STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, August 2025, pp 808–850.
Published online in International Academic Press (www.IAPress.org)

An Efficient Machine Learning Framework for Disease Gene Prediction in Parkinson's Disease and Bladder Cancer

Noura Mohammed A. Abdelwahed [1,*], Gh.S. El-Tawel[2], M. A. Makhlouf [1]

[1] *Information Systems Department,Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt*
[2] *Computer Science Department,Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt*

Abstract Machine learning (ML) has been increasingly used in disease prediction, leveraging both phenotype and genotype data. However, genotype data have received comparatively less attention due to limited availability, whereas phenotype data have been more extensively studied. While breast cancer research is abundant, studies on other cancers, such as bladder cancer, and neurological diseases like Parkinson's disease, remain limited. High-dimensional datasets pose challenges, including lengthy processing times, overfitting, an excess of features, and difficulties in classification. This study introduces a framework that integrates phenotype and genotype data for cancer prediction, aiming for high accuracy with a minimal number of relevant features. The framework consists of three main procedures: feature selection (FS), cancer prediction (CP), and identification of cancer-associated genes/features (CAG/F). FS employs a hybrid LEDF approach, combining the empirical distribution function (EDF) with three embedded methods: lasso regression selection (LRS), ridge regression selection (RRS), and random forest selection (RFS). EDF acts as a resampling tool with external (EEDF) and internal (IEDF) components that merge as E/IEDF. Features are selected based on classification accuracy using both union and intersection methods. CP applies multiple ML models with cross-validation to enhance prediction accuracy. Lastly, CAG/F identifies cancer-associated genes/features following the FS and CP steps. The algorithms E/IEDF-RFS, E/IEDF-LRS, and E/IEDF-RRS demonstrated excellent performance for RNA gene and dermatology datasets, achieving 100% accuracy. E/IEDF-RFS reached 94.58% accuracy for Parkinson's Disease2, while EEDF-LRS performed best for DNA data with 94.85% accuracy. E/IEDF-RRS showed 96.43% accuracy for Parkinson's Disease1 using RF classifiers, and IED-RFS and E/IEDF-LRS achieved 98.42% accuracy for the BreastEw dataset.

Keywords Human Cancers and Genes; Machine Learning; Gene and Feature Selection; Embedded Methods; Overfitting and High Dimensional Dataset; Classification Algorithms

DOI: 10.19139/soic-2310-5070-2517.

1. Introduction

In the realm of ML and artificial intelligence (AI), the analysis of high-dimensional datasets is considered a crucial step. AI has revolutionized feature selection (FS) limitations resolution, establishing a significant link between computer science and data, particularly in the healthcare field. Its purpose is to emulate human decision making [1, 2, 3]. Since the rapid advancements in computer science, the abundance of healthcare datasets, and the development of arithmetic algorithms, AI applications have been extensively employed in this field since 2000 [1, 2, 3]. Moreover, AI has significantly improved and addressed numerous issues pertaining to human cancer diseases. It serves as a valuable tool for specialists, providing a second opinion to assist in their final decision-making due to its effectiveness and robustness [2]. AI possesses remarkable capabilities in acquiring more precise information compared to manual methods. This accurate information supports specialists in making informed decisions [1]. Furthermore, AI applications require less labor than manual methods, reducing patient

*Correspondence to: Noura Mohammed A. Abdelwahed (Email: malekmalek20131988@gmail.com). Information Systems Department,Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

ISSN 2310-5070 (online) ISSN 2311-004X (print)

burden and saving time and costs [1, 2, 3]. Cancer has the ability to spread very quickly in recent times causes many diseases—not only common illnesses but also cancer—leading to death [1, 2, 3, 4]. Cancer can be defined as the uncontrolled growth of abnormal cells or changes in gene sequences due to various factors, and it can develop in any part of the body[1, 5, 6]. There are many types of cancer, but in our work, we focused on specific types.

