms without dimensionality reduction yield better results. In 2022, Chen et al. [25] proposed a multi-tasking particle swarm optimization (PSO) approach for high-dimensional datasets (including many clinical datasets) to achieve higher classification accuracy in a shorter time than other state-of-the-art FS methods on high-dimensional classification. Interestingly, Hu et al. [26] introduced a multi-participant federated evolutionary FS algorithm for imbalanced data under privacy protection.

Very recently, the graph-based methods, including graph theory [27–29], spectral embedding [30], spectral clustering [31], and semi-supervised learning [32], are being significantly used in many problems for FS because of their capability of encoding similarity relationships among the features. Interestingly, these techniques may be applied in the medical field, since most of the medical datasets consist of images. Alelyani proposed one bagging-based ensemble approach to improve stability of feature selection in clinical datasets using data variance reduction [33]. In 2021, Xie et al. [34] developed a standard deviation and cosine similarity-based FS approach to tackle the challenges in genomic data analysis caused by their tens of thousands of dimensions while having a small number of examples and unbalanced examples between classes. In 2020, Sarkar proposed a two-step knowledge extraction framework for faster and accurate detection of disease [35]. The model used the entropy reduction approach to select few best relevant features from each dataset, but the issue is that several features in selected set may be correlated (i.e., redundant) among themselves which may often degrade the performance of the developed model. A few more standard published studies are listed in Tables 1 and 2, comparing their performances with the present work.

1.1. Research Scope.

As of now, there is extensive literature on feature selection in the medical domain. But most of them are disease specific or research focuses very less on generalizability case. Further, deciding *threshold value* criteria sufficient to identify minimal feature set is another issue in feature reduction. Also, dimension reduction often leads to information loss. It may be noted that, for dimension reduction, researchers prefer principal component analysis (PCA), but retaining the number of components is a big issue in PCA. Further, feature selection is viewed as a search optimization problem. More specifically, minimum feature subset selection (MFSS) is proved to be an NP problem [36, 37]. The MFSS NP problem is mathematically explained below.

1.1.1. Minimum Feature Subset Selection (MFSS) as NP Problem.

The search space in context of MFSS includes all possible feature subsets to discover the best feature subset, and the total number of possible ways to select feature subsets will be

$$N = \sum_{s=0}^{n} \binom{n}{s} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n, \tag{1}$$

where n is the dimensionality (quantity of original features) and s denotes the size of the chosen current feature subset.

Certainly, selection of zero (0) features (*i.e.*, ${}^{n}C_0$) may be ignored.

1.1.2. Challenges.

The problem to *discover the ideal feature subset* seems to be NP-hard because the analysis of all the feature subsets is costly in a computational manner, time-consuming, and inefficient even in case of small sizes. In fact, the exhaustive search can find the optimal solution, if the number of variables is not too large. In particular, there is still no effective way to deal with this problem. That is why, the problem is attempted to solve sustainably by using statistical or information theory-based or search-based strategies including best-first, branch-and-bound, simulated annealing, genetic algorithms, and so on.

Table 1: Feature reduction (% age) comparison of the proposed work with some standard existing works over clinical datasets.

Dataset	Total no. of features	Number of features and (% age) reduced by approaches adopted in the Ref.									
		[19]	[53]	[54]	[55]	[56]	[57]	[58]	[14]	[59]	Proposed work
Lung Cancer	56	—	—	30 (54%)	—	—	52 (92%)	—	—	—	36 (65%)
Wisconsin Breast Cancer	10	03 (30%)	05 (50%)	—	—	—	—	6 (60%)	—	—	6 (60%)
Hepatitis	19	09 (47%)	—	—	**(0%)	—	18 (94%)	—	—	—	12 (63%)
Colon Cancer	2000	—	1995 (99%)	—	—	—	—	—	1981 (98.9%)	—	1685 (64%)
Indian Liver Patient	10	—	—	—	**(0%)	—	9 (90%)	—	—	—	5 (50%)
Cleaveland Heart Disease	13	10 (77%)	—	—	**(0%)	1 (8%)	—	—	—	3 (23%)	6 (46%)
Parkinson’s	22	—	—	—	—	—	7 (31%)	—	—	—	17 (77%)
Dermatology	34	—	—	—	—	—	—	26 (76%)	—	—	23 (67%)

Note. **“∗∗” implies all features are selected (i.e., no features reduced). Further, “—” means result not available due to experiment not conducted over the dataset. Percentage shown within parenthesis in columns is the % age amount reduced by the works.

Table 2: Accuracy (% age) comparison of the proposed work with some surveyed works.

Cited paper	Methods	Performance measuring metrics (best results) shown for dataset								Limitations
		Lung Cancer	Wisconsin Breast Cancer	Hepatitis	Colon	Indian Liver Patient	Cleveland Heart Disease Dataset	Parkinson's	Dermatology	
[19]	***Feature selection***: filter-based neural network. ***Used learner***: RBF.	—	CA = 97.28%, specificity = 0.99, sensitivity = 0.94, AUROC = 0.98	CA = 85.16%, specificity = 0.66, sensitivity = 0.90, AUCROC = 0.86	—	—	CA = 84.46%, specificity = 0.66, sensitivity = 0.90, AUROC = 0.81	—	—	Disease specific and data reduction is less as compared to the *proposed method.*
[53]	***Feature selection***: ML-based ensemble feature selection. ***Used learners***: DT, RF, KNN, SVM.	—	CA = 97.00%	—	CA = 84.00%	—	—	—	—	Disease specific and performance wise not good as compared to the *proposed method.*
[56]	***Feature selection***: the features are chosen using different combinations of feature based on hit-and-trial approach. ***Used learners***: KNN, decision tree, NB, logistic regression, SVM, NN, voting (ensemble of Naïve Bayes and logistic regression).	—	—	—	—	—	Accuracy = 86.87%, F-measure = 88.22% precision = 95.00%	—	—	Not generic. Finding the proper combinations of feature from a large number of features is a *time-consuming task.*
[57]	***Feature selection***: chaotic crow search algorithm. ***Used learner***: KNN, where $K = 3$.	CA = 100%	—	CA = 100%	—	CA = 71.68%	—	CA = 90.78%	—	Performance is measured on the basis of only one learner.
[58]	Feature selection: binary grasshopper optimization algorithm. Used learner: KNN.	—	CA = 97.43%	—	—	—	—	—	CA = 100%	Amount of data reduction is less compared to the proposed method.

Table 2: Continued.

Cited paper	Methods	Performance measuring metrics (best results) shown for dataset								Limitations
		Lung Cancer	Wisconsin Breast Cancer	Hepatitis	Colon	Indian Liver Patient	Cleveland Heart Disease Dataset	Parkinson's	Dermatology	
[14]	Feature selection: filter-based multi-objective PSO algorithm and node centrality technique. Used learners: SVM, NB, and AdaBoost.	—	—	—	CA = 81.87%	—	—	—	—	Heuristic approach is adopted. It may increase time complexity. CA is not up to mark.
[59]	Feature selection: wrapper-based genetic algorithm. Used learner: SVM.	—	—	—	—	—	CA = 88.34%	—	—	Not generic. The developed method is applied on only one dataset.
[54]	Feature selection: binary coded Jaya optimization algorithm. Used learner: KNN.	CA = 75.00%	—	—	—	—	—	—	—	Performance is not good as compared to the other approaches in the literature.
[55]	Feature selection: CSO, KH, BFO-based super Learner. Used learner: SVM.	—	—	CA = 90%, sensitivity = 91%, specificity = 88%, precision = 91% F-score = 0.92	—	CA = 70.00%, sensitivity = 84.21%, specificity = 63.41%, precision = 51.61%, F-score = 0.64	CA = 84%, sensitivity = 80%, specificity = 86%, precision = 80%, F-score = 0.80	—	—	Not generic. Developed method is applied on limited number of datasets.
Proposed work in this study	Feature selection: information gain and correlation-based hybrid feature selection approach. Used learners: J48, JRip, KNN, ANN, NB, SVM, J48 + JRip	CA = 84.37%, (TPR, FPR) = (0.844, , 0.332), AUROC = 0.862	CA = 97.89%,(TPR, FPR) = (0.979, 0.017), AUROC = 0.996	CA = 83.92%, (TPR, FPR) = (0.839, 0.261), AUROC = 0.795	CA = 89.00%, (TPR, FPR) = (0.928, 0.200), AUROC = 0.516	CA = 72.14%, (TPR, FPR) = (0.721,0.518), AUCROC = 0.810	CA = 57.09%, (TPR, FPR) = (0.571, 0.206), AUROC = 0.809	CA = 97.82% (TPR, FPR) = (0.978, 0.037), AUROC = 0.97	CA = 97.29% (TPR, FPR) = (0.973, 0.005), AUROC = 0.994	Less number of big datasets is experimented.

Note. "—"means result not available due to experiment not conducted over the dataset.

1.1.3. Present Work. In the present study, a two-step hybrid generic model is proposed to identify an ideal subset of features for medical datasets, taking the strength of the existing statistical measures—*information gain* and *correlation coefficient*. The hybrid approach is a polynomial time approximation approach to tackle MFSS problem. More specifically, the concept of information gain is used in the first step to identify the most *relevant* non-target attributes with more information gain, whereas a correlation coefficient-based approach is used in the subsequent step to search the non-target features having maximum dependency with class attribute but minimum dependency among the non-target attributes. Optimal threshold criteria are decided based on the *trial-and-error* approach in Step II to select the most *informative* features from the datasets. However, threshold value is deterministically set in Step I. Generally, threshold limits are determined by expert knowledge, but such decision may not often result in good solution. In addition, it may vary from problem to problem. That is why, threshold values in Step II are decided based on the trial-and-error approach. The approach also includes provision of decremental and incremental scope of threshold values dynamically. Hence, we may claim that this approach has capability of high compression of storage and much time reduction, resulting in minimum information loss (since improvement is observed in performance metrics). Anyway, both the steps follow *backward elimination* technique to retain the best features.

1.2. Contributions of the New Hybrid Approach

(i) Operating MFSS (an NP problem) in polynomial time by the *method of hybridization* and setting optimal *threshold values* through *trial-and-error approach* to get the maximum accuracy and minimum false rate. The time complexity of the approach is $O(n^3)$ with n attributes/features in the dataset.

(ii) It is a *generic* feature reduction model for medical datasets, i.e., it targets medical datasets *irrespective* of any medical disease dataset. However, the model may show better performance for datasets other than medical datasets but nothing wrong is there.

(a) The speciality about operating disease datasets is claimed by *adaptation* of information gain-based approach and correlation coefficient-based approach in sequence. Medical datasets usually possess more *irrelevant* and *redundant* features. The information gain approach in the first step aims to eliminate the *irrelevant features,* whereas the correlation coefficient-based approach in the second step aims to eliminate both *irrelevant* and *redundant* features (not losing attributes with maximum information gain).

(b) The speciality about disease generalability in the approach is the *provision* of changing threshold values decided for non-target and target and non-target and non-target pairs in the datasets.

(iii) Preventing information loss due to feature reduction, since the model aims not to lose the informative features while processing features in Steps I and II as well. Prevention of information loss is validated through performance measuring metrics like accuracy, true positive rate (TPR), false positive rate (FPR), and area under the receiver operating characteristics (AUROC).

(iv) Datasets of different dimensions and rarely considered datasets like Arrhythmia, Lower back pain, Malaria, and Parkinson's are experimented in this study.

(v) The percentage of feature reduction by the new model is high enough.

