EAI Endorsed Transactions
on Scalable Information Systems Research Article **EAI.EU**

Modified Filter Based Feature Selection Technique for Dermatology Dataset Using Beetle Swarm Optimization

J. Rajeshwari[1,*] and M. Sughasiny[2]

[1]Research Scholar
[2]Assistant Professor
[1,2]Department of Computer Science, Srimad Andavan Arts and Science College,
Tiruchirappalli, Tamil Nadu 620005, India.

Abstract

INTRODUCTION: Skin cancer is an emerging disease all over the world which causes a huge mortality. To detect skin cancer at an early stage, computer aided systems is designed. The most crucial step in it is the feature selection process because of its greater impact on classification performance. Various feature selection algorithms were designed previously to find the relevant features from a set of attributes. Yet, there arise challenges in selecting appropriate features from datasets related to disease prediction.
OBJECTIVES: To design a hybrid feature selection algorithm for selecting relevant feature subspace from dermatology datasets.
METHODS: The hybrid feature selection algorithm is designed by integrating the Latent Semantic Index (LSI) along with correlation-based Feature Selection (CFS). To achieve an optimal selection of feature subset, beetle swarm optimization is used.
RESULTS: Statistical metrics such as accuracy, specificity, recall, F1 score and MCC are calculated.
CONCLUSION: The accuracy and sensitivity value obtained is 95% and 92%.

Keywords: Skin cancer, Feature selection algorithm, LSI, CFS, Beetle swarm optimization, Classification performance.

Received on 22 February 2022, accepted on 12 July 2022, published on 15 July 2022

doi: 10.4108/eetsis.vi.1998

1. Introduction

World Health Organization (WHO) conducted a study and found that around one-third of the identified cancer patients are suffering from skin cancer, and further the rate of affected person kept on increasing all around the world [1]. The major reason for the increase in the number of skin cancer patients is the increased reduction of the ozone layer. This ozone layer serves as a barrier to the passage of ultraviolet radiation [2]. Normally, skin cancer is classified into two kinds such as non-melanoma and melanoma. According to a survey conducted in the US, nearly five million non-melanoma patients and eighty-seven thousand melanoma patients are diagnosed every year. Among these two types of skin cancer, melanoma is detected to be dangerous and start forming in cell pigment. Nearly, 75% of death is reported due to presence of a malignant type of skin cancer [3]. Medical experts are trying their level best to cure skin cancer by means of utilizing advanced treatment such as chemotherapy, radiotherapy and immunotherapy along with surgery. The analysis of the survival ratio over the past five years showed that the survival rate of advanced stages of cancer is under 15%. On the other hand, the survival rate of the initial stage of cancer is around 95%. This difference proved that the survival rate of cancer patients could be increased by providing early detection and treatment of cancer [4, 5].

[*] Corresponding author. Email: rajeshwarijcs@gmail.com

Therefore, the major objective to improve the survival rate of cancer is to offer early-stage diagnosis [6]. The traditional method used for the diagnosis of cancer is a detailed study conducted by doctors belonging to dermatology. But this process is found to be inaccurate, time-consuming and cost consuming [7]. To address these above-mentioned issues, numerous computer-based methods have been developed in recent years. These computer-based methods perform automatic detection and do not require any medical expert system [8]. So, it will be found to be time-consuming, cost consuming and accurate. The most commonly used computer-based methods are data mining techniques. The significant process involved in data mining technique is pre-processing, feature selection and classification [9, 10]. In order to achieve accurate detection of diseases, prior steps such as pre-processing and feature selection are considered vital. Still, there exist major challenges in finding the best feature selection technique [11, 12].

An effective feature selection technique is required for detaching redundant features from the dataset in order to reduce the complexity of processing [13]. By means of removing the irrelevant attributes from the dataset, the learning rate and predictive accuracy of the classifier can be enhanced. The significant factor used for defining the feature selection technique is evaluating measures and search procedure [14, 15]. The feature selection process can be divided into three types such as filter-based approach, wrapper-based approach and hybrid approach. However, in any type of feature selection technique, searching for an optimal subset can lead to better training and testing performance [16, 17]. The in-depth search of every feature subspace can provide an optimal and effective solution. However, it is quite impossible even to use the medium of feature subsets. Some of the existing feature selection technique includes conditional entropy, chi-squared test, Mutual information [18], and information gain and so on. Yet, these existing system does not provide an effective selection of feature subset. So, in the proposed research effective feature selection process is designed to improve classification performance.

1.1 Contribution of Research Work

The major contribution of the research work is given below,

- Early-stage detection of skin cancer is achieved using a hybrid feature selection technique through merging feature transformation along with a filter-based feature selection technique.
- Reduction in process complexity and improved accuracy rate is attained through reducing dimensionality of the dataset using Latent Semantic Indexing (LSI)
- Effective selection of relevant features is done using correlation-based feature selection, which is a kind of filter-based approach.
- Optimal number of feature subsets is selected using a swarm-based optimization algorithm, namely beetle swarm optimization, is used.
- To optimize a number of feature subsets, error rate and classification performance are considered as the fitness function.

1.2 Organization of the Paper

The upcoming portion of the manuscript is organized as follows, and section 2 describes some of the research articles related to the existing feature selection process used for disease prediction. Section 3 provides a detailed description of the proposed methodology. Section 4 explains the results gathered through the implementation of the proposed framework. Section 5 finally concludes the entire research work.

2. Related Works

Numerous feature selection algorithms have been introduced for selecting the most relevant features necessary for better classification. The most commonly used feature selection technique is filter-based and wrapper based. Some of the existing feature selection technique used in data mining approach is reviewed below.

M. J. Abinash et al. [19] performed an analysis on wrapper related feature selection algorithm used for selecting features in the leukaemia dataset. At present, data mining plays a significant role in almost every field. The volumes of structured, semi-structured, and unstructured data were in particular large. It was the basis for the mining of data. Thus it was very difficult to discover knowledge using these data. The interdisciplinary biological and IT sciences were used by bioinformatics, the expression of genes or microarray data is analyzed by some software. These gene data were increasing and higher so that the cancer classification data were analyzed. The present work designed a selection of cancer classification SVM-based wrapper features. Two functional selection algorithms were applied to the cancer dataset. Of them, the wrapper-based SVM method was best used for the cancer classification function selection.