Parkinson's disease (PD) is a prevalent neurodegenerative disorder [7]. Detecting and diagnosing it in the early stages is particularly challenging [8]. PD is recognized as the second most common neurodegenerative cause of death after Alzheimer's disease [8, 9]. A primary factor contributing to PD is a decrease in dopamine production, a critical chemical produced by neurons in the brain [10]. Dopamine facilitates key brain functions, and its deficiency leads to the onset of PD [10]. While insufficient dopamine production is a known cause, the precise reasons for the disease remain unclear. Early diagnosis is crucial for preventing severe complications, yet it remains difficult due to several factors: (1) a shortage of specialists who can accurately diagnose PD in many regions, (2) the significant workload and pressure on doctors [10], (3) the influence of genetic and environmental factors, (4) age-related risks and stress [11]. PD affects both men and women, although men are more likely to develop it [12, 13]. The disease's symptoms include impaired movement and walking, which worsen progressively over time [12]. To improve early detection, it is critical to identify the key genes associated with PD. While machine learning (ML) techniques have been applied to address this, relatively few studies focus on PD-specific methods [10].

Bladder cancer (BLC) is a rapidly spreading disease and ranks as the fifth most common cancer globally [14, 15] as well as the ninth most frequent malignant tumor [16]. It is classified into two types: (1) Muscle-invasive bladder cancer (MIBC), (2) Non-muscle invasive bladder cancer (NMIBC) [16, 17]. BLC usually starts in the bladder's inner lining and, in some cases, spreads to the surrounding muscle tissue. Once the tumor reaches this stage, it may metastasize to other body parts through the lymphatic system [18]. Smoking is one of the most prominent risk factors for BLC, with its impact depending on the duration and intensity of tobacco use [14, 19, 20]. Men are at a higher risk of developing BLC compared to women [20, 21]. Genetic factors also significantly contribute to the likelihood of BLC [22]. Moreover, dietary habits are crucial, as unhealthy diets are linked to an increased risk of the disease [20, 23]. Repeated bladder inflammation, urinary obstructions, or catheter-related injuries can further exacerbate the condition [20, 24]. Although medicinal treatments are widely used, they often come with serious side effects [20, 25]. Obesity has also been identified as a notable risk factor for BLC [26]. To reduce the risk of BLC, the following guidelines are recommended: (1) avoid smoking and limit exposure to tobacco smoke [20, 27], (2) reduce exposure to harmful chemicals [20], (3) maintain a diet rich in fruits and vegetables, (4) incorporate regular exercise into your routine [20, 28]. Individuals diagnosed with BLC should prioritize regular health monitoring and undergo necessary medical examinations to manage their condition effectively[20, 29, 30].

Big data presents numerous challenges, prompting the use of machine learning (ML) to address these complexities without explicit programming [1, 31]. Some of the most pressing issues include selecting relevant features, reducing fitting time, improving classification accuracy, and ensuring robust model validation [1, 31]. Feature selection (FS), the process of identifying and retaining the most informative features while discarding irrelevant ones, is essential to solve these challenges [1, 31, 32, 33, 34, 35]. FS techniques are broadly categorized into three types: Filter methods: These evaluate features based on statistical scores and select subsets accordingly. Filters are popular for their simplicity and speed but face limitations such as overfitting, lack of ML integration, and occasional failure to identify the most relevant features [1, 35, 36]. Wrapper methods use ML algorithms to evaluate and select features, often using predictive models for performance optimization. Although effective, wrappers are computationally intensive and susceptible to overfitting [1, 35]. Embedded methods incorporate feature selection into the model training process, often during the classification phase. Embedded techniques are preferred for their ability to reduce overfitting, optimize computational costs, and identify the most significant features [37]. In this study, we adopted the embedded FS method to address these challenges. This approach offers a comprehensive solution, enhancing classification accuracy, reducing data dimensionality, lowering processing time, and mitigating over-fitting [1, 37]. Furthermore, it significantly improves cancer prediction (CP) performance, making it an optimal choice for big data applications.