1.3. Organization of the Article. The rest of the paper is organized as follows. Section 2 includes previous works related to the present work. Section 3 discusses *the proposed methodology* in detail. The implementation of the method, the obtained results, and analysis of the results are illustrated in Section 4. The conclusions and the future scope are presented in Section 5.

2. Previous Works

Prior to the model description, previous works related to the proposed model are included in this section. It may be noted that basic knowledge on dimensionality reduction and its very common categories, namely, feature selection and feature extraction, are already included in the Introduction section. Truly, before feature reduction using machine learning approaches, few features may be simply ignored as follows:

(i) Domain expert may reduce unnecessary features.

(ii) Feature exceeding certain threshold value of missing data may be removed.

Next, suspecting interdependence among features and less contributary features, the machine learning-based feature reduction approaches need to be applied. Now based on labelled, unlabelled, and partially labelled data, the standard feature selection methods are usually divided into *three* main categories: *supervised*, *unsupervised*, and *semi-supervised* [38]. Any supervised method selects and evaluates convenient features based on labelled data. Entropy-based technique is a supervised FS technique. On the other hand, unsupervised FS techniques ignore the target variable and remove redundant variables. The correlation coefficient-based approach is usually considered as an example of unsupervised FS method. Evaluating and selecting features in the unsupervised method are made based on the ability to meet some of the dataset's properties, like locality preserving ability and variance. However, a small amount of labelled data (not all) is available in many datasets, and finding their labels is costly. So, semi-supervised or constrained methods are used in such cases. In particular, the semi-supervised FS method uses both labelled and unlabelled data.

Further, based on the evaluation methods adopted for feature selection, the methods may be categorized as *filter*, *wrapper*, *embedded, ensemble*, and *hybrid* approaches [39–41]. In the filter-based method, four types of evaluation criteria, namely, dependency, information, distance, and consistency (i.e., unambiguous), are used. These methods are classifier independent. So, such technique has better generalization property but ignores interactions between classifiers. For more details about filter-based approaches, one may refer to the studies [42, 43]. On the contrary, a learning algorithm in the wrapper-based method is iteratively employed to evaluate the quality of feature subsets in the search space. This method interacts with classifier frequently and focuses on minimizing the prediction error. So, the major issue of the wrapper method is the computational complexity. Some common examples of wrapper methods are *forward feature selection*, *backward feature elimination*, *recursive feature elimination*, etc. In this regard, one may refer to the recent studies [44, 45]. The embedded method is a built-in FS mechanism that embeds the FS in the learning algorithm and uses its properties to good feature evaluation. Therefore, ensemble approaches are often the best way to tackle the limitations of the individual approaches. In general, the ensemble model aims to construct a group of feature subsets and then produce an aggregated result out of the group. It interacts with classifier. So, it is classifier specific but better than wrapper method, since it interacts with classifier once (not frequently). LASSO and RIDGE regressions are some popular examples of this method. These have inbuilt penalization functions to reduce overfitting. The time complexity of this model is also high. Finally, approaches based on hybridization employ the wrapper model's proper performance and the filter model's computational efficiency. However, the accuracy issue may be challenging in the hybrid model, since the filter and wrapper models are considered two separate steps [46]. So, we need to develop new ideas in order to design a new model (hybrid model) that will be able to improve performance of the learners, taking a smaller number of computational resources. The hybrid method can be formed by combining two or more different methods (usually filter method). It attempts to inherit the strengths of the individual methods.

3. Proposed Hybrid Feature Selection Approach

The conceptual view of the proposed *selection-based* feature reduction approach is depicted in Figure 1. The hybrid approach is, indeed, Phase II of the entire work carried out in the present study. More specifically, Phase I of the model deals with original datasets (drawn directly from several sources) and the performance measures of the selected competent learners over the chosen datasets. Phase II attempts to search for accurate features from each original dataset. Finally, Phase III uses the same learners and the same infrastructure (as applied in Phase I) to measure their performances over the datasets with reduced features.

3.1. Definition

(i) *Original Dataset.* Medical datasets with original features (drawn directly from the data repository) are termed here as the original datasets. In such a dataset, features/attributes are recommended by experts (physicians).

(ii) *Relevant Attribute.* As per the existing literature, a non-target feature (x) is relevant with target attribute (C) if these two are highly dependent (or correlated), otherwise irrelevant

(iii) *Redundant Attribute.* As per the existing literature, a non-target feature (x) is relevant with another non-target attribute (y) if these two are highly dependent (or correlated). Here, same information is carried by both x and y about the dataset

Details of Phase II. The very common but unimportant nominal attributes (e.g., id number, zip code, eye colour, and so on) in medical dataset are first discarded from it. Next, the suggested FS method is applied in Phase II. In fact, this phase consists of *two* steps: Step I and Step II. An entropy-based approach is employed in Step I to choose the most informative features from each original dataset. In Step II, the feature set decided by Step I (for each dataset) is passed to a correlation-based approach that finds association between target and non-target attribute pairs and then non-target and non-target attribute pairs in order to remove *irreverent* and *redundant* non-target attributes. In particular, two supervised feature selection approaches are *fused* here. Truly, the entropy-based approach emphasizes to identify the most relevant informative features, but some of these may have strong dependency among themselves and this results in redundant attributes. Certainly, finding the redundant attributes may not be resolved through this approach. However, the correlation-based approach has the capability to tackle this limitation. That is why, the entropy-based approach is applied in the first step and the correlation-based approach is applied in the subsequent step.

Importantly, optimal *threshold criteria* in Step II are decided based on the *trial-and-error* approach in order to retain the most informative as well as the essential features of the datasets, whereas the threshold value in Step I is set deterministically to select only the informative features. More specifically, two threshold values in correlation-based approach are set by applying the trial-and-error approach—one for checking relationship between non-target and target attribute pairs and the other for non-target and non-target attribute pairs. Actually, the threshold values in the proposed approach are decided based on the trial-and-error approach to yield maximum improvement or no loss in performance of the learners over almost all the chosen clinical datasets. More details about threshold values are described in the respective algorithm sections. It may be noted that the selection of threshold values (through trial-and-error approach) assists to solve MFSS NP problem approximately.

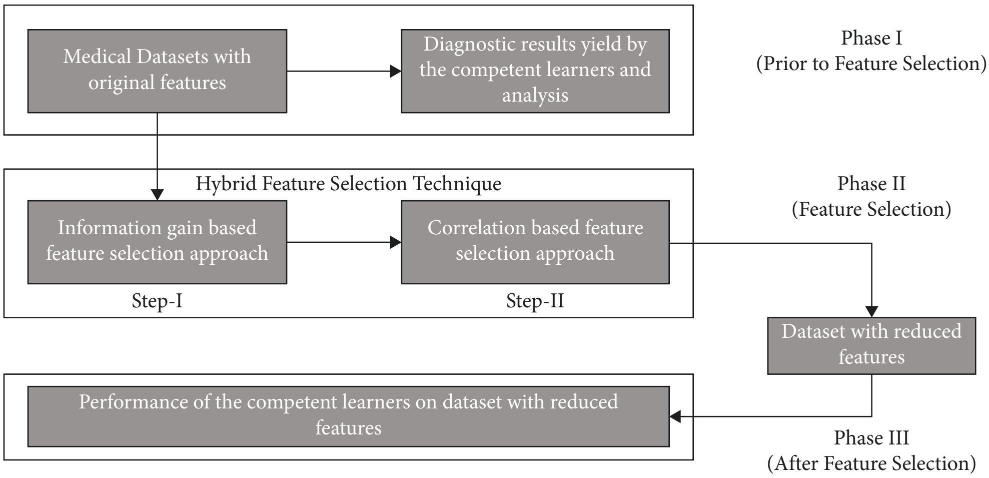
Medical Datasets with original features
Diagnostic results yield by the competent learners and analysis
Phase I
(Prior to Feature Selection)
Hybrid Feature Selection Technique
Information gain based feature selection approach
Correlation based feature selection approach
Step-I
Step-II
Phase II
(Feature Selection)
Dataset with reduced features
Performance of the competent learners on dataset with reduced features
Phase III
(After Feature Selection)

Figure 1: Block diagram of the proposed model.

3.1.1. Concept Adopted in Step I Using Information Gain Measure. The approach first computes *information gain* of each attribute and then finds their *mean* (i.e., mean_Gain = $\sum_{i=1}^{n}$ Gain $(A_i)/n$) and *standard deviation* (s.d.). Next, parameter Threshold_value is set as

$$\text{Threshold\_value} = \text{mean\_Gain} - \text{sd}. \tag{2}$$

Here, sd is the Gain standard deviation (later denoted as Gain_std.)

Finally, it filters out the attributes as follows. *If* any attribute has information gain *less than* Threshold_value, *then* it is discarded from the set of attributes. Thus, it reduces *search space* and enables to filter out the right informative attributes. The algorithmic version of this logic is presented in Algorithm 1.

Complexity Analysis. The algorithm is very simple and straightforward. Its running time (including entropy calculation time) is simply $O(n^2)$, where n is the number of attributes in the dataset. The algorithm is implemented in Python 3.9.

Note.The decided *Threshold_value* = (mean_ Gain – Gain_std) is statistically appropriate for filtering features. In particular, feature possessing *less* than the Threshold_value is assumed to have very *low* contribution in constructing expert system and may be ignored/removed from the feature set. This strategy is a kind of search-based filtration technique for dimensionality reduction task. Now, the joint entropy of the discarded features is checked to conclude if they are statically independent by using the inequality: $H(X_1, X_2, \ldots, X_k) \leq H(X_1) + H(X_2) + \cdots + H(X_k)$. If the inequality is satisfied, then they are statistically independent; otherwise, they are dependent. One may note that Shannon's joint entropy formula for two ensemble variables X and Y is defined as $H(X, Y) = -\sum_x \sum_y p(x, y) \log(p(x, y))$.

Now, if X and Y are dependent, we may not directly find the measure of dependency level by using the inequality. However, it can be obtained from correlation measures, and so correlation measure is used in the subsequent step of the ensemble approach.

Importantly, information gain (in comparison to Gini index) is preferred here to remove irrelevant attributes, since Gini index facilitates the bigger distributions not for lesser distributions having small count with multiple specific values.

3.1.2. Concept Adopted in Step II Using Correlation Coefficient Measure. The statistical measure, correlation coefficient, represents the strength of association between the variables. Its values lie in [−1, 1]. In this study, Pearson's product moment correlation coefficient is employed. The adopted correlation coefficient-based logic to reduce features is first graphically shown in Figure 2 (a wheel of complete graph) for easy understanding. The entire logic is described in 2 parts, namely, substeps I and II (as shown in Figure 2).