Laith Abualigah et al. [20] had designed a feature selection method for data mining tasks using a hybrid Sine Cosine Algorithm and Genetic Algorithm. The selection of features (FS) was an actual problem that could be solved with optimization techniques. These techniques have created a predictive model which minimizes the prediction errors of classification by choosing the important or informative features by removing in the original dataset redundant, noisy and irrelevant attributes. Using Sine Cosine Algorithm (SCA) and Genetic Algorithm (GA), called SCAGA, a new hybrid feature selection method was presented. The SCAGA was developed using: classification accuracy, worse fitness, mean fitness, best fitness, average features and standard deviations to the following assessment criteria.

Hongqiang Lyu et al. [21] had designed a filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining. Recent risk has been revealed that a certain part of the redundancy, known as irrelevance redundancy, could be involved in the minimum-redundancy component of this method in the classical filter process of minimum-redundancy-maximal-pertinence (mRMR). The present work was conceived with a filter approach to overcome this problem. To check the performance of OMICFS, two types of biomedical datasets were used with other filter selection methods, both as regards classification accuracy and computational efficiency. The results have shown that OMICFS, in most cases, exceeds the other methods.

Smita Chormunge et al. [22] had designed correlation-based feature selection with clustering for high dimensional data. Traditional algorithms for the selection of features are not large in size. This work introduced a method for solving the problem of dimensionality where the clustering was incorporated into correlation measures to produce good functional subsets. First, irrelevant features were eliminated by the k-medium clustering method and then by correlation measure from each cluster, non-redundant features were selected. Microarray and Text datasets have evaluated the current method and compared the results to other renowned functional selection methods using the Naive Bayes classification system. In percentage terms, criteria have been used to check the accuracy of the present method with a variety of relevant features.

Indu Jain et al. [23] had designed correlation-based feature selection integrated with particle swarm optimization for the selection of gene expression and cancer prediction. DNA microarray technology for the detection and classification of cancer was recently used. It provides better insight into many cancer-related genetic mutations in a cell. However, a great challenge was posed by thousands of gene expressions measured by microarray for every biological test. Before cancer classification, many statistical and machine learning methods were employed to obtain the most relevant genes. In this present work, a two-phase hybrid model was developed for cancer classification. In the developed framework, experimental analyses were carried out, and 100% accuracy was obtained.

Erick Odhiambo Omuya et al., [24] designed a hybrid feature selection technique through merging information gain and principal component analysis. Feature selection was considered as a significant process in any classification process. Because the performance of classifier mainly depend on the selected features sets. However the irrelevant data present in the dataset creates negative influence on the classification performance. So, along with selecting optimal feature sets reducing the dimension of the dataset was also essential. In this present study principal component analysis was used for reducing dimension and information gain was used for selecting necessary attributes for classifying using naive bayes classifier.

S Sivaranjani et al., [25] had developed machine learning technique through incorporating PCA for dimension reduction and step forward and backward technique for feature selection in order to predict diabetics. Diabetics was a common disease which occurred due to various factors such as age factor, stress, genetic and lifestyle. Various machine learning algorithms such as SVM and RF was used for diabetics' prediction. Selected features after dimensionality reduction was sent for classification.

Md. Nurul Ahad Tawhid et al., [26] had designed classification framework through incorporating spectrogram images and textural feature extraction technique to detect brain disease. In this present framework initially the input signals were filtered to remove noise and artifacts. Next the signals were segmented into small chunks and from those chunks histogram was generated using short time Fourier transform. Then, histogram based textural features are extracted and optimal features were selected using PCA. Finally the features were sent into ML classifiers for prediction.

From the above-mentioned article, it is found that various feature selection technique is designed based on filter, wrapper and optimization technique. But accurate prediction with a lesser error rate is not achieved using these existing systems. At the same time, the selection of optimal feature subset is also not achieved in any of the existing algorithms. Therefore, a hybrid feature selection algorithm is designed in this proposed research for the effective selection of features from the dataset.

3. Proposed Methodology

Skin cancer ruined the whole world due to its harmful effects. Because of this, several people are affected. Numerous research exposes that all the tumour related problem is cured only if it's identified in the initial stage especially skin tumour is diagnosed in the primitive stage crucially diminishing the mortality. In fact, Oncologists also suffocating to prognosis the initial stage of skin tumour is a consequential issue. Thus, an easier process to diagnose skin cancer in the primitive stage is furthering to the dermatologists. Software assisted technology play a vital role in the implementation of data mining approaches in software formation effectively nowadays. These methods enhance the rapidity in prognosis steps than a human-made diagnosis. Human-made diagnosis has some mistakes, but in this, machine-made techniques are not. It can ease the work of radiologists and dermatologists to the estimation of skin tumours is very effective and comfortable. Pre-processing, feature selection and classification are the phases elaborated significantly in the data-mining process.

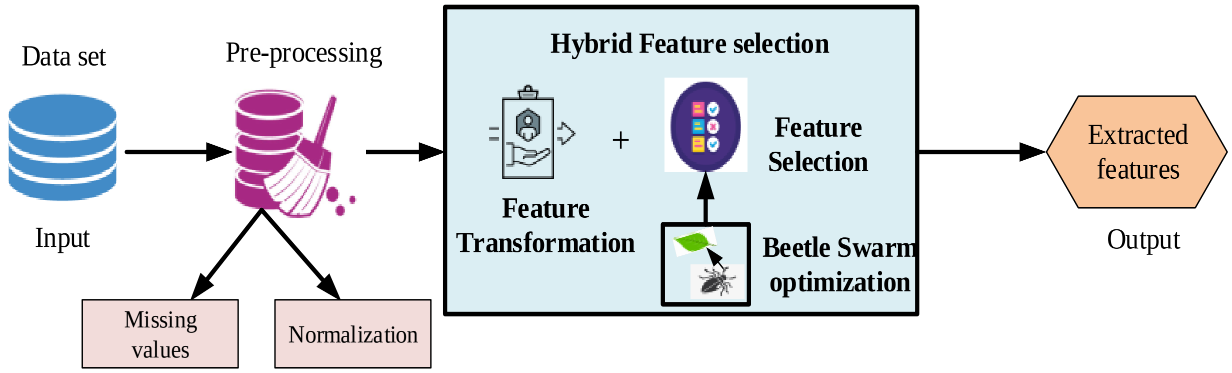

Figure 1. Architecture of Proposed method

Analyze the dataset and further move to pre-processing technique. In this pre-processing, the proposed method fill the missing data values with numerous techniques and also perform the normalization process. The missing values are filled by the skewness technique, and the z-score normalization is used. And next step is to feature the selection approach. In this feature, selection has two methodologies that are feature transformation and feature selection. So it is known as Hybrid feature selection. Feature transformation used the Latent Semantic Indexing technique. Hybrid feature selection is used to optimize the k-subset features using Beetle swarm optimization.