Recently, machine learning (ML) models have been widely applied across various fields to detect and diagnose diseases [1, 38]. These models have demonstrated superior predictive performance compared to manual methods [39]. However, big data poses significant challenges, such as high dimensionality, which can lead to inaccurate

results and wasted computational resources. To address these issues, many researchers have focused on leveraging ML algorithms for disease detection and diagnosis, with particular emphasis on Parkinson's disease (PD) and bladder cancer (BLC). Filter models are commonly used in feature selection (FS) due to their simplicity and efficiency in saving time. However, they often face issues such as over-fitting. For instance in [40], the authors proposed a filtering algorithm combining the Information Gain (IG) algorithm with a distance metric known as IGD to identify key features. Their process involved five steps and employed three classifiers: K-nearest neighbors (KNN), neural networks, and Naive Bayes. Similarly, in [34], the authors developed an IG-based feature selection (IGF) approach to analyze various types of human cancer using a DNA copy number variation (CNV) dataset. This method successfully identified 16,381 features out of 23,000, followed by classification using multiple algorithms. Another study [41], explored several filtering algorithms, including correlation-based feature selection (CBS), fast correlation-based feature selection (FCBF), and mutual information for feature selection (MIFS). The classification was carried out using the KNN classifier. In [42], the researchers implemented multiple filter-based FS algorithms, such as IGF, principal component analysis (PCA), and CBS. These selected features were classified using multilayer perceptron, decision trees, random forest (RF), and KNN classifiers.

In [43], the Chi-Square test was used to identify key features associated with malaria Likewise. In [44], a filter model based on IGF was proposed, evaluated using J48, support vector machine (SVM), and Bayes Net classifiers. The study used 32 real-world datasets in [45], and eight FS filter algorithms were explored, including the Gini index, ReliefF, spectral feature selection (SPEC), conditional mutual information maximization (CMIM), minimum redundancy maximum relevance (mRMR), joint mutual information (JMI), efficient and robust feature selection (RFS), and CBS. Four classifiers were applied to the FS output. Additionally, in [46], the authors introduced a Bat algorithm-based RF (BbRF) approach to enhance disease classification performance, incorporating fuzzy values to improve accuracy. Another method proposed in [47] used an extensive feature selector (EFS) across four datasets and compared it with nine established FS techniques. Finally, in [48] a filter method. An approach integrating IGF, Chi-Square, and inter-correlation algorithms was proposed. This approach addressed imbalanced classes, implemented feature ranking, and analyzed feature-to-feature correlations.

Wrapper methods enhance model performance and generate a meaningful subset of features. For these reasons, many studies have applied them to the feature selection (FS) process. Despite their advantages, wrapper methods face several challenges, with one of the biggest being computational cost. In [49], the authors proposed a multilayer feature subset selection method (MLFSSM), where features are divided into subsets with equal weights. Numerous feature subsets are generated to obtain diverse feature combinations, with each subset having its own classifier. The final results are based on the highest subset accuracy from the last layer. In [50], the authors modified the genetic algorithm (GA) to create GA-based Feature Selection (GbFS), which was applied to develop firewalls and intrusion detection systems (IDSs), using three benchmark network traffic datasets. In [34], the authors introduced a wrapper method that combines a genetic algorithm (GA) and particle swarm optimization (PSO) for the feature selection process, employing various classifiers to evaluate model performance. To address the challenges posed by wrapper methods, many authors have explored embedded methods. In [51], the Salp Swarm Algorithm (SSA) was used to improve classification accuracy and convergence speed, with an inertia weight incorporated to refine the final outcomes. On the other hand, the authors in [35] introduced a modified algorithm for feature selection that incorporates chaotic maps to enhance the performance of the SSA. This approach was tested on 27 different datasets, with the selected features subsequently used in a KNN classifier.