Logic to Decide Initial Threshold Values. The initial threshold values primarily decided by the *trial-and-error* approach (based on 10 trials) are here set as

$$\text{Threshold\_value1} = 0.4 * \max\{\text{Correlation\_coefficient}(A_i, C), \quad i = 1, \ldots, m\}. \tag{3}$$

Here, A_i is the i-th attribute and C is the class attribute.

$$\text{Threshold\_value2} = 0.75 * \max\left\{\text{Correlation\_coefficient}\left(A_i, A_j\right), i \neq j; i, j = 1, \ldots, k; k \leq m\right\}. \tag{4}$$

```
Suppose a dataset (DS) of classification problem has $n$ attributes, say $A_i$, ($i = 1, \ldots, n$) and $N$ instances. So, DS $\in\mathbb{R}^{N\times n}$ refers to the given dataset with $N$ instances and $n$ dimensions (attributes). Now let $F$ be the set of original features of DS, where $F = \{A_1, A_2, \ldots, A_n\}$. Further, let $F_s$ denote the set of features, consisting of the most relevant informative features taken from F. Initially, $F_s = F = \{A_1, A_2, \ldots, A_n\}$.
Goal: elimination of the non-informative features from $F_s$.
**Input**: DS $\in \mathbb{R}^{N\times n}$ //Dataset with $n$ features and $N$ instances
**Output**: $F^n \longrightarrow F^m$ //Feature set with $n$ features to $m$ features, $m \leq n$
**Parameter:** Threshold_value
**Variables:** Gain_measure[1, ..., $n$], Gain_sum = 0, mean_Gain, Gain_suqare_diff = 0, Gain_sdt
  begin
  1. for each attribute: $A_i$ ($i = 1, \ldots, n$) of DS do
     begin
       1.1. Compute the entropy reduction measure for $A_i$ as: Gain ($E$, $A_i$) = Entropy ($E$) $-\sum_{v_j \in A_i}|E_{v_j}|/|E|$ . Entropy ($E_{v_j}$), where
            $v_j$ ($j = 1, \ldots, k$) denotes values of attribute $A_i$ and Entropy ($E$) = , where $|E|$ returns the number of
            examples in DS, and $p_m$ = $|E_m|/|E|$, where $|E_m|$ is the number of $m$-th class examples, out of $c$ classes.
       1.2 Gain_measure[$i$] = Gain($E$, $A_i$)//Stores $i$-th attribute's ($A_i$) information in $i$-th location of Gain_maesure[ ] array
     endfor
  2. for $i$: = 1 to $n$ do
       Gain_sum = Gain_sum + Gain_measure[$i$]
     endfor
  3. mean_Gain = Gain_sum/$n$//finds mean value (mean_Gain) of information gain measures
  4. for $i$: = 1 to $n$ do
     4.1Gain_square_diff = Gain_square_diff + square(Gain_measure[$i$] – mean_Gain)//square is the math function
     endfor
  5. Gain_sdt = $\sqrt{\text{Gain\_square\_diff}/n}$)//finds standard deviation (Gain_sdt) of information gain measures
  6. Threshold_value = mean_Gain – Gain_std
  7. for each attribute: $A_i$ ($i = 1, \ldots, n$) of DS do
     7.1 If Gain ($E$, $A_i$), ($i = 1, \ldots, n$) < Threshold_value, then discard $A_i$ from $F_s$, i.e., $F_s = F_s - \{A_i\}$//It is backward elimination.
     endfor
  end//of the algorithm
```

Algorithm 1: Algorithm for information gain-based feature selection approach.

Here, A_i and A_j are the i-th attribute and j-th attribute, respectively.

This logic is implemented using Python 3.9.

Note. Here, selection of the best features is done via removal of the irrelevant and redundant features from the feature set.

The high-level description of the logic is presented in Algorithm 2.

3.2. Complexity Analysis. The algorithm is very simple and it uses two iterations (in cascaded fashion), each continuing for a maximum of m times. So, its running time (including correlation coefficient computation time) is simply $O(m^3)$, where m is the number of attributes in the dataset. The approach is implemented in Python 3.9.

4. Experimental Results and Discussion

To assess the performance of the proposed feature selection model, several extensive experiments are performed over *seventeen* publicly available datasets drawn from several machine learning repositories (e.g., UCI [47], Kaggle [48], and OpenML [49]). In particular, the values of the *performance metrics* obtained (before and after applying the suggested hybrid approach) by six state-of-the-art and well-known learners over the datasets are presented in Tables 3–5, respectively. Importantly, each learner belongs to one specific type of learning strategy, such as J48 (a decision tree-based rule inducer [2]), JRip (Java version of Repeated Incremental Pruning to Produce Error Reduction (RIPPER)) [50] (a sequential covering algorithm), nature-inspired artificial neural network [51] (here, a 3-layer NN with $\lceil ((n+k)/2) \rceil$ neurons in hidden layer is taken, where input layer has n neurons—one for each input parameter, whereas output layer has k neurons—one for each class and each neuron uses *sigmoid* function), KNN [3] (a distance/instance-based learner), Naïve Bayes [4] (a probability-based learner), and support vector machine (SVM) [52] (with popular radiant basis function (RBF) kernel). In fact, to show the performance of the proposed feature reduction model rigorously, learners are chosen based on different strategies. The experiments over the learners are performed in Weka (Waikato environment for knowledge analysis) platform (http://www.cs.waikato.ac.nz/ml/weka). On the other hand, the proposed combined feature selection model is implemented using Python 3.9.

Used Performance Measuring Metrics. The results of the standard performance metrics—*prediction accuracy*, TPR, FPR, and AUROC obtained by the machine learning algorithms (applied before and after the proposed hybrid feature selection approach), are used to assess the

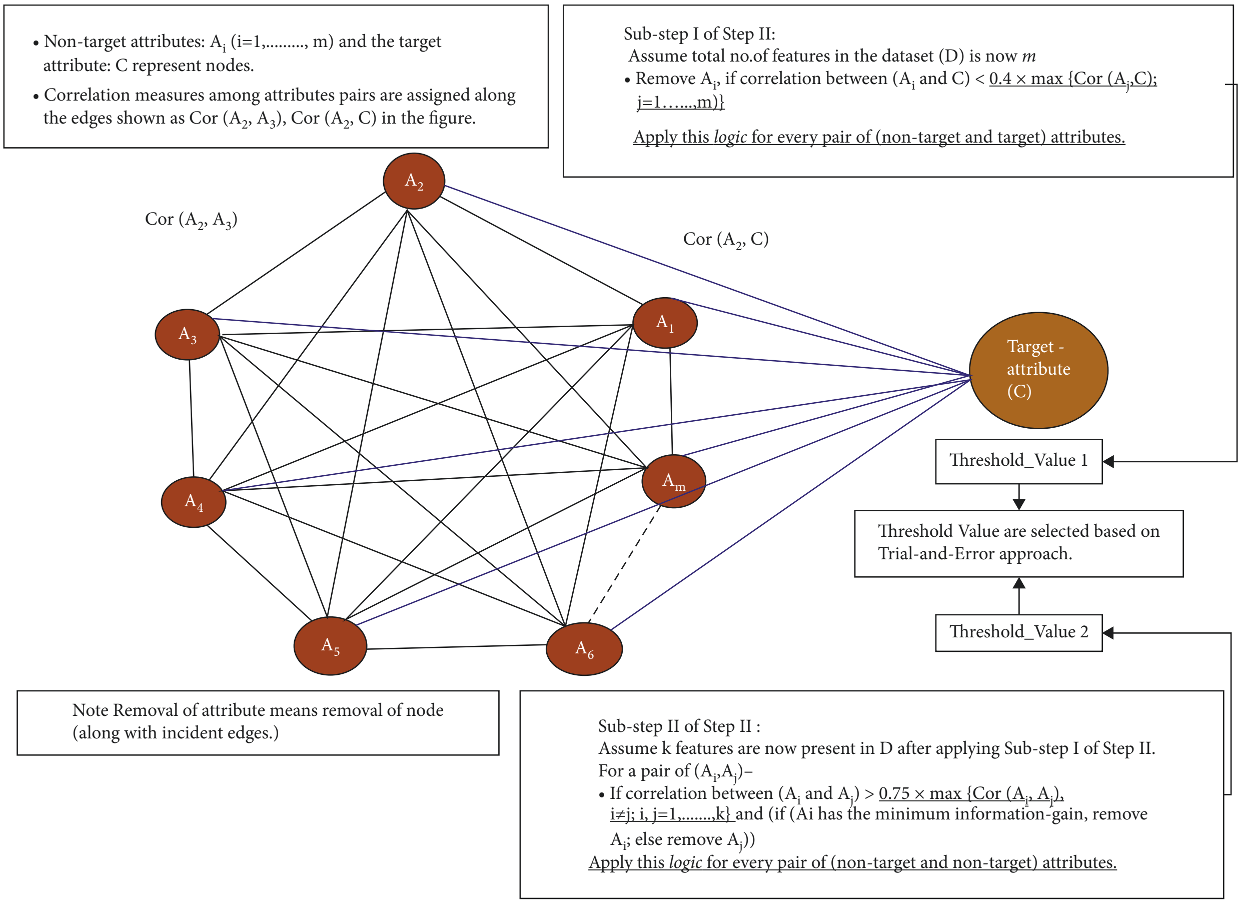
• Non-target attributes: A_i (i=1,........, m) and the target attribute: C represent nodes.
• Correlation measures among attributes pairs are assigned along the edges shown as Cor (A_2, A_3), Cor (A_2, C) in the figure.
Sub-step I of Step II:
Assume total no.of features in the dataset (D) is now *m*
• Remove A_i, if correlation between (A_i and C) < 0.4 × max {Cor (A_j,C); j=1…...,m)}
Apply this *logic* for every pair of (non-target and target) attributes.
Cor (A_2, A_3)
Cor (A_2, C)
A_2
A_3
A_1
A_4
A_m
A_5
A_6
Target - attribute (C)
Threshold_Value 1
Threshold Value are selected based on Trial-and-Error approach.
Threshold_Value 2
Note Removal of attribute means removal of node (along with incident edges.)
Sub-step II of Step II :
Assume k features are now present in D after applying Sub-step I of Step II.
For a pair of (A_i,A_j)–
• If correlation between (A_i and A_j) > 0.75 × max {Cor (A_i, A_j), i≠j; i, j=1,.......,k} and (if (Ai has the minimum information-gain, remove A_i; else remove A_j))
Apply this *logic* for every pair of (non-target and non-target) attributes.

Figure 2: A graphical representation of correlation-based feature selection logic.

effectiveness of the approach. In brief, the classification accuracy (CA) is the ability to predict the right classes correctly. TPR is the proportion of actual positives that were classified as positives. High TPR indicates that most of the positive cases in (TP + FN) are correctly labelled as positive. In medical models, we always expect high recall and low FPR. In fact, FPR explains the number of negative cases incorrectly predicted as positive. This measure together with a related measure, namely, the false negative rate, is extremely important in medical testing. Undoubtedly, FPR increases the mental worries, and so high FPR is always bad, and it is necessary to *minimize* those. Lastly, AUROC measures the quality of predictions irrespective of the chosen classification threshold. AUROC close to 1 is desirable.

Description of the Datasets. In this study, several datasets with different properties are used in the experiments to demonstrate the robustness and effectiveness of the introduced hybrid model. The primary characteristics of the datasets like *no. of non-target features*, *no. of classes*, *no. of instances*, and presence/absence of missing values are presented in Table 6.

A brief description of each the selected datasets is presented in Table 7.

Logic to Handle Missing/Null Values in the Datasets. Missing or null values exist in the datasets. To process the missing values in the attributes, the following strategy is adopted.

(i) If any attribute in a dataset possesses more than 60% missing values, then that attribute is simply dropped from the dataset.

On the other hand, any attribute with less than 60% missing values is processed as follows.

(ii) Missing values are replaced by *mean* value if the attribute type is *continuous*; otherwise, they are replaced by the value with *maximum* frequency.

Data analysis from Table 6:

(i) Number of datasets with features less than 10 : 2 (11%); these are, namely, *E. coli* and New Thyroid.

(ii) Number of datasets with features greater than 10 but less than 50 : 13(76%)

(iii) Number of datasets with features greater than 50 : 3(17.6%); these are, namely, Arrhythmia, Colon Cancer (2000 features), and Lung cancer.

(iv) Number of binary class datasets: 10 (58.8%); number of multi-class datasets: 7 (41%).