3.1. Pre-processing

In this proposed method, the pre-processing technique is used to fill the lost or missing values by skewness equation (1). Missing values arise for several bases in a dataset that leads to partial attribute values. Ignorance, equipment failure, data corruption, and Measurement error are a few of the major causes for missing values in a dataset [27]. In the data mining approach, it is difficult to find the missing value in a dataset.

Missing Values (Skewness)

The missing values in a dataset are filled by the following equation,

$$\mathrm{Skew}\left(\mathrm{d_i^{obs}}\right) = \frac{\sum_{\mathrm{k=1}}^{\mathrm{N_i}}\left(\mathrm{d_{ik}^{obs}} - \mathrm{\bar{d}_i^{obs}}\right)^3}{\mathrm{N_i * var}\left(\mathrm{d_i^{obs}}\right)^{1.5}} \qquad (1)$$

where, d_{ik}^{obs} is the kth non-missed value in the ith feature and d_i stands for i^{th} attribute, $var(d_i^{obs})$ is the variance of the ith feature, $\bar{d}_i^{obs}$ is the average of non-missed values in the ith feature, and N_i is the number of those instances that don't contain missed values for ith attribute.

Normalization (z-score)

Normalization is one of the finest pre-processing methods [28]. For that, numerous techniques of normalization such as min-max normalization, z-score normalization, decimal scaling, standardized moment etc. utilized in dataset normalization. This proposed method using z-score normalization.

The following equation (2) determines the z-score normalization attribute of the normalized value.

$$z = \frac{\upsilon - \mu_{\mathrm{A}}}{\sigma_{\mathrm{A}}} \qquad (2)$$

The mean and standard deviation of the attribute is μ_{A} and σ_{A} respectively. The υ and z are demonstrating the original and normalized feature values correspondingly.

3.2. Feature selection algorithm

The method of acquiring a subset from the original feature set corresponding to few standards is known as feature selection; however, it chooses the appropriate features of the dataset. Repeated and unrelated features are eliminated from playing a vital role in compacting the data processing, simplifies learning results, learning accuracy and reduce learning time in the feature selection method. Both dimension reduction and filter approach are coupled to enhance, so it's known as hybrid feature selection algorithm.

Feature transformation

The information is still maintaining when the data will be altered, known as Feature transformation. The features are converted from one form to another form is a general function of Feature transformation.

(i) Latent Semantic Indexing (LSI):

The dimensionality of the dataset is deducing in accordance with enhancing the efficiency of the method is known as feature transformation. In this work, the dimension of the dataset is deducing with the help of Latent Semantic Indexing (LSI) [29]. The dimension of the Term Document Matrix (TDM) is diminishing by the idea of singular value decomposition (SVD) utilized by Latent Semantic Indexing (LSI). The issue of synonymy and polysemy words are noticed by LSI is the main objective Eqn(3). The singular value decomposition (SVD) method develops the semantic space of LSI in linear algebra. Retrieving and demonstrating the uniqueness of semantic of phrases for vector spacing and deduction of dimension utilized the LSI.

$$TDM = USV^T \qquad (3)$$

where, V is $n \times n$ unitary matrix which is the right singular vector of TDM, U is an $m \times m$ unitary matrix which its column is the left-singular vector of TDM, and S is $m \times n$ diagonal matrix of singular values.

LSI employs by maintaining the first k diagonal value of S and the remainder is zero. The diagonal value of S is sorted descending is mentioned. Also, U and T matrix is deduced by conserving its first k columns and rows correspondingly. The compact form of TDM is evaluated as following Eqn (4).

$$TDM_k = U_K S_K V_K^T \qquad (4)$$

Feature selection

In the feature selection method, the input is received that is deduced dataset dimension. From existing features, a correlation-based feature selection method is created to choose the high relevant feature. To identify the feature subset for optimal number in this method, incorporating the beetle swarm optimization is an algorithm for optimization. The error rate is deduced, and classification performance is high in the optimization problem is the main goal of this process.

(i) Filter-based Feature Selection

The significant feature incorporating with the feature subset is evaluated by the main features of the data. There are two distinct parts. They are Rank-based and subset Evaluation based Univariate statistical methods to detect the rank of a single feature lacking inter-correlation among features [30]. This method fails to evaluate the redundant features and the multivariate statistical method to determine the rank of the whole feature subset based on subset feature selection.

(ii) Correlation-based Feature Selection (CFS)

A single feature is not detected by Information Gain Ratio (IGR). The feature subset is not determined by the IGR. The univariate method is less effective than multivariate feature selection to detect the diagnosis of tumour about datasets. Univariate filters are an Information gain that doesn't make communication among features. These problems were conquered by Correlation-based Feature selection in multivariate filters. One of the multivariate filter approaches is Correlation-based Feature Selection (CFS) to choose the better feature subset that has a high correlation with the target class, however uncorrelated together in single feature selection [31]. So, this proposed method uses correlation-based feature selection. On the basis of the following hypothesis, "Good feature subsets contain features highly correlated with classes, yet uncorrelated to each other", rank the feature subsets rather than individual features.

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (5)$$

where, $Merit_s$ is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ represent the average feature-class correlation and $\overline{r_{ff}}$ represent the average feature-feature inter-correlation. Between the features, CFS examines the feature subsets based on the redundancy degree. The single feature subset is less inter-correlation with the class. However, its correlation is huge, according to the evaluator. Subset evaluator utilizes a certain amount of numeric in conditional entropy. Therefore, the highest correlation in the class and increase the features with the support of replication of search. Determine the advantage of feature subset is investigated by correlation-based feature selection. To diminishing the feature repeatability and the relevancy of the feature is enhanced by the correlation-based feature selection method in Eqn(5). Each and every feature subset redundancy degree is determined by correlation-based Feature Selection (CFS). Evaluating the class and coefficient of correlation is utilized by attributes in the correlation between subset and the inter-correlations among the class by the features. The correlation between the classes and the features is proliferated.

Correlation-based Feature Selection (CFS) evaluates the best feature subset hence deduce the improving inter-correlation. Best-first search, backward elimination, genetic search, bidirectional search and forward selection are the techniques to identify the best feature subset in CFS used by Eqn (6)

$$r_{zc} = \frac{\overline{K r_{z\iota}}}{\sqrt{K + K(K-1) r_{ii}}} \qquad (6)$$

Where $r_{ii}$the average inter-correlation among subset is featured, k is the number of subset features, r_{zi} is the average of the correlation among the subset features and the class variable, and r_{zc} is the correlation among the summed feature subsets and class variables [32].