Embedded algorithms are employed to address and enhance the challenges associated with wrapper methods. In this context, the author in [52] proposed a modified Random Forest algorithm, named xRF, which eliminates less important features based on p-values and identifies a subset of unbiased features. In [62], the authors improved the performance of the Random Forest (RF) algorithm by applying it to gastric cancer patient data from the Surveillance, Epidemiology, and End Results (SEER) database. To improve RF voting, they implemented out-of-bag (OOB) evaluation to assess decision tree performance, also introducing a three-level weighted random forest (TLWRF) that replaces OOB with training data. On the other hand, many authors have applied LRS algorithms in the feature selection process to choose relevant features for model construction. In [54], the authors applied an LRS based feature selection model to a crime dataset, using the caret package for preprocessing. The selected features were then fed into Naive Bayes classifiers, with Autoregressive Integrated Moving Average (ARIMA)

used for forecasting. In [55], the authors introduced a feature selection process that combines LRS with elastic net regularized generalized linear models (glmnet in R) and mRMR, applied to public health nursing documentation. Additionally, in [56], a multi-level feature selection algorithm based on LRS coefficient threshold (Coe-Thr-Lasso) was proposed. This method removes features with low correlation to classification results using t-tests and variance, and eliminates redundant features with a low coefficient threshold. The proposed method was compared to other classifiers, such as RF, LR, and SVM. To improve survival prediction, the authors in [57] applied LRS for selecting important features from microarray datasets in gynecologic cancer research, using 10-fold cross-validation and calculating the area under the ROC curve to validate accuracy. In [58], the authors used a dataset for air quality prediction and proposed a selection method that compared LRS with Correlation-based Adaptive LASSO (CbAL), which enhances LRS by evaluating adaptive weights. The selected features were forwarded to various classifiers. To optimize diabetes diagnosis, the authors in [59] applied LRS, citing its ability to (1) select the most relevant features, (2) improve classification accuracy, (3) minimize over-fitting, and (4) maximize model interpretability [59, 60]. They used Analysis of Variance (ANOVA) for their analysis. Several authors have applied Regularized Regression Selection (RRS) algorithms due to their advantages. In [61], a method using RRS was proposed for both synthetic and real-world datasets. The method was compared with several algorithms, including information gain, the single-set spectral sparsification algorithm with leverage-score sampling, random feature selection, and rank-revealing QR factorization (RRQR). In [62], the RRS algorithm was applied in the medical field using a genotype dataset called GTEx RNAseq expression, based on Bayesian methods (B-GEX). Correlation coefficients between target genes and preselected feature genes in peripheral blood were captured; with feature reduction performed using the cosine similarity approach, and linear regression as the baseline method. Furthermore, in [63], the authors used the RRS algorithm to diagnose diabetes at an early stage, integrating it with the Ridge-Adaline Stochastic Gradient Descent (RASGD) classifier. High correlation features were selected, and Adaline was added to the Stochastic Gradient Descent method to enhance the classification model.

Hybrid algorithms have become central to the prediction process. Numerous authors have implemented various hybrid algorithms for feature selection (FS). For instance, in [64], the authors combined filter and wrapper methods to create hybrid features for five cancer microarray datasets. They utilized a gain ratio (GR) filter, with the selected features passed to a forward selection algorithm and then evaluated using several classifiers. In [65], a hybrid method was proposed that combines IGF with best-first search, rank search, and greedy stepwise approaches for cancer datasets. The selected features were classified using KNN, Naïve Bayes, RF, SVM, and stacking ensemble methods. Similarly, the authors in [66] developed a hybrid method incorporating autoregressive (AR) models and empirical mode decomposition (EMD). The selected features were processed using CBS methods and RF, with RMSE calculated to measure performance. In [67], a hybrid FS approach was employed for survival prediction in hepatocellular carcinoma datasets. This method combined wrapper and embedded algorithms, using LRS and RRS based on the LR classifier for embedding and RFE with gradient boosting and RF. Various classifiers were then applied. To address dimensionality reduction and class imbalance, the authors in [68] developed a hybrid FS method combining LRS with random oversampling, applied to a diabetes dataset, with the features fed to an ANN classifier. On another note, the authors in [69] proposed a hybrid FS method for economic datasets by combining correlation analysis ANN, RRS, LRS and Elastic-net. The selected features were evaluated using different classifiers. In [70], a two-stage hybrid FS method was introduced. The first stage combined GA with IGF, while the second stage used mRMR algorithms. The features selected in both stages were forwarded to various classifiers. The authors in [71] developed a hybrid method incorporating RLS and RRS for heart disease datasets, using multiple classifiers to evaluate performance. Similarly, in [72], a hybrid method using LRS and RRS was proposed for diabetes datasets, with RMSE and median RMSE calculated to assess accuracy. In [73], a comprehensive hybrid FS approach was applied to a heart disease dataset, combining methods such as ANOVA, chi-square, MIFS, relief, forward feature selection, backward feature elimination, RFE, exhaustive feature selection, RLS, and RRS. The selected features were tested across various classifiers.