From the data analysis, it may be reported that the chosen datasets are of different sizes with diverse number of features (smaller to larger like big data). *For instance*, Arrhythmia,

```
Suppose the dataset (DS) has now $m$ attributes, say $A_{i,}i = 1, \ldots, m$ (after applying Step I) and $m \le n$. Therefore, DS $\in \mathbb{R}^{N \times m}$ now refers to the given dataset with N instances and m dimensions (features), and the feature set (Fs) of DS is now described as $F_s = \{A1, A2, \ldots, A_m\}$.
    (i) Correlation value between two variables: $x$ and $y$ (denoted as Cor($x$, $y$)), is computed by the formula: Cor($x$, $y$) $= \text{cov}(x, y)/\sqrt{\text{var}(x)\text{var}(y)}$, where cov($x$, $y$) $= \sum_{l=1}^{m} (x_i - \overline{x})(y - \overline{y})/m$ and var(x) $= \sum_{i=1}^{m} (x_i - \overline{x})^2/m$. Likewise, we get variance for $y$ (i.e., var($y$)).
    (ii) The approach includes provision of decremental and incremental scope of threshold values dynamically.
Goal: removal of irrelevant features (via non-target to target co-relationship) and removal of redundant (with less information gain) features (via non-target to non-target co-relationship)
Input: DS $\in \mathbb{R}^{N \times m}$ //Dataset with $m$ features and $N$ instances
Output: $F^m \longrightarrow F^k$ //Feature set with $m$ features to $k$ features and $k \le m$
Parameter: Threshold_value1, Threshold_value2//For storing threshold values
Variables:
  rnc[1, ..., $m$], rnn[1, ..., ($m - 1$)] [1, ..., ($m - 1$)], Atemp//Used matrices and temporary variable, Atemp
  Max_rnc = 0/* for capturing the maximum correlation value between class(target) and non-target attributes */
  Max_rnn = 0/* for capturing the maximum correlation value between non-target attribute and non-target attributes */
Step 1. Find correlation coefficient matrix ($C$) of size $(m + 1) \times (m + 1)$ for dataset DS with total $(m + 1)$ attributes including the class (target) attribute (placed at the last column of the matrix).
  Actually, the matrix: $C(m + 1) \times (m + 1)$, is represented by the following two arrays
    (i) rnc[1, ..., $m$]: a 1-D array to store correlation measures between non-target and class attribute pairs:
      (e.g., rnc[1] = Cor(1,class) (i.e., correlation measure between attribute 1 and the class), rnc[2] = Cor(2,class),...)
  and
    (ii) rnn[1, ..., $m$][1, ..., $m$]: a 2-D array to store correlation measures between non-target and non-target attribute pairs
      (e.g., rnn[1][2] = Cor(1, 2), i.e., correlation measure between attributes 1 and 2), ...
Step 2. Find the maximum value from rnc[$i$], $i = 1, \ldots, m$ and store that at Max_rnc
Step 3. Find the maximum value from rnn[$i$][$j$], $i = 1, \ldots, (m - 1)$; $j = (i + 1), \ldots, m$ and store that at Max_rnn.
  Substep I of Step II./* Removal of irrelevant non-target features from $F_s$ using Threshold_value1 set for (non-target and class) attributes pairs */
  Step 4. Threshold_value1 = 0.4 * Max_rnc//i.e., (40% of Max_rnc)
  Step 5. For each attribute: $A_i$ ($i = 1,2, \ldots, m$) of the current $D$ do
    Step 5.1. If rnc[$i$] < Threshold_value1, then discard $A_i$ from $F_s$, i.e., $F_s = F_s - \{A_i\}$.//Backward elimination
      endfor
  Step 6. If all the attributes are discarded from FS (i.e., $F_s = \Phi$), then perform the following substeps, else go to Step 7.
    Step 6.1. Threshold_value1 = Threshold_value1 – 0.1 * Max_rnc and take $F_s = \{A_1, A_2, \ldots, A_m\}$.
    Step 6.2. If Threshold_value1 > 0, then go to Step 5.
  Substep II of Step II./* Removal of redundant non-target features from the current Fs using Threshold_value2 set for non-target and non-target attribute pairs */
  Step 7. Threshold_value2 = 0.75 * Max_rnn//i.e., (75% of Max_rnn)
  Step 8. For each attribute $A_i$ ($i = 1, 2, \ldots, k - 1$) in the current DS do//$D$ with reduced features
    Step 8.1. For each attribute $A_j$ ($j = i + 1, 2, \ldots, k$) in the current DS do
      Step 8.1.1. If rnn[$i$][$j$] > Threshold_value2, then find Atemp = min_information_gain($A_i$,$A_j$) and
                  discard Atemp from $F_s$, (i.e., $F_s = F_s - \{\text{Atemp}\}$) if not already discarded.
      /* min_information_gain($A_i$, $A_j$) returns the attribute with minimum information gain between two attributes */
      endfor
    endfor
  Step 9. If all the attributes are discarded from $F_S$ (i.e., $F_s = \Phi$), then perform the following substeps, else got o Step 10.
    Step 9.1. Threshold_value2 = Threshold_value2 + 0.1 * Max_rnn and take the Fs obtained after Step 6.
    Step 9.2. If Threshold_value2 < Max_rnn, then go to Step 8.
  Step 10. Stop
```

Algorithm 2: Algorithm for correlation coefficient-based feature selection approach.

Colon cancer, and Lung cancer are significantly high dimensional datasets with small sample size; however, COVID-19 is an example of low dimensional database with a large number of samples. On the other hand, Primary Tumour is a multi-class dataset with twenty-two different kinds of classes.

4.1. Experimental Results. This section first describes the experiments conducted in the present research over the selected clinical datasets. The obtained results are presented in the tables, and the results are then analysed.

In the experiments, the number of selected features (including % age of dimension reduction), the classification accuracy (% age), TPR, FPR, and AUROC are used as the performance measures to evaluate the performance of the proposed model. First, a list of number of features reduced from the original datasets is reported in Table 8 on applying the proposed feature selection approach and its individuals. More specifically, the table describes, respectively, the *name of dataset* (DN), *number of instances* (NI), *number of features* (NF) in each original dataset, and

Table 3: Mean accuracy (% age) obtained over the datasets prior to and after feature selection.

Name of dataset	J48	JRip	KNN	ANN	Naïve Bayes	SVM	J48 + JRip
Arrhythmia	77.21, (78.31, 74.11, **78.98**)	75.44, (74.11, 75.22, **77.43**)	64.60, (65.92, 70.13, **71.09**)	68.14, (72.12, 76.99, **77.21**)	77.21, (77.01, 75.22, **77.85**)	75.88, (78.09, 77.21, **78.09**)	79.42, (77.65, 74.55, **81.07**)
Breast Cancer Wisconsin	94.42, (94.42, 94.70 **94.70**)	93.27, (93.84, 95.42, **94.70**)	95.56, (95.56, 95.27, 95.56)	95.99, (95.99, 95.13, **96.12**)	97.42, (97.28, 95.85, **97.89**)	95.70, (95.85, 95.70, 95.32)	94.84, (94.42, 94.87, **94.98**)
Colon Cancer	64.51, (65.47, 66.23, **67.12**)	64.51, (65.47, 66.23, **67.12**)	62.90, (63.19, 64.21, **64.75**)	***74.00, (79.00, 84.00, **89.00**)	35.48, (35.48, 35.48, 35.48)	64.51, (65.47, 66.23, **67.12**)	64.51, (65.47, 66.23, **67.12**)
COVID-19	98.17, (97.95, 96.06, **98.66**)	97.77, (97.62, 94.18, **97.87**)	98.08, (97.92, 94.79, 97.05)	98.14, (98.08, 97.27, **98.58**)	96.54, (96.72, 96.17, **97.07**)	96.74, (96.85, 96.36, **97.36**)	98.14, (97.88, 97.91, **98.37**)
Dermatology	94.53, (94.26, 93.98, **94.98**)	88.79, (90.16, 87.57, **90.05**)	94.80, (94.26, 94.35, **95.12**)	96.44, (96.44, 95.35, 95.62)	97.26, (97.81, 96.17, **97.27**)	96.17, (96.44, 96.23, **96.78**)	95.08, (93.98, 93.67, **95.82**)
E. coli	84.22, (84.22, 82.73, 83.73)	80.35, (81.25, 81.84, **81.54**)	80.35, (80.05, 79.46, 79.16)	85.71, (86.60, 84.82, **86.01**)	85.41, (85.41, 86.01, **86.01**)	83.63, (83.63, 84.22, **84.22**)	83.63, (82.73, 82.14, 82.84)
Heart (Cleveland)	52.14, (57.75, 52.80, **54.45**)	54.12, (53.79, 53.79, 54.12)	55.11, (53.46, 53.46, 54.82)	51.48, (56.10, 53.46, **57.09**)	55.77, (56.76, 53.46, **56.89**)	54.78, (55.77, 49.50, **55.77**)	53.13, (53.79, 53.79, **53.46**)
Heart (Hungarian)	81.29, (81.29, 81.29, 81.29)	79.93, (80.61, 78.23, **80.87**)	80.61, (78.51, 79.55, **81.07**)	81.29, (80.83, 80.12, **82.17**)	83.33, (82.41, 78.23, 83.08)	81.29, (80.61, 80.61, **81.95**)	80.27, (79.25, 80.61, **80.61**)
Heart (Swiss)	39.02, (39.02, 39.02, 39.02)	39.83, (41.46, 41.46, **42.27**)	32.52, (27.64, 36.58, **37.39**)	25.20, (25.20, 40.65, **36.58**)	26.82, (30.89, 39.02, **39.83**)	26.82, (34.14, 39.83, **40.65**)	39.83, (41.46, 41.46, **42.27**)
Hepatitis	63.22, (71.57, 76.42, **81.72**)	71.61, (74.96, 78.54, **80.37**)	59.35, (61.93, 63.87, **65.82**)	66.45, (70.80, 76.22, **79.17**)	69.67, (74.35, 78.41, **83.92**)	64.51, (68.93, 75.65, **81.76**)	69.67, (74.38, 78.42, **82.90**)
Indian Liver Patient	68.95, (66.72, 68.95 68.95)	69.81, (69.12, 71.18, **71.18**)	64.49, (65.35, 63.97, 63.97)	69.12, (69.92, 70.32, **70.32**)	55.74, (55.74, 55.23, **57.12**)	71.35, (71.35, 71.35, 71.35)	70.84, (68.27, 70.49, **72.14**)
Lower Back Pain	81.61, (81.61, 73.54, 80.51)	80.96, (82.25, 76.45, **82.17**)	62.25, (81.61, 57.41, **81.02**)	75.48, (84.51, 66.45, **84.13**)	77.74, (77.74, 74.19, **78.01**)	77.41, (78.70, 68.70, **78.78**)	81.61, (81.61, 76.12, **82.17**)
Lung Cancer	78.12, (78.12, 84.37, **84.37**)	78.12, (75.00, 81.25, **84.37**)	68.75, (71.87, 78.12, **78.12**)	65.62, (68.75, 65.62, **75.00**)	78.12, (77.01, 78.12, 78.12)	65.62, (68.75, 56.25, **71.87**)	81.25, (80.81, 84.37, **84.37**)
Malaria	65.57, (65.57, 65.57, 65.57)	62.90, (63.50, 64.39, **64.58**)	56.97, (56.37, 62.01, **62.47**)	55.48, (54.70, 56.08, **56.71**)	63.79, (63.17, 62.61, **64.01**)	59.94, (61.72, 61.42, **61.72**)	65.57, (65.28, 65.57, **65.68**)
New Thyroid	69.76, (69.76, 69.76, 69.76)	75.34, (75.34, 75.81, **75.81**)	80.46, (80.46, 79.87, **80.78**)	88.37, (88.37, 87.81, 87.81)	91.62, (91.62, 91.62, 91.62)	88.83, (88.83, 87.51, **88.94**)	72.09, (72.09, 72.23, **72.23**)
Parkinson	80.51, (80.51, 82.05, **84.10**)	87.69, (85.64, 86.66, **89.23**)	96.41, (95.81, 95.38, **97.82**)	90.76, (91.79, 92.30, **92.59**)	69.23, (68.71, 73.84, **80.51**)	87.17, (86.66, 86.15, **87.41**)	82.56, (81.53, 84.61, **85.12**)
Primary Tumour	71.09, (71.09, 73.15, **74.92**)	71.68, (71.68, 78.76, **78.92**)	73.15, (73.15, 71.68, 73.15)	69.91, (69.91, 73.74, **79.05**)	77.58, (77.58, 75.51, **78.17**)	75.51, (75.51, 79.94, **80.23**)	71.38, (71.38, 73.45, **74.33**)

Note: accuracy values obtained by step I only, step II only, and their combination are shown within parenthesis.

the *number of features* (*NF*) *reduced* individually by Step I and Step II and by their combination. Next, the significance of the introduced approach is affirmed through the standard performance metrics attained by the classifiers, namely, J-48, NB, JRip, KNN, ANN, NBs, SVM, and J48+JRip over the chosen benchmark datasets. Importantly, NB learner is chosen because it works better on datasets with independent features and the suggested approach focuses on identifying such features. In particular, the accuracy results for each dataset are shown in Table 3 as follows:

(i) Results obtained prior to applying the proposed hybrid approach.