The Subset Features tool divides the data into two subsets. Subset one will have L features, and subset two will have N - L features (with N being the number of features in the original dataset). The features are divided by generating random values from a uniform [0, 1] distribution.

With the combination of both filter-based and beetle swarm optimization, the filter-based feature selection is to choose the significant features from the original features, and the Beetle swarm optimization is to choose the feature subset from the deduced set of features [33]. Here,k is the number of subset features to optimizing by using Beetle swarm optimization.

3.3 Beetle Swarm Optimization

Insects are sensing by fluctuating chemical sensory systems to numerous atmospheric stimuli, and their manner is monitored. The main chemical receptor of insects is the antennae: olfactory and tactile effects, and few of that having an auditory function mostly. Opposite gender, food searching and select spawning sites to identify by these antennae. People frequently exploit insects' ability to emit chemicals with adverse reactions scents in order to attract or repel insects that are damaging to plants. Tremendously long antennae establish in long-horned beetle has four times the height of its body. There are two kinds of features in this type of long antennae. The first one is to analyze the surrounding circumstances. The next one is to determine the possible mates to fluctuating the body's antenna. The beetle will fluctuate in a unique direction, therefore, determined the enormous amount of smell. If not, it will fluctuate on the opposite side. The optimization method of the Beetle Antennae Search (BAS) Algorithm is devoid of a particular function and incline details [34]. The BAS has lower complexity in design and capability to fix the issue is time complexity are the main cons of BAS algorithm. Therefore, the higher-measurement implementation is not appropriate, and the redundant outcome is based on the beetle's earlier position. The optimization accuracy and effectiveness highly disrupt the selection of earlier positions. Proceeding enhancements to the BAS algorithm by increasing from individual to the group. This is the Beetle Swarm Optimization (BSO).

Step 1: Initialization

The concept of particle swarm algorithm imitated in the mathematical formation

$$X = (X_1, X_2, \dots . X_n) \quad (7)$$

$$B_i = (b_{i1}, b_{i2}, \dots \dots \dots . . b_{is})^T \quad (8)$$

B_i Demonstratses the speed of the beetle

Step 2: Fitness Function

The fitness formula is

$$\text{Fitness value} = f(X) = f(X_1, X_2, \dots X_n) \quad (9)$$

Step 3: Updating

The following equation (9) is a mathematical model for simulating behaviour is

$$X_{is}^{k+1} = X_{is}^{k} + \lambda V_{is}^{k} + (1-\lambda)\xi_{is}^{k} \quad (10)$$

Where λ is a positive constant, the speed of beetle is represented by V_{is} , ξ_{is}^{k} signifies the increase in beetle position movement, $s = 1,2, \dots . , S$; $i = 1,2 \dots, n$; k is the present number of iterations.

The velocity of the Beetle swarm optimization is

$$V_{is}^{k+1} = \omega V_{is}^{k} + c_1 r_1 \left(P_{is}^{p} - X_{is}^{p}\right) + c_2 r_2 \left(P_{gs}^{k} - X_{gs}^{K}\right) \quad (11)$$

Where $r_1 \, and \; r_2$ are two random functions in the range [0, 1], c_1 and c_2 are two positive constants, ω is the inertia weight, P represents the individual extremity.

Step 4: Terminating the value

Then the process will be terminated once the best solution is obtained.

BSO based Optimal Feature Selection

BSO is a new experimental optimization algorithm used for feature selection in the proposed part. The reason for using this BSO algorithm is that it converges faster when compared to other optimization algorithm. Both the Beetle Antennae Search (BAS) algorithm which is a new heuristics-based algorithm, and the Particle Swarm Optimization (PSO) algorithm are together to solve numerous real-time issues. The method of discovering a beetle group in the resolution area is the optimization process considered in this algorithm. The solution to the issue is demonstrated by the position of the beetle. The strength and velocity of the BSO algorithm are excellent. Also, it has high accuracy when it is trading with non-linear conditions [35].

Further, execute the particular process of the BSO algorithm. First, the initial position K of each beetle that is the number of subsets and the Fitness value F(x) will be created based on the number of subsets in Eqn(7).

$$AUC_{max} = \phi\left(\sqrt{(\mu_1 - \mu_0)^T (\Sigma_0 + \Sigma_1)^{-1} (\mu_1 - \mu_0)}\right) \quad (12)$$

Where μ_0, μ_1 are the population mean vector and $\Sigma_{0,} \, \Sigma_1$ are the covariance matrix. The efficiency to explore the error rate and evaluate data mining explores the accuracy of this method for best predictive analysis.

Mean square Error is for minimizing the error rate for Fitness function is

$$MSE = \frac{1}{N} \sum_{n=1}^{N} |x(n) - y(n)|^2 \quad (13)$$

Where the original and denoised N-length partial discharge signal is $x(n)$ and $y(n)$.

By using this Area under the Curve (AUC) and Mean square Error (MSE) formula, extracting feature selection of k number of subsets

Step1: Initialization

The initial position of each beetle k is the number of subsets in the following Eqn(14). The n beetle's population is demonstrated as

$$K = (K_1, K_2, \dots . K_n) \quad (14)$$

Step 2: Fitness Function

The largest value of Area under the Curve is the Fitness function of the number of subsets based on beetle swarm optimization. AUC is used to get high classification performance [36, 37]. The fitness formula is

$$Fitness\ value = Max(AUC) \quad (15)$$

$$Fitness\ value = Min(Error\ rate) \quad (16)$$

Step3: Updating the Value

Next will represent the position for each beetle is created based on the Eqn(9).

$$K_{is}^{p+1} = K_{is}^{p} + \lambda B_{is}^{p} + (1-\lambda)\xi_{is}^{p} \quad (17)$$

Where λ is a positive constant, the velocity of beetle is represented by B_{is} , ξ_{is}^{k} signifies the increase in beetle position movement, $s = 1,2, \dots . , S$; $i = 1,2 \dots, n$; p is the current number of iterations.

The velocity formula is Eqn(11)

$$B_{is}^{p+1} = \omega B_{is}^{p} + c_1 r_1 \left(M_{is}^{p} - K_{is}^{kp}\right) + c_2 r_2 \left(M_{gs}^{p} - K_{gs}^{p}\right) \quad (18)$$

Where $r_1 \, and \; r_2$ are two random functions in the range [0, 1], c_1 and c_2 are two positive constants, ω is the inertia weight, M represents the individual extremity. k_{gs} represents the global optimization of the number of subsets.

Step 4: Termination

Then the process will be iterated until getting the best solution. Once the K value gets the best subset solution, then the process will be terminated.