(ii) Results obtained after applying the proposed approach and these results are shown within parenthesis as (*results obtained by applying* Step I separately, *results obtained by applying* Step II separately, and *results obtained by applying* Step I and Step II combined).

Likewise, the (TPR, FPR) and AUC (ROC) metrics obtained from the employed learners over the datasets are presented in Tables 4 and 5, respectively, following the *same order* as adopted in case of accuracy result presentation.

For better estimation of the performance metrics of the learners, each experiment is repeated 10 times based on 10-fold cross validation scheme. Thus, each entry of Tables 3–5 denotes the *mean* value of the findings obtained from 10 independent runs, where each run applies 10-fold cross validation scheme. Particularly, in each column corresponding to each row of the performance tables, the best *mean* value (if obtained by any learner after feature reduction) is marked in ***bold***. A head-to-head comparison of dimensionality reduction achieved by Step I singly, Step II singly, and their combination is reported in Table 9.

Recall that the trial-and-error approach is used here to decide the initial *threshold values* (mainly for Step II) to operate feature selection NP problem in polynomial time. In this model, 10 trials for each dataset are conducted. At each trial, 10% increment/decrement of max correlation (as

Table 4: Mean TPR and FPR in the format (TPR, FPR) obtained over the datasets prior to and after feature selection.

Name of dataset	J48	JRip	KNN	ANN	Naïve Bayes	SVM	J48 + JRip
Arrhythmia	(0.772, 0.232), (0.783, 0.222), (0.741, 0.266), (**0.790**, **0.212**)	(0.754, 0.246), (0.741, 0.258), (0.752, 0.248) (0.774, 0.231)	(0.646, 0.381), (0.659, 0.366), (0.701, 0.336), (**0.697**, **0.338**)	(0.681, 0.341), (0.721, 0.309), (0.770, 0.236), (**0.772**, **0.236**)	(0.722, 0.252), (0.770, 0.253), (0.752, 0.270), (**0.779**, **0.248**)	(0.759, 0.259), (0.781, 0.241), (0.772, 0.249), (**0.781**, **0.239**)	(0.794, 0.210), (0.777, 0.228), (0.746, 0.255), (**0.811**, **0.205**)
Breast Cancer Wisconsin	(0.944, 0.065), (0.944, 0.065), (0.947, 0.063), (**0.947**, **0.063**)	(0.933,0.079), (0.938, 0.076), (0.954, 0.059), (**0.947**, **0.063**)	(0.956, 0.063), (0.956, 0.063), (0.953, 0.064), (0.955, 0.063)	(0.960, 0.043), (0.960, 0.043), (0.951, 0.059), (0.946, 0.068)	(0.974, 0.019), (0.973, 0.022), (0.959, 0.043), (0.969, 0.032)	(0.957, 0.054), (0.959, 0.049), (0.957, 0.056), (0.953, 0.061)	(0.948, 0.059), (0.944, 0.067), (0.949, 0.058), (**0.950**, **0.053**)
Colon Cancer	(0.645, 0.645), (0.654, 0.621), (0.662, 0.615), (0.671, 0.607)	(0.645, 0.645), (0.654, 0.634), (0.662, 0.628), (0.671, 0.621)	(0.629, 0.654), (0.631, 0.647), (0.642, 0.642), (0.647, 0.637)	***(0.888, 0.400), (0.923, 0.500), (0.833, 0142), (0.928, 0.200)	(0.355, 0.355), (0.355, 0.355), (0.355, 0.355), (0.355, 0.355)	(0.645, 0.645), (0.654, 0.641), (0.662, 0.637), (0.671, 0.627)	(0.645, 0.645), (0.654, 0.639), (0.662, 0.636), (0.671, 0.631)
COVID-19	(0.982, 0.030), (0.980, 0.035), (0.951, 0.162), (0.967, 0.080)	(0.978, 0.032), (0.976, 0.038), (0.942, 0.142), (0.966, 0.060)	(0.981, 0.031), (0.979, 0.033), (0.948, 0.127), (0.971, 0.048)	(0.981, 0.028), (0.981, 0.030), (0.949, 0.152), (0.964, 0.109)	(0.965, 0.099), (0.967, 0.082), (0.930, 0.253), (0.925, 0.239)	(0.967, 0.126), (0.969, 0.131), (0.949, 0.173), (0.944, 0.234)	(0.981, 0.032), (0.979, 0.038), (0.949, 0.173), (0.966, 0.084)
Dermatology	(0.945, 0.013), (0.943, 0.013), (0.940, 0.018), (**0.950**, **0.009**)	(0.888, 0.025), (0.902, 0.024), (0.850, 0.040), (0.850, 0.039)	(0.948, 0.010), (0.943, 0.011), (0.913, 0.016), (0.907, 0.017)	(0.964, 0.007), (0.964, 0.007), (0.954, 0.009), (0.956, 0.009)	(0.973, 0.005), (0.978, 0.004), (0.962, 0.008), (0.973, 0.005)	(0.962, 0.007), (0.964, 0.007), (0.945, 0.012), (0.961, 0.008)	(0.951, 0.010), (0.940, 0.013), (0.918, 0.022), (0.940, 0.013)
E. coli	(0.842, 0.040), (0.842, 0.040), (0.827, 0.044), (0.837, 0.042)	(0.804, 0.068), (0.813, 0.056), (0.818, 0.062), (**0.815**, **0.052**)	(0.804, 0.054), (0.801, 0.054), (0.795, 0.051), (0.792, 0.051)	(0.857, 0.038), (0.866, 0.035), (0.848, 0.040), (**0.860**, **0.037**)	(0.854, 0.036), (0.854, 0.036), (0.860, 0.035), (**0.860**, **0.035**)	(0.836, 0.052), (0.836, 0.052), (0.842, 0.047), (**0.842**, **0.047**)	(0.836, 0.050), (0.827, 0.047), (0.821, 0.057), (0.828, 0.045)
Heart (Cleveland)	(0.521, 0.397), (0.578, 0.318), (0.528, 0.470), (**0.545**, **0.369**)	(0.541, 0.531), (0.538, 0.528), (0.538, 0.521), (0.541, 0.524)	(0.551, 0.247), (0.535, 0.252), (0.535, 0.267), (0.548, 0.250)	(0.515, 0.266), (0.561, 0.192), (0.535, 0.210), (**0.571**, **0.206**)	(0.558, 0.193), (0.568, 0.192), (0.535, 0.193), (**0.569**, **0.187**)	(0.548, 0.203), (0.558, 0.203), (0.495, 0.252), (**0.558**, **0.196**)	(0.531, 0.475), (0.538, 0.507), (0.538, 0.521), (**0.535**, **0.511**)
Heart (Hungarian)	(0.813, 0.254), (0.813, 0.254), (0.813, 0.254), (0.813, 0.254)	(0.799, 0.253), (0.806, 0.237), (0.782, 0.271), (**0.809**, **0.235**)	(0.806, 0.245), (0.785, 0.294), (0.776, 0.295), (0.796, 0.288)	(0.813, 0.225), (0.808, 0.238), (0.786, 0.277), (0.810, 0.232)	(0.833, 0.222), (0.824, 0.222), (0.782, 0.250), (0.831, 0.227)	(0.813, 0.225), (0.806, 0.253), (0.806, 0.266), (**0.820**, **0.215**)	(0.803, 0.263), (0.793, 0.290), (0.806, 0.266), (**0.806**, **0.266**)
Heart (Swiss)	(0.390, 0.390), (0.390, 0.390), (0.390, 0.390), (0.390, 0.390)	(0.398, 0.353), (0.415, 0.348), (0.415, 0.355), (**0.423**, **0.351**)	(0.325, 0.283), (0.276, 0.304), (0.366, 0.328), (**0.374**, **0.316**)	(0.252, 0.329), (0.252, 0.304), (0.407, 0.298), (**0.366**, **0.329**)	(0.268, 0.326), (0.309, 0.350), (0.390, 0.322), (**0.398**, **0.373**)	(0.268, 0.319), (0.341, 0.319), (0.398, 0.334), (**0.407**, **0.363**)	(0.398, 0.353), (0.415, 0.348), (0.415, 0.355), (**0.423**, **0.342**)
Hepatitis	(0.632, 0.424), (0.632, 0.424), (0.658, 0.375), (0.621, 0.437)	(0.716, 0.304), (0.710, 0.310), (0.735, 0.278), (**0.735**, **0.278**)	(0.594, 0.433), (0.594, 0.446), (0.619, 0.422), (**0.639**, **0.383**)	(0.665, 0.339), (0.658, 0.352), (0.632, 0.373), (0.659, 0.350)	(0.697, 0.320), (0.677, 0.344), (0.671, 0.347), (**0.708**, **0.312**)	(0.645, 0.363), (0.639, 0.373), (0.600, 0.413), (0.645, 0.359)	(0.697, 0.328), (0.697, 0.323), (0.735, 0.278), (**0.729**, **0.289**)
Indian Liver Patient	(0.690, 0.519), (0.667, 0.650), (0.690, 0.673), (0.690, 0.673)	(0.698, 0.562), (0.691, 0.590), (0.712, 0.571), (**0.712**, **0.553**)	(0.645, 0.458), (0.654, 0.483), (0.640, 0.517), (0.640, 0.517)	(0.691, 0.554), (0.700, 0.626), (0.703, 0.624), (**0.703**, **0.624**)	(0.557, 0.206), (0.557, 0.206), (0.552, 0.212), (**0.571**, **0.197**)	(0.714, 0.714), (0.714, 0.714), (0.714, 0.714), (0.714, 0.714)	(0.708, 0.540), (0.683, 0.676), (0.705, 0.674), (**0.721**, **0.518**)
Lower Back Pain	(0.816, 0.281), (0.816, 0.276), (0.735, 0.215), (0.805, 0.261)	(0.810, 0.274), (0.823, 0.221), (0.765, 0.275), (**0.822**, **0.224**)	(0.623, 0.452), (0.816, 0.203), (0.574, 0.517), (**0.810**, **0.218**)	(0.755, 0.316), (0.845, 0.231), (0.665, 0.458), (**0.841**, **0.237**)	(0.777, 0.179), (0.777, 0.174), (0.742, 0.332), (**0.780**, **0.172**)	(0.774, 0.385), (0.787, 0.379), (0.687, 0.620), (**0.788**, **0.374**)	(0.816, 0.266), (0.816, 0.239), (0.761, 0.255), (**0.822**, **0.232**)
Lung Cancer	(0.580, 0.580), (0.781, 0.424), (0.844, 0.332), (**0.844**, **0.332**)	(0.525, 0.578), (0.750, 0.436), (0.813, 0.412), (**0.844**, **0.332**)	(0.568, 0.436), (0.719, 0.448), (0.781, 0.356), (**0.781**, **0.356**)	(0.542, 0.527), (0.688, 0.460), (0.656, 0.540) (**0.750**, **0.368**)	(0.586, 0.433), (0.750, 0.436), (0.781, 0.356), (**0.781**, **0.356**)	(0.614, 0.407), (0.688, 0.460), (0.563, 0.645), (**0.719**, **0.381**))	(0.533, 0.600), (0.808, 0.392), (0.844, 0.332), (**0.844**, **0.332**)
Malaria	(0.656, 0.656), (0.656, 0.656), (0.656, 0.656), (0.656, 0.656)	(0.629, 0.649), (0.635, 0.613), (0.644, 0.618), (**0.646**, **0.608**)	(0.570, 0.582), (0.564, 0.622), (0.620, 0.568), (**0.620**, **0.565**)	(0.555, 0.533), (0.547, 0.554), (0.561, 0.566), (**0.567**, **0.559**)	(0.638, 0.555), (0.632, 0.558), (0.626, 0.544), (**0.640**, **0.547**)	(0.599, 0.587), (0.617, 0.586), (0.614, 0.592), (**0.617**, **0.590**)	(0.656, 0.656), (0.653, 0.659), (0.656, 0.656), (**0.657**, **0.654**)

Table 4: Continued.

Name of dataset	J48	JRip	KNN	ANN	Naïve Bayes	SVM	J48 + JRip
New Thyroid	(0.698, 0.698), (0.698, 0.698), (0.698, 0.698), (0.698, 0.698)	(0.753,0.421), (0.753, 0.421), (0.758, 0.450) (**0.758**, **0.450**)	(0.805, 0.431), (0.805, 0.431), (0.799, 0.432), (**0.807**, **0.428**)	(0.884, 0.249), (0.884, 0.249), (0.878, 0.283), (0.878, 0.283)	(0.916, 0.174), (0.916, 0.174), (0.916, 0.174), (0.916, 0.174)	(0.888, 0.248), (0.888, 0.248), (0.875, 0.274), (**0.889**, **0.246**)	(0.721, 0.585), (0.721, 0.585), (0.722, 0.582), (**0.722**, **0.582**)
Parkinson	(0.805, 0.344), (0.805, 0.330), (0.821, 0.297), (**0.841**, **0.290**)	(0.877, 0.251), (0.856, 0.313), (0.867, 0.366), (**0.892**, **0.246**)	(0.964, 0.040), (0.958, 0.057), (0.954, 0.043), (0.958, 0.051)	(0.908, 0.128), (0.918, 0.111), (0.923, 0.151), (**0.926**, **0.149**)	(0.692, 0.157), (0.687, 0.172), (0.738, 0.324), (**0.805**, **0.400**)	(0.872, 0.379), (0.867, 0.380), (0.862, 0.410), (**0.874**, **0.375**)	(0.826, 0.324), (0.815, 0.313), (0.846, 0.289), (**0.851**, **0.315**)
Primary Tumour	(0.711, 0.428), (0.711, 0.428), (0.732, 0.693), (**0.749**, **0.578**)	(0.717, 0.370), (0.717, 0.370), (0.788, 0.352), (**0.789**, **0.350**)	(0.732, 0.345), (0.732, 0.345), (0.717, 0.481), (0.732, 0.414)	(0.699, 0.370), (0.699, 0.370), (0.737, 0.459), (**0.791**, **0.378**)	(0.776, 0.276), (0.776, 0.276), (0.755, 0.432), (**0.782**, **0.368**)	(0.755, 0.320), (0.755, 0.320), (0.799, 0.395), (**0.802**, **0.381**)	(0.714, 0.422), (0.714, 0.422), (0.735, 0.699), (**0.743**, **0.634**)

Note: mean TPR and FPR obtained by step I only, step II only, and their combination are shown within parenthesis.

specified in Substeps I and II of correlation-based algorithm) is done. Based on 10 trials, the threshold values (as shown in equations (3) and (4) in Section 3) produce better performance over almost all the selected datasets, resulting in acceptable amount of data reduction.

Referring to Table 8, we get the dimension reduction (% age) of the used clinical datasets by Step I and Step II separately and by the combination of Step I and Step II, and the corresponding measures are presented in Table 9.

4.2. Discussion on the Experimental Results. Based on the empirical results yielded by the applied learners over the chosen datasets, some significant findings about the proposed hybrid feature selection approach are listed below.

(i) From Table 8, we may claim that the proposed approach is good enough to reduce *noise* from medical data. The justification behind its strength is analysed below from the results presented in Tables 3–5 and 9.

  (a) Table 3 reveals that each learner's classification accuracy over almost all the clinical datasets is improved after applying the combined approach. More clearly, the competent classifiers' mean accuracy (%) (presented in Table 3) increases in almost all cases after removing the non-informative, irrelevant, and redundant features. Notably, the performance of the NB classifier improves sufficiently over almost all the datasets, and it indicates that the introduced approach is good enough for reducing redundant features. The reason is that the NB learner works better over the dataset with independent features and the suggested approach can select such attributes. Thus, we may claim that the redundant features are removed from the datasets by applying this approach.

  (b) Table 4 reports that the metrics TPR and FPR bagged by the chosen learners over the datasets are improved considerably (i.e., TPR increases and FPR decreases almost over all the datasets) after applying the introduced feature reduction approach.

  (c) Table 5 deals with ROC-AUC metric, the most desirable performance metric of learners for clinical datasets. The head-to-head comparison of AUC values achieved by the learners between original datasets and the datasets with reduced features shows that AUC has increased over almost all the datasets after applying hybrid FS approach—it indicates a positive signal of the proposed approach.

  (d) From Table 9, it is clear that the proposed model results in more than 50% reduction of the features in 15 datasets (except *E. coli* and New Thyroid). In particular, 80% or more data reduction is made in 4 datasets, namely, Arrhythmia, Breast Cancer, Colon Cancer, Heart (Hung.), Heart (Swiss), and Lower Back Pain. Further, it is worth noting that the reduction in the datasets does not affect the classification accuracy, rather performance metrics are improved. More specifically, the original datasets contain about 50% to 80% redundant attributes, but the current hybrid approach is competent enough in removing these redundant attributes without affecting the classification accuracy. Consequently, removal of the features enables learning algorithms to speed up. Further, the comparison results presented in Table 9 exhibit that the proposed system is efficient in terms of data reduction as compared to the sole information gain-based approach and correlation coefficient-based approach. However, it may be noted that the individual correlation coefficient-based approach is better than the information gain-based approach alone. The reason is that perfectly correlated variables are truly redundant in the sense that no additional information

Table 5: AUROC obtained on the datasets prior to and after feature selection.

Name of dataset	J48	JRip	KNN	ANN	Naïve Bayes	SVM	J48 + JRip
Arrhythmia	0.774, (0.767, 0.735, **0.793**)	0.786, (0.735, 0.777, **0.788**)	0.631, (0.642, 0.690, **0.697**)	0.740, (0.762, 0.835, **0.825**)	0.809, (0.807, 0.844, **0.812**)	0.750, (0.770, 0.761, **0.771**)	0.840, (0.824, 0.821, **0.848**)
Breast Cancer Wisconsin	0.955, (0.955, 0.958, **0.958**)	0.947, (0.944, 0.948, **0.958**)	0.983, (0.984, 0.987, **0.987**)	0.988, (0.988, 0.983, 0.977)	0.993, (0.993, 0.987, 0.991)	0.952, (0.955, 0.951, 0.948)	0.981, (0.982, 0.977, **0.984**)
Colon Cancer	0.464, (0.469, 0.473, 0.479)	0.464, (0.471, 0.475, 0.481)	0.460, (0.465, 0.471, 0.478)	***0.500, (0.500, 0.500, 0.500)	0.518, (0.506, 0.513, 0.516)	0.500, (0.515, 0.523, 0.529)	0.464, (0.469, 0.473, 0.480)
COVID-19	0.995, (0.994, 0.959, 0.986)	0.980, (0.981, 0.923, 0.968)	0.998, (0.998, 0.988, 0.997)	0.996, (0.995, 0.984, 0.988)	0.990, (0.991, 0.953, 0.954)	0.921, (0.919, 0.855, 0.855)	0.995, (0.993, 0.970, 0.991)
Dermatology	0.976, (0.976, 0.968 **0.978**)	0.966, (0.902, 0.932, 0.950)	0.990, (0.987, 0.970, 0.967)	0.998, (0.998, 0.997, 0.996)	0.999, (0.999, 0.998, 0.998)	0.985, (0.987, 0.981, **0.989**)	0.991, (0.990, 0.975, 0.984)
E. coli	0.920, (0.920, 0.906, 0.906)	0.906, (0.902, 0.927, **0.913**)	0.878, (0.877, 0.869, 0.868)	0.953, (0.956, 0.870, **0.959**)	0.960, (0.960, 0.961 **0.961**)	0.943, (0.942, 0.939, **0.948**)	0.941, (0.941, 0.943, 0.938)
Heart (Cleveland)	0.597, (0.684, 0.545, **0.627**)	0.505, (0.496, 0.496, 0.491)	0.749, (0.763, 0.741, **0.768**)	0.701, (0.772, 0.757, **0.734**)	0.793, (0.807, 0.792, **0.809**)	0.718, (0.711, 0.679, 0.711)	0.603, (0.696, 0.551, **0.623**)
Heart (Hungarian)	0.708, (0.708, 0.708, 0.708)	0.786, (0.755, 0.751, 0.789)	0.833, (0.782, 0.844, 0.847)	0.855, (0.828, 0.864, 0.835)	0.897, (0.883, 0.868, 0.893)	0.794, (0.776, 0.770, 0.798)	0.757, (0.777, 0.755, 0.718)
Heart (Swiss)	0.455, (0.455, 0.455, 0.455)	0.487, (0.507, 0.504 **0.527**)	0.501, (0.473, 0.503, **0.543**)	0.462, (0.509, 0.570 **0.544**)	0.461, (0.449, 0.540, 0.453)	0.504, (0.515, 0.551, **0.550**)	0.478, (0.499, 0.496, **0.521**)
Hepatitis	0.577, (0.577, 0.615, 0.572)	0.668, (0.670, 0.685, **0.685**)	0.615, (0.568, 0.657, **0.664**)	0.674, (0.703, 0.676, **0.699**)	0.728, (0.739, 0.701, **0.742**)	0.641, (0.615, 0.594, **0.648**)	0.690, (0.680, 0.667 **0.692**)
Indian Liver Patient	0.678, (0.629, 0.584, 0.584)	0.582, (0.545, 0.586, **0.586**)	0.573, (0.568, 0.535, 0.535)	0.710, (0.728, 0.729, **0.729**)	0.726, (0.733, 0.734, **0.810**)	0.500, (0.500, 0.500, 0.500)	0.696, (0.630, 0.640, **0.721**)
Lower Back Pain	0.816, (0.838, 0.730, 0.807)	0.804, (0.822, 0.741, **0.818**)	0.585, (0.807, 0.529, **0.802**)	0.852, (0.925, 0.689, **0.912**)	0.879, (0.880, 0.804, **0.884**)	0.695, (0.704, 0.533, **0.716**)	0.870, (0.897, 0.755, **0.914**)
Lung Cancer	0.708, (0.708, 0.862, **0.862**)	0.589, (0.577, 0.582, **0.638**)	0.396, (0.570, 0.676, **0.725**)	0.676, (0.758, 0.681 **0.758**)	0.773, (0.792, 0.802 **0.821**)	0.558, (0.614, 0.459 **0.669**)	0.766, (0.785, 0.848, **0.843**)
Malaria	0.488, (0.488, 0.488, 0.488)	0.478, (0.497, 0.511, **0.515**)	0.507, (0.463, 0.484, 0.486)	0.538, (0.514, 0.513, 0.517)	0.567, (0.565, 0.600, **0.605**)	0.506, (0.516, 0.511, **0.514**)	0.478, (0.497, 0.511 **0.517**)
New Thyroid	0.480, (0.480, 0.480, 0.480)	0.690, (0.690, 0.744, **0.744**)	0.890, (0.890, 0.884, **0.892**)	0.929, (0.929, 0.921, 0.921)	0.968, (0.968, 0.968, 0.968)	0.852, (0.852, 0.842, **0.854**)	0.689, (0.689, 0.692, **0.692**)
Parkinson	0.769, (0.773, 0.799, **0.813**)	0.846, (0.757, 0.741, 0.805)	0.967, (0.962, 0.953, 0.964)	0.947, (0.953, 0.955, **0.957**)	0.858, (0.863, 0.858 0.817)	0.747, (0.743, 0.726, 0.751)	0.856, (0.831, 0.829, **0.872**)
Primary Tumour	0.670, (0.670, 0.512, **0.681**)	0.686, (0.686, 0.682, **0.687**)	0.756, (0.756, 0.691, 0.708)	0.740, (0.740, 0.703, **0.796**)	0.803, (0.803, 0.749, **0.808**)	0.717, (0.717, 0.702, 0.711)	0.714, (0.714, 0.689, **0.725**)