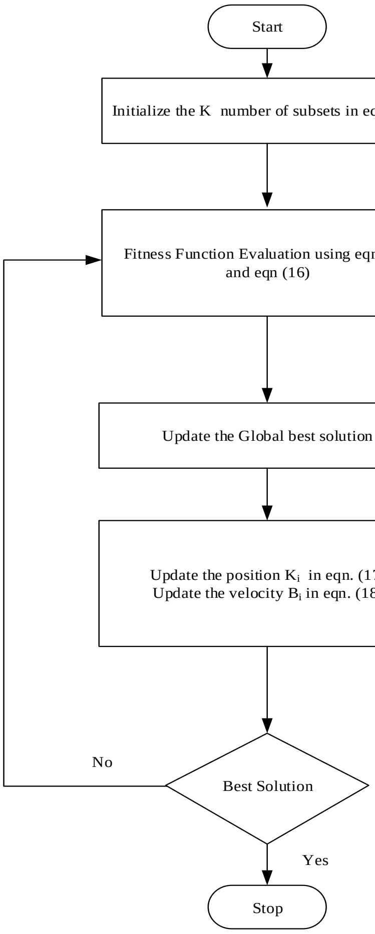
Start
Initialize the K number of subsets in eqn. (14)
Fitness Function Evaluation using eqn.(15) and eqn (16)
Update the Global best solution
Update the position K_i in eqn. (17)
Update the velocity B_i in eqn. (18)
No
Best Solution
Yes
Stop

Figure 2. Flowchart for BSO Optimization for Feature Subset

Table 1. Pseudocode for overall system

Pseudocode for Proposed Feature Selection Algorithm

```
Input Raw Dataset =X
{
For all data in dataset
   #Pre-processing
    {
    Step 1: Missing Values based on skewness using eqn. (1)
    Step 2: Normalization based on z-score using eqn. (2)
      }
     Pre-processed data →Pre-data
   #Feature Selection
     {
       Step 1: Feature Transformation based on Latent Semantic Indexing
               using eqn. (3) and eqn. (4)
       Step 2: Correlation based Feature Selection using eqn. (5) and eqn. (6)
            #Beetle Swarm Optimization
             {
              Step 1: Initialization using eqn. (14)
              Step 2:  Fitness evaluation using eqn. (15) and eqn. (16)
              Step 3: Updating using eqn. (17) and eqn. (18)
               }
             }
    }
    Outcome: Extracted Features
```

4. Results and Discussion

The execution of the proposed feature selection algorithm is done on Python 3.8 software to validate its functioning. The testing is performed with the help of Intel (R) Core ™ i5-3330s processor, CPU @ 2.70 GH, 8.00 Gb (7.88 Gb usable) Memory (RAM) and System type of 64-bit operating system. Initially, in this proposed algorithm, the dataset related to dermatology is collected from the UCI repository [38]. The acquired dataset contains around thirty four attributes, out of which thirty three attributes are linear values, and one of the attributes is nominal. The information contained in the attributes is of two types. Some of the attributes contain clinical information, while the rest of the attributes contains histological information. The total number of instances in the dataset is 366 which signifies the number of rows. This collected dataset is taken as input for the proposed algorithm in the initial step. Following that, in the second step, the input data is pre-processed using two pre-processing techniques, such as missing value imputation and normalization. The pre-processed data is obtained as output from this step. In the third step, the pre-processed data is sent as input for feature reduction. In this present work, a feature reduction technique, namely LSI, is used. Using this feature reduction technique, the attributes in the dataset get reduced. From thirty four attributes, only twelve attributes are taken for further processing. The data obtained after performing feature reduction using LSI is given in Figure 3.

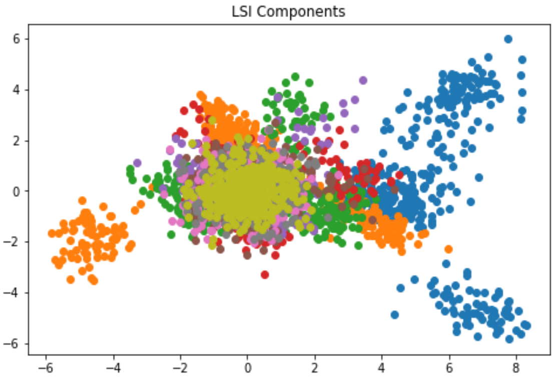

Figure 3. Dimension Reduction Using LSI Technique

From the above plot, it is viewed that the data's obtained after carrying out dimensionality reduction is scattered in both x-axes and Y-axes. Each colour represents the number of instances present in every attribute. From a total of thirty-four attributes, only twelve attributes were selected. These twelve attributes are quite significant, and it is essential for prediction. The remaining attributes in the dataset do not have a major impact on the classification process. The selected twelve attributes are sent for the feature selection process. In this present work, the feature selection process is done using filter-based techniques, namely correlation related feature selection. In this correlation-based feature selection to optimize the number of feature subsets, a Metaheuristic algorithm named beetle swarm optimization algorithm is utilized. The final output obtained as a result of the feature selection process is extracted features. These extracted features are finally sent for the classification process. The feature selection is a significant process performed prior to classification, and it must be carried out effectively in order to improve the classification performance and to reduce the error rate.

4.1. Comparison Analysis

The performance of the proposed feature selection process is compared with some of the existing features selection approaches. The proposed feature selection algorithm is designed by merging LSI and CFS to achieve better extraction of features. Some of the existing features selection techniques used for comparison study are Principal Component Analysis (PCA), Independent Component Analysis (ICA), T- Distributed Stochastic Neighbour Embedding (TSNE) and Isometric mapping (ISO). The performance attained using these existing techniques is compared with the proposed feature selection algorithm. Some of the performance metrics used for comparison are accuracy, specificity, recall, F-1 score, precision, False Positive Rate (FPR), False Negative Rate (FNR), Negative Predictive Value (NPV), Error and Mathew Correlation Coefficient (MCC). Table 2 illustrates the comparison study done using the proposed and existing feature selection algorithm.

Table 2. Comparison Investigation between Proposed and Existing Feature Selection Algorithm

Performance Metrics	PCA-CFS	TSNE-CFS	ICA-CFS	ISO-CFS	LSI-CFS
Accuracy	0.9	0.87	0.88	0.73	0.95
Specificity	0.8	0.74	0.68	0.17	0.85
recall	0.88	0.79	0.81	0.80	0.92
F-1 score	0.92	0.90	0.92	0.83	0.94
Precision	0.89	0.88	0.86	0.71	0.94
False Positive Rate (FPR)	0.3	0.25	0.31	0.82	0.14
False Negative Rate (FNR)	0.12	0.16	0.2	0.24	0.05
Negative Predictive Value (NPV)	0.9	0.72	0.76	0.72	0.92
Error	0.09	0.12	0.11	0.26	0.05
Mathew Correlation Coefficient (MCC).	0.75	0.69	0.74	0.35	0.85

The above table displays the value obtained for various metrics using the proposed and existing feature selection algorithm. The value for a parameter such as error, false-negative rate, false-positive rate and negative predictive value is smaller for the proposed feature selection algorithm in contrast with existing approaches. The rest of the metrics, such as accuracy, specificity, recall, F1 score, precision and MCC, is greater for the proposed feature selection algorithm in comparison to conventional techniques. The graphical representation for this comparison investigation is given below.

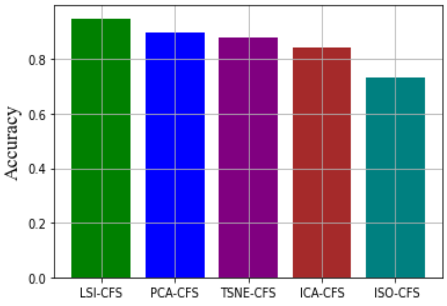

Figure 4. Comparison of Accuracy Metrics

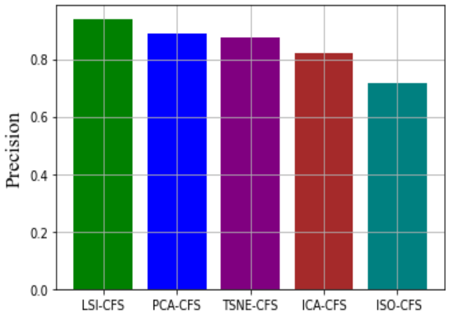

Figure 5. Comparison of Precision Metrics

The comparison investigation done using the accuracy metric is given in figure 4. The graph is plotted against different feature selection algorithm on X-label and obtained accuracy value on Y-label respectively. The accuracy value for the proposed feature selection algorithm LSI-CFS is 95%, and it is seen to be greater in comparison to conventional feature selection techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values for accuracy are 90%, 88%, 87% and 73%. This shows better performance of the proposed feature selection algorithm in comparison with the traditional algorithm. Figure 5 illustrates the comparison study carried out using precision metrics. In this analysis, too, the graph is drawn using various feature selection algorithm on X- coordinate and value for precision on Y-coordinate. The acquired precision value for the proposed feature selection technique is 94%, and it is found to be greater in contrast with an existing technique like PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values are 89%, 88%, 86% and 71%.

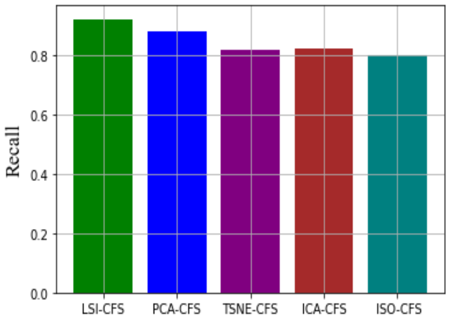

Figure 6. Comparison of Recall Metric

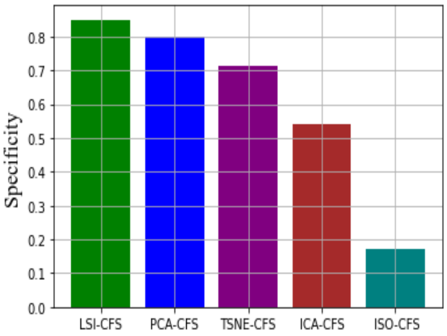

Figure 7. Comparison of Specificity Metric

The comparison investigation done using the recall metric is given in figure 6. The graph is plotted against different feature selection algorithm on X-label and obtained sensitivity or recall value on Y-label, respectively. The recall value for the proposed feature selection algorithm LSI-CFS is 92%, and it is seen to be greater in comparison to conventional feature selection techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values for accuracy are 88%, 79%, 81% and 80%. Figure 7 illustrates the comparison study carried out using specificity metric. In this analysis, too, the graph is drawn using various feature selection algorithm on X- coordinate and value for specificity on Y-coordinate. The acquired specificity value for the proposed feature selection technique is 85%, and it is found to be greater in contrast with an existing technique like PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values are 80%, 74%, 68% and 17%. This shows better performance of the proposed feature selection algorithm in comparison with the traditional algorithm.

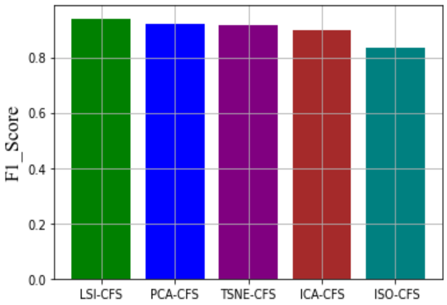

Figure 8. Comparison of F1 Score Parameter

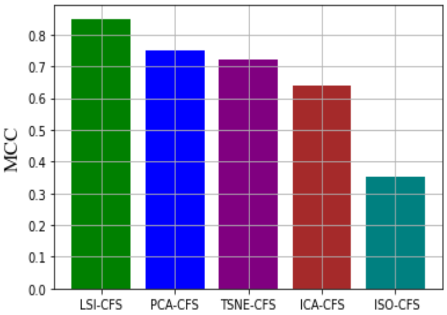

Figure 9. Comparison of MCC metric

The comparison investigation done using the F1 score parameter is given in figure 8. The graph is plotted against different feature selection algorithm on X-label and obtained F1 score value on Y-label respectively. The value of the F1 score for the proposed feature selection algorithm LSI-CFS is 94%, and it is seen to be greater in comparison to conventional feature selection techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values for accuracy are 92%, 90%, 92% and 83%. Figure 9 illustrates the comparison study carried out using the MCC metric. In this analysis, too, the graph is drawn using various feature selection algorithm on X- coordinate and value for MCC on Y-coordinate. The acquired MCC value for the proposed feature selection technique is 85%, and it is found to be greater in contrast with an existing technique like PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values are 75%, 69%, 74% and 35%. The evaluation of the F1 score and MCC metric revealed that the proposed feature selection algorithm functions effectively and improves the classification performance.