***Architectural limitation of WEKA is unable to support ANN learner to execute Colon cancer dataset with more features. That is why, 3-layer NN learner is implemented using Python 3.9 and the performance metrics of the learner in the tables corresponding to the Colon dataset are filled. AUROC obtained by step I only, step II only, and their combination are shown within parenthesis.

is gained by adding them. Besides, the correlation coefficient-based approach's data reduction capability demonstrates that clinical datasets often possess more redundant features, and removal of those is possible via a correlation-based approach.

(ii) It is well accepted to the researchers that the SVM learner is comparatively appropriate for datasets with high dimension but small number of classes. Examples include here Arrhythmia, Lung cancer, and COVID-19. But the improvement in the performance of this learner over the datasets is observed here after reducing the redundant features.

(iii) In some datasets, improvement in the used metrics yielded by some learners (not for all the chosen learners) is observed to be unchanged or very less or acceptably down, but amount (% age) of dimension reduction (i.e., noise reduction) is considerably good. This may be due to the adopted learning strategies by the learners that usually desire more features while training. Again, increasing the number of features in a dataset may not be always helpful to increase the classification performance of the data. In other way, increasing the number of features may often result in reduction of classification rate after a peak.

The proposed approach is not compared with other standard works in the literature, since almost all the approaches in the literature have chosen only few medical datasets (not a list of datasets) and few of them are disease specific. Of greater interest, the data reduction (% age) and classification accuracy performances of the present two-step system are compared with some standard studies for few specific clinical datasets, and these are presented, respectively, in Tables 1 and 2.

Referring Tables 1 and 2, the following *insights* may be highlighted in favor of the proposed approach.

(i) The presented model is generic (i.e., not disease specific). From Tables 1 and 2, it is observed that

Table 6: Summary of the selected datasets.

Sl. No.	Problem name	No. of non-target attributes	No. of classes	No. of examples	Missing values
1.	Arrhythmia	279	2	452	No
2.	Breast Cancer Wisconsin	10	2	699	Yes
3.	Colon Cancer	2000	2	62	No
4.	COVID-19	20	2	5434	No
5.	Dermatology	34	6	366	Yes
6.	*E. coli*	8	8	336	No
7.	Heart (Cleveland)	13	5	303	Yes
8.	Heart (Hungarian)	13	2	294	Yes
9.	Heart (Swiss)	13	5	123	Yes
10.	Hepatitis	19	2	155	Yes
11.	Indian Liver Patient	10	2	583	No
12.	Lower Back Pain	12	2	310	No
13.	Lung Cancer	56	3	32	Yes
14.	Malaria	17	2	337	No
15.	New Thyroid	5	3	215	Yes
16.	Parkinson	22	2	195	No
17.	Primary Tumour	17	22	339	Yes

Table 7: Dataset description.

Sl. No.	Dataset	Description
1.	Arrhythmia	This dataset aims to identify the heart arrhythmia or irregular heartbeat from ECG recordings.
2.	Breast Cancer Wisconsin	This dataset aims to identify whether a breast sample taken from a patient is cancerous or benign.
3.	Colon Cancer	The purpose of this dataset is to analyse tumour and normal colon tissues.
4.	COVID-19	The purpose of this dataset is to predict whether COVID-19 is possibly present or not from the symptoms.
5.	Dermatology	The goal of this dataset is to figure out what kind of erythemato-squamous disease there is.
6.	*E. coli*	This dataset aims to predict the cellular localization sites of *E. coli* proteins.
7.	Heart (*Cleveland, Hungarian*, and *Swiss*)	This dataset seeks to determine if a patient has the cardiac disease, which is represented by an integer value scale from zero to four.
8.	Hepatitis	The goal is to determine whether the person dies or lives.
9.	Indian Liver Patient	The goal of this dataset is to identify liver disease at early stages.
10.	Lower Back Pain	This dataset determines if a person is abnormal or normal using acquired physical spine details/data.
11.	Lung Cancer	The goal of this dataset is to predict whether or not someone has lung disease.
12.	Malaria	The main aim of this dataset is to identify from the symptoms whether the patient is suffering from malaria or not.
13.	New Thyroid	The aim of this dataset is to predict whether a person has thyroid or not.
14.	Parkinson	The aim of this dataset is to predict whether a person has Parkinson or not.
15.	Primary Tumour	The aim of this dataset is to identify the location of the tumour.

most of the feature reduction models are disease specific and they used heuristic/metaheuristic/combinatorial strategy to tackle the MFSS NP problem. Therefore, lack of generability and time consideration are the main drawbacks of the described studies. Actually, due to application of heuristic/metaheuristic/combinatorial strategy, the chance of increasing time may increase.

(ii) It is a good alternative of the standard dimension reduction models for clinical datasets. Data compression by the proposed model (as compared to most of the standard clinical dimension reduction models) is *noticeable* and performances over the datasets are quite encouraging. Some datasets of the comparison tables are analysed below.

(a) For Breast cancer dataset, the present approach exhibits performance wise better (or equal) to the studies [53, 58].

(b) In case of Colon cancer, the proposed approach attains better performance as compared to the methods [14, 53], reducing considerable amount of data reduction.

(c) About Lung cancer dataset, the % age of data reduction and the CA achieved by [54] are, respectively, 54% and 75%, whereas our approach bags these measures as 65% and 84.34% (although, not better than [58]).

Table 8: Dataset with reduced feature set after Step I and Step II.

Sl. No.	Name of dataset (DN)	Number of instances (NI)	No. of features (NF)	No. of nominal features	No. of features reduced by Step I only	No. of features reduced by Step II only	No. of features reduced by Step I and Step II in sequence (hybrid approach)
1.	Arrhythmia	452	279	Nil	94	243	247
2.	Breast Cancer (Wisconsin)	699	10		2	5	6
3.	Colon Cancer	62	2000	Nil	876	1638	1685
4.	COVID-19	5434	20		7	12	14
5.	Dermatology	366	34	Nil	3	21	23
6.	*E. coli*	336	8		1	1	2
7.	Heart (Cleveland)	303	13	Nil	5	5	6
8.	Heart (Hungarian)	295	13		5	8	10
9.	Heart (Swiss)	123	13	Nil	6	8	10
10.	Hepatitis	155	19		4	10	12
11.	Indian Liver Patient	583	10	Nil	3	5	5
12.	Lower Back Pain	310	12		4	7	9
13.	Lung Cancer	32	56	Nil	8	33	36
14.	Malaria	337	17		8	10	12
15.	New Thyroid	215	5	Nil	Nil	1	1
16.	Parkinson	195	22		4	14	17
17.	Primary Tumour	339	17	Nil	Nil	10	10

Note. Here, number of reduced features = number of features in the original dataset (drawn directly from the data repositories) − number of important features selected by any feature selection approach for further use. (i) *Actual features* are the features in a dataset available in the data repository and these are the features/attributes recommended by experts (physicians). (ii) *Selected features* are the features identified usually by the feature reduction model. (iii) *Nominal features* (e.g., id number, zip code, eye colour, and so on) are very common in the medical dataset. In particular, before applying the suggested ensemble feature selection approach over any medical dataset, such unimportant features are simply removed from the dataset.

Table 9: Comparison of dimension reduction (% age).

Sl. No.	Problem name	Dimension reduction (%) age by the information gain-based approach only (Step I)	Dimension reduction (%) age by the correlation coefficient-based approach only (Step II)	Dimension reduction (%) age by the hybrid approach (Step I + Step II)
1.	Arrhythmia	34	87	89
2.	Breast Cancer Wisconsin	20	50	60
3.	Colon Cancer	44	82	84
4.	COVID-19	35	60	70
5.	Dermatology	9	62	68
6.	*E. coli*	13	13	25
7.	Heart (Cleveland)	38	38	46
8.	Heart (Hungarian)	38	61	77
9.	Heart (Swiss)	46	61	77
10.	Hepatitis	21	53	63
11.	Indian Liver Patient	30	50	50
12.	Lower Back Pain	33	58	69
13.	Lung Cancer	14	59	64
14.	Malaria	47	59	71
15.	New Thyroid	0	20	20
16.	Parkinson	18	64	77
17.	Primary Tumour	0	59	59

(d) For Indian liver dataset, the study [57] reduces 90% data and attains 71.68% CA but the present study results in 50% data reduction and 72.14% CA. On the other hand, the work [55] yields only 70% CA without removing any feature.

(e) Parkinson disease is rarely experimented. The cited study [57] reduces only 31% data reduction and attains 90.78% CA, whereas the presented model results in 77% data reduction and yields 97.82% CA.

Few vital reasons in favor of achieving good performance by the proposed strategy are stated below.

(i) Note that Step I of the presented approach removes the irrelevant attributes, whereas Step II removes the correlated redundant attribute. However, while removing redundant attribute, the approach emphasizes to retain the informative attribute in between the correlated attributes. This prevents not only information loss of the data but also stops high dimensionality reduction. Practically, expected results are not achieved due to high dimensionality reduction.

(ii) The idea to decide threshold value criteria is conceptually justifiable. Introduction of these threshold values in the approach enhances the strength of the model.

5. Conclusions and Future Scope

Conclusions. Over the last 10 years, the growth of computer and database technologies has led to the rapid growth of large-scale datasets. Importantly, large-scale datasets give more accurate and valuable results. But they require high speed to process. One reason is that the number of dimensions in such datasets is very high, i.e., the key issue is the curse of dimensionality, which is mostly faced in the applications like pattern recognition, classification, and clustering.