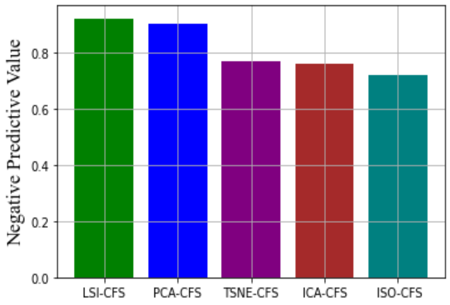

Figure 10. Comparison of Negative Predictive Value

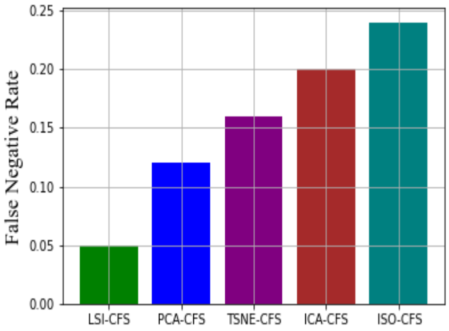

Figure 11. Comparison of False Negative Rate

The comparison analysis done based on negative predictive value is given in figure 10. In this figure also the graph is drawn between several feature selection algorithm on X-axis and negative predictive values on Y-axis. The negative predictive value attained for the proposed feature selection algorithm LSI-CFS is 92%, and it is found to be greater in comparison to traditional feature selection techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose corresponding values are 90%, 72%, 76% and 72% respectively. Figure 11 illustrates the comparison study done based on the false-negative rate. In the case of figure 11, the graph is drawn between several feature selection techniques on X-axis and false-negative rates on Y-axis. The false-negative rate attained for the proposed feature selection algorithm is 0.05%, and it is viewed to be lesser in comparison with existing techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values for false-negative rate is 0.12%, 0.16%, 0.2% and 0.24%.

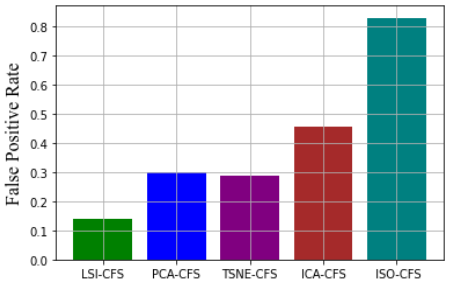

Figure 12. Comparison of False Positive Rate

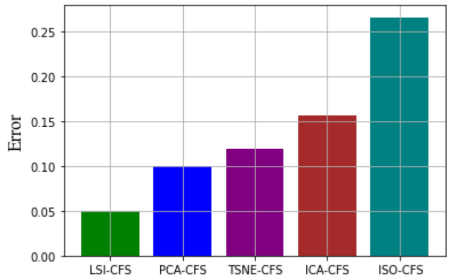

Figure 13. Comparison of Error

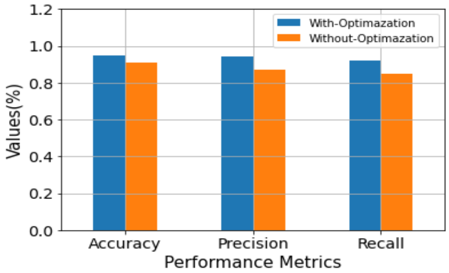

Figure 14. Comparison of Proposed Technique with and without Optimization

The comparison analysis done based on a false positive rate is given in figure 12. In this figure also the graph is drawn between several feature selection algorithm on X-axis and false-positive rates on Y-axis. The false-positive rate attained for the proposed feature selection algorithm LSI-CFS is 0.14%, and it is found to be lesser in comparison to traditional feature selection techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose corresponding values are 0.3%, 0.25%, 0.31% and 0.82% respectively. Figure 13 illustrates the comparison study done based on error metrics. In the case of figure 13, the graph is drawn between several feature selection techniques on X-axis and value for error on Y-axis. The value of error obtained for the proposed feature selection algorithm is 0.05%, and it is viewed to be lesser in comparison with existing techniques such as PCA-CFS, TNSE-CFA, ICA-CFS and ISO-CFS, whose values for false-negative rate is 0.9%, 0.12%, 0.11% and 0.26%. Following that, in figure 14, the performance of the proposed algorithm is analyzed in the presence and absence of the optimization technique. In this figure, the graph is plotted between various performance metrics and their corresponding values in both X and Y labels, respectively. From the graph, it is inferred that the value for performance metrics such as accuracy, sensitivity and precision in the presence of optimization is 0.95%, 0.92% and 0.92, respectively. On the other hand, in the absence of optimization the value for performance metrics such as accuracy, sensitivity and precision is 0.89%, 0.82% and 0.85%, respectively. This revealed that the performance of the proposed feature selection algorithm gets enhanced in the presence of the optimization technique.

Table 3. Comparison Analysis

Performance Metrics	CFS [19]	CFS-PSO [20]	PCA-IG [21]	PCA-SFB[22]	LSI-CFS
Accuracy	0.87	0.89	0.90	0.93	0.95
Specificity	0.73	0.77	0.80	0.83	0.85
recall	0.81	0.85	0.87	0.90	0.92
F-1 score	0.81	0.84	0.88	0.91	0.94
Precision	0.80	0.84	0.88	0.91	0.94
False Positive Rate (FPR)	0.21	0.19	0.18	0.15	0.14
False Negative Rate (FNR)	0.32	0.21	0.12	0.08	0.05
Negative Predictive Value (NPV)	0.115	0.102	0.98	0.95	0.92
Error	0.32	0.21	0.12	0.08	0.05

Table 3 illustrates the comparison study performed between proposed method and some of the existing techniques mentioned in the literature review section. 95% accuracy was attained by the proposed method. Whereas 93%, 90%, 89% and 87% accuracy was attained by the existing technique such as CFS [19], CFS-PSO [20], PCA-IG [21] and PCA-SFB [22]. This greater accuracy attained by the proposed method proves its effective functioning over existing technique. From the entire analysis, it is understood that the functioning of the proposed feature selection algorithm is superior in comparison with the existing technique.

5. Conclusion

This paper aims at designing the best feature selection algorithm in order to achieve accurate prediction of skin cancer disease with a lesser error rate. Over the past years, the number of people suffering from skin cancer is increasing due to ozone layer depletion. So, early stages detection of skin cancer is mandatory in order to increase the survival rate of skin cancer patients. For that purpose, the present experts are focusing on designing automated techniques to achieve effective prediction of skin cancer. Considering this idea, the data mining technique is introduced. The data mining approach relies on a machine learning algorithm to classify the patient suffering from skin cancer. So, to improve the performance of the machine learning algorithm, a better feature selection algorithm must be designed. So, in this research hybrid feature selection algorithm is designed using LSI and CFS techniques. The dimensionality reduction in the dataset is achieved using LSI, and the selection of the best features is done using CFS. Within CFS, to optimize the value of k, beetle swarm optimization is utilized. Further, the performance of the proposed feature selection algorithm is validated and compared with the existing feature selection technique. The comparison analysis revealed that excellent feature selection is achieved using the proposed feature selection algorithm.

Acknowledgements.

Funding

There is no funding provided to prepare the manuscript.

Conflict of Interest

There is no conflict of Interest between the authors regarding the manuscript preparation and submission.

Data availability statement

If all data, models, and code generated or used during the study appear in the submitted article and no data needs to be specifically requested.

Code availability

No code is available for this manuscript.

References

[1] Li Y, Li T, Liu H. Recent advances in feature selection and its applications. Knowledge and Information Systems, 2017, 53(3):551-77.

[2] Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. Journal of biomedical informatics, 2018, 85:189-203.

[3] Chen K, Zhou FY, Yuan XF. Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. Expert Systems with Applications, 2019, 128:140-56.

[4] Selvakumar B, Muneeswaran K. Firefly algorithm based feature selection for network intrusion detection. Computers & Security, 2019, 81:148-55.

[5] Rao H, Shi X, Rodrigue AK, Feng J, Xia Y, Elhoseny M, Yuan X, Gu L. Feature selection based on artificial bee colony and gradient boosting decision tree. Applied Soft Computing, 2019, 74:634-42.

[6] Brezočnik L, Fister I, Podgorelec V. Swarm intelligence algorithms for feature selection: a review. Applied Sciences, 2018, 8(9):1521.

[7] Bayati H, Dowlatshahi MB, Paniri M. MLPSO: a filter multi-label feature selection based on particle swarm optimization. In 2020 25th International Computer Conference, Jan 1 Computer Society of Iran (CSICC) IEEE 2020. pp. 1-6.

[8] Li M, Wang H, Yang L, Liang Y, Shang Z, Wan H. Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. Expert Systems with Applications, 2020, 150:113277.

[9] Gokulnath CB, Shantharajah SP. An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 2019, 22(6):14777-87.

[10] Abdel-Basset M, El-Shahat D, El-henawy I, de Albuquerque VH, Mirjalili S. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Systems with Applications, 2020, 139:112824.

[11] Maldonado S, Bravo C, López J, Pérez J. Integrated framework for profit-based feature selection and SVM classification in credit scoring. Decision Support Systems, 2017, 104:113-21.

[12] Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. Biophysical reviews, 2019, 11(1):31-9.

[13] Jadhav S, He H, Jenkins K. Information gain directed genetic algorithm wrapper feature selection for credit rating. Applied Soft Computing, 2018, 69:541-53.

[14] Moslehi F, Haeri A. An evolutionary computation-based approach for feature selection. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(9):3757-69.

[15] Pashaei E, Aydin N. Binary black hole algorithm for feature selection and classification on biological data. Applied Soft Computing, 2017, 56:94-106.

[16] Zhou, H., Zhang, J., Zhou, Y., Guo, X., & Ma, Y. (2021). A feature selection algorithm of decision tree based on feature weight. Expert Systems with Applications, 164, 113842.

[17] Hosseini, S., & Seilani, H. (2021). Anomaly process detection using negative selection algorithm and classification techniques. Evolving Systems, 12(3), 769-778.

[18] Liu, S., Wang, H., Peng, W., & Yao, W. (2022). A surrogate-assisted evolutionary feature selection algorithm with parallel random grouping for high-dimensional classification. IEEE Transactions on Evolutionary Computation.

[19] Abinash MJ, Vasudevan V. A study on wrapper-based feature selection algorithm for leukemia dataset. InIntelligent Engineering Informatics 2018 (pp. 311-321), Springer, Singapore.

[20] Abualigah L, Dulaimi AJ. A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm. Cluster Computing, 2021, 24(3):2161-76.

[21] Lyu H, Wan M, Han J, Liu R, Wang C. A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data mining. Computers in biology and medicine, 2017, 89:264-74.

[22] Chormunge S, Jena S. Correlation based feature selection with clustering for high dimensional data. Journal of Electrical Systems and Information Technology, 2018, 5(3):542-9.

[23] Jain I, Jain VK, Jain R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing, 2018, 62:203-15.

[24] Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. Expert Systems with Applications, 174, 114765.

[25] Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021, March). Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 141-146). IEEE.

[26] Tawhid, M., Ahad, N., Siuly, S., Wang, K., & Wang, H. (2021, October). Data Mining Based Artificial Intelligent Technique for Identifying Abnormalities from Brain Signal Data. In International Conference on Web Information Systems Engineering (pp. 198-206). Springer, Cham.

[27] Jenghara MM, Ebrahimpour-Komleh H, Rezaie V, Nejatian S, Parvin H, Yusof SK. Imputing missing value through ensemble concept based on statistical measures. Knowledge and Information Systems. 2018, 56(1):123-39.

[28] Jain S, Shukla S, Wadhvani R. Dynamic selection of normalization techniques using data complexity measures. Expert Systems with Applications. 2018, 106:252-62.

[29] Adinugroho S, Sari YA, Fauzi MA, Adikara PP. Optimizing K-means text document clustering using latent semantic indexing and pillar algorithm. In2017 5th international symposium on computational and business intelligence (ISCBI) 2017 Aug 11 (pp. 81-85), IEEE.

[30] Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 2019 Jun 25.

[31] Ma J, Gao X. A filter-based feature construction and feature selection approach for classification using Genetic Programming. Knowledge-Based Systems, 2020, 196:105806.

[32] Abinash MJ, Vasudevan V. A study on wrapper-based feature selection algorithm for leukemia dataset. InIntelligent Engineering Informatics 2018 (pp. 311-321), Springer, Singapore.

[33] Albashish D, Hammouri AI, Braik M, Atwan J, Sahran S. Binary biogeography-based optimization based SVM-RFE for feature selection. Applied Soft Computing, 2021, 101:107026.

[34] Wang T, Yang L. Beetle swarm optimization algorithm: Theory and application. arXiv preprint arXiv:1808.00206, 2018 Aug 1.

[35] Wang L, Wu Q, Lin F, Li S, Chen D. A new trajectory-planning beetle swarm optimization algorithm for trajectory planning of robot manipulators. IEEE access, 2019, 7:154331-45.

[36] Cunha CF, Carvalho AT, Petraglia MR, Amorim HP, Lima AC. Proposal of a novel fitness function for evaluation of wavelet shrinkage parameters on partial discharge denoising. IET Science, Measurement & Technology, 2018, 12(2):283-9.

[37] Xue JH, Hall P. Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis?. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(5):1109-12.
[38]Dataset Link:
https://archive.ics.uci.edu/ml/datasets/dermatology