A very natural question that may arise now is that earlier feature reduction was a very much active field due to hardware limitation. But, computational resource is now unlimited. So, at present, we may keep the larger datasets, since the larger the dataset is, the better it is for machine learning and knowledge discovery. However, there may still be redundant and irrelevant attributes in the large dataset which need to be removed from the dataset for achieving more effective results. Further, recently advanced machine learning approaches are able to handle the curse of dimension and large datasets. But these approaches are more suitable for large dataset (not for small size data). That is why, feature reduction is always welcome.

In this paper, a novel hybrid feature selection approach is proposed to predict the disease in a cost-effective way. We compare the classification accuracy, TPR, FPR, and AUC over the chosen seventeen datasets with the selected features using six individual well-known state-of-the-art learning approaches (namely, C4.5 (J48), JRip, ANN, KNN, Naïve Bayes, and SVM) and one hybrid learning approach (J48 + JRip).

(i) The list of the datasets (collected from several standard web data repositories) consists of both communicable and non-communicable disease datasets of smaller to larger dimension. The list includes the new dreadful disease—COVID-19.

(ii) Out of 17 datasets, 4 datasets, namely, Arrhythmia, Lower back pain, Malaria, and Parkinson, are rarely considered by the researchers.

(iii) In terms of the selected performance metrics, the overall performance of our method has been found to be very good for almost all these datasets. In summary, the presented approach works well for all the chosen medical datasets (i.e., it is not disease specific), and it can be an excellent alternative to the well-known data reduction approaches.

(iv) The approach is simple to implement, and computational complexity is $O(n^3)$, where n is the number of attributes in the dataset.

(v) The percentage of feature reduction by the new model is high.

(vi) The article gives a solid background information (including literature review) for researchers who are not familiar enough with feature reduction (specifically for medical datasets).

(vii) It assists to collect information data, saving data collection time.

Undoubtedly, with the help of the proposed method, redundant attributes can be removed efficiently from the datasets without sacrificing the classification performance. The proposed method of feature selection was also shown to perform well against feature selection with information gain.

5.1. Limitations

(i) The proposed method is not applied on more number of big medical datasets.

(ii) Two variables that are useless can be useful, but they are simply removed here.

Future Scope. We are in the process of searching the following.

(i) A variable that is completely useless by itself can provide a significant performance improvement when taken with others.

(ii) Two variables that are useless by themselves can be useful together.

Data Availability

The data used to support the findings of the study are included within the article.

Additional Points

Highlights. (1) Operating dimension is reduced in polynomial time. (2) Model is generic for medical datasets. (3) Data loss is prevented and diagnostic accuracy is improved. (4) Learning and diagnostic time is reduced. (5) Storage space is compressed.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors express great thanks to the Birla Institute of Technology for providing good infrastructure to carry out this research work.

References

[1] M. A. Hall, *Especially*, Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.

[2] J. R. Quinlan, *Programs for Machine Learning*, Morgan Kaufman, San Mateo, CA, U.S.A, 2014.

[3] E. Fix and J. Hodges, “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties,” Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX, U.S.A, 1951.

[4] R. O. Duda and P. E. Hurt, *Pattern Classification and Scene Analysis*, John Wiley & Sons, Hoboken, NJ, U.S.A, 1973.

[5] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, “Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success,” *Biomedical Journal*, vol. 330, no. 7494, pp. 765–772, 2005.

[6] S. Abdullah, N. R. Sabar, M. Z. A. Nazri, and M. Ayob, “An exponential Monte-Carlo algorithm for feature selection problems,” *Computers & Industrial Engineering*, vol. 67, pp. 160–167, 2014.

[7] D. K. Bhattacharyya and J. K. Kalita, *Network Anomaly Detection, A Machine Learning Perspective*, CRC Press (Book ch.), Boca Raton, FL, U.S.A, 2013.

[8] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[9] M. A. Hall, “Correlation-based feature selection for discrete and numeric class machine learning,” *Proceedings of Seventeenth International Conference on Machine Learning*, pp. 359–366, San Francisco, CA, USA, June 2000.

[10] N. Hoque, D. Bhattacharyya, and J. Kalita, “MIFS-ND: a mutual information-based feature selection method,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, 2014.

[11] R. W. Swiniarski and A. Skowron, “Rough set methods in feature selection and recognition,” *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.

[12] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[13] R. E. Schapire, “A brief introduction to boosting,” in *Proceedings of the 16th international joint conference on Artificial intelligence*, pp. 1401–1406, Sweden, July 1999.

[14] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, “Integration of multi-objective PSO based feature selection and node centrality for medical datasets,” *Genomics*, vol. 112, no. 6, pp. 4370–4384, 2020.

[15] Y. Saeys, I. Inza, and P. Larranaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[16] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[17] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[18] H. H. Inbarani, A. T. Azar, and G. Jothi, “Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis,” *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 175–185, 2014.

[19] P. Jaganathan and R. Kuppuchamy, “A threshold fuzzy entropy based feature selection for medical database classification,” *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2222–2229, 2013.

[20] A. Fernandez, S. D. Rio, N. V. Chawla, and F. Herrera, “An insight into imbalanced big data classification: outcomes and challenges,” *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 105–120, 2017.

[21] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, “Big data analytics in bioinformatics: a machine learning perspective,” *JOURNAL OF LATEX CLASS FILES*, vol. 13, no. 9, pp. 1–20, 2015.

[22] S. Majid and A. H. Maryam, “Distributed ensemble feature selection framework for high-dimensional and high-skewed imbalanced big dataset,” in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), INSPEC Accession Number: 21524210*, Orlando, FL, U.S.A, December 2021.

[23] D. López, S. R. Gallego, N. Xiong, F. Herrera, and S. García, “BELIEF: a distance-based redundancy-proof feature selection method for Big Data,” *Information Sciences*, vol. 558, pp. 124–139, 2021.

[24] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., “Analysis of dimensionality reduction techniques on big data,” *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[25] K. Chen, B. Xue, M. Zhang, and F. Zhou, “Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 3, pp. 446–460, 2022.

[26] Y. Hu, Y. Zhang, D. Gong, and X. Sun, “Multi-participant federated feature selection algorithm with particle swarm optimizaiton for imbalanced data under privacy protection,” *IEEE Transactions on Artificial Intelligence*, p. 1, 2022.

[27] K. Berahmand, A. Bouyer, and M. Vasighi, “Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes,” *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 1021–1033, 2018.

[28] K. Berahmand and A. Bouyer, “A link-based similarity for improving community detection based on label propagation algorithm,” *Journal of Systems Science and Complexity*, vol. 32, no. 3, pp. 737–758, 2019.

[29] K. Berahmand and A. Bouyer, “LP-LPA: a link influence-based label propagation algorithm for discovering community structures in networks,” *International Journal of Modern Physics B*, vol. 32, no. 6, Article ID 1850062, 2018.

[30] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” *Neural Inform Process System*, vol. 14, pp. 585–592, 2002.

[31] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[32] F. Chung, “Spectral graph theory,” *CBMS Regional Conference Series in Mathematics*, vol. 92, no. 92, pp. 1–212, 1997.

[33] S. Alelyani, “Stable bagging feature selection on medical data,” *Journal of Big Data*, vol. 8, no. 1, p. 11, 2021.

[34] J. Xie, M. Wang, S. Xu, Z. Huang, and P. W. Grant, “The unsupervised feature selection algorithms based on standard deviation and cosine similarity for genomic data analysis,” *Frontiers in Genetics*, vol. 12, Article ID 684100, 2021.

[35] B. K. Sarkar, “A two-step knowledge extraction framework for improving disease diagnosis,” *The Computer Journal*, vol. 63, no. 3, pp. 364–382, 2020.

[36] K. S. Fu, P. J. Min, and T. Li, “Feature selection in pattern recognition,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 6, no. 1, pp. 33–39, 1970.

[37] E. Amaldi and V. Kann, “On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems,” *Theoretical Computer Science*, vol. 209, no. 1-2, pp. 237–260, 1998.

[38] C. Tang, M. Bian, X. Liu et al., “Unsupervised feature selection via latent representation learning and manifold regularization,” *Neural Networks*, vol. 117, pp. 163–178, 2019.

[39] M. A. Hall and L. A. Smith, “Practical feature subset selection for machine learning,” in *Computer Science ’98*, C. McDonald, Ed., pp. 181–191, Springer, Berlin, Germany, 1998.

[40] K. Kira and LA. Rendell, “A practical approach to feature selection,” in *Machine Learning Proceedings’92*, pp. 249–256, Elsevier, Netherlands, 1992.

[41] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, Elsevier Inc, Burlington, MA, U.S.A, 2005.

[42] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[43] J. Tang, S. Alelyani, and H. Liu, “Feature Selection for Classification: A Review, Data Classification,” *Algorithms and applications*, p. 37, crc press, Boca Raton, FL.U.S.A, 2014.

[44] V. B. Semwal, J. Singha, P. K. Sharma, A. Chauhan, and B. Behera, “An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification,” *Multimedia Tools and Applications*, vol. 76, no. 22, pp. 24457–24475, 2017.

[45] Y. Masoudi-Sobhanzadeh, H. Motieghader, and A. Masoudi-Nejad, “Feature Select: a software for feature selection based on machine learning approaches,” *BMC Bioinformatics*, vol. 20, no. 1, p. 170, 2019.

[46] S. Fernández, J. A. C. Ochoa, and J. F. M. Trinidad, “A new hybrid filter–wrapper feature selection method for clustering based on ranking,” *Neurocomputing*, vol. 214, pp. 866–880, 2016.

[47] C. Blake, E. Koegh, and C. J. Mertz, *Repository of Machine Learning*University of California, Los Angeles, LA, U.S.A, 1999.

[48] Kaggle: https://www.kaggle.com/datasets/sammy123/lower-back-pain-symptoms-dataset.

[49] Openml: https://www.openml.org/search?type=data&sort=runs&id=1017&status=active.

[50] W. W. Cohen, “Fast effective rule induction,” in *Proceedings of the Machine Learning Proceedings 1995*, pp. 115–123, California, CA, U.S.A, July 1995.

[51] L. M. Fu, “Knowledge discovery based on neural networks,” *Communications of the ACM*, vol. 42, no. 11, pp. 47–50, 1999.

[52] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[53] N. Hoque, M. Singh, and D. K. Bhattacharyya, “EFS-MI: an ensemble feature selection method for classification,” *Complex & Intelligent Systems*, vol. 4, no. 2, pp. 105–118, 2017.

[54] P. D. Sheth, S. T. Patil, and M. L. Dhore, “Evolutionary Computing for Clinical Dataset Classification Using a Novel Feature Selection Algorithm,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, 2020.

[55] S. Murugesan, R. S. Bhuvaneswaran, H. N. Khanna, S. Keerthana, and Y. J. Nancy, “Feature selection and classification of clinical datasets using bioinspired algorithms and super learner,” *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–18, Article ID 6662420, 2021.

[56] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, pp. 82–93, 2019.

[57] G. I. Sayed, A. E. Hassanien, and A. T. Azar, “Feature selection via a novel chaotic crow search algorithm,” *Neural Computing & Applications*, vol. 31, no. 1, pp. 171–188, 2017.

[58] M. Mafarja, I. Aljarah, H. Faris, A. I. Hammouri, A. M. Zoubi, and S. Mirjalili, “Binary grasshopper optimisation algorithm approaches for feature selection problems-,” *Expert Systems with Applications*, vol. 117, pp. 267–286, 2019.

[59] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease,” *Cluster Computing*, vol. 22, no. S6, pp. 14777–14787, 2018.