Currently, the detection and diagnosis of various cancer types leverage genotype and phenotype datasets, which form the primary motivation for our work, particularly for Parkinson's disease (PD) and bladder cancer (BLC). To validate the proposed strategy, we employ a hybrid feature selection (FS) approach that integrates LEDF, RFS, LRS, and RRS algorithms. The EDF equation is incorporated for several reasons, most notably its ability to minimize

over-fitting among features. It is applied at three stages alongside embedded algorithms, ultimately identifying the most relevant features for cancer detection. The selected features are evaluated using multiple classifiers, including Bagg, SVM, RF, and LR. The performance of the LEDF method is compared against other standalone algorithms as well as methods from state-of-the-art research. This cancer prediction framework demonstrates exceptional results across six datasets.

This work contributes significantly in the following ways:

1. Development of a comprehensive framework for diagnosing various types of cancer using phenotype and genotype data across six datasets.
2. Proposal of a novel FS methodology combining the hybrid EDF equation with embedded algorithms, specifically LEDF-RFS, LEDF-RLS, and LEDF-RRS, to address existing FS limitations.
3. Highlighting the importance of FS in enhancing predictive accuracy for bladder cancer and Parkinson's disease.
4. Utilizing union and intersection operations to identify the most critical features or genes associated with the progression of human cancers.

The structure of the work is organized as follows:

- Introduction: This section highlights the challenges in feature selection and reviews prior research efforts to tackle these issues.
- Materials and Methods: This section details the hybrid algorithm proposed to improve feature selection and resolve the identified challenges.
- Results: This section provides the numerical findings of the proposed methods and compares them with outcomes from existing studies on the same datasets.
- Biological Interpretation of Key Features: This section provides n in-depth analysis of the key features identified by our models, explaining their biological significance and relevance to the studied conditions.
- Discussion: This section analyzes the implementation of the proposed methods and their potential practical applications.
- Conclusions: This section summarizes the key aspects of the proposed methods and evaluates their effectiveness in overcoming feature selection challenges.
- Limitations: This section outlines the key limitations and obstacles faced during the course of this research.

2. Materials and Methods

In this section, we introduce the proposed LEDF method, which integrates three embedded algorithms. The approach is applied in three configurations: external, internal, and a combination of both. Furthermore, we provide detailed descriptions of six distinct datasets related to various types of human cancer.

2.1. Datasets

In our study, we utilized diverse datasets, encompassing both phenotype and genotype data. Six distinct cancer datasets from various sources were employed. Five of these datasets were obtained from the UCI Machine Learning Repository [74], while the sixth was sourced from CBioPortal for Cancer Genomics [75, 76, 77]. A detailed description of these datasets is provided in Table 1.

Table 1. Summary of The Six Datasets Integrated in the Present Study

Category Type	DS No.	Datasets	#Features	#Samples	#Class
Small <100	D1	BreastEW	30	569	2
	D2	Dermatology	34	366	6
	D3	Parkinson's2	46	240	2
Medium <100 <D 1000	D4	Parkinson's1	753	756	2
Large <1000 <D 21000	D5	DDNA CNV	16381	2916	6
	D6	RNA gene	120531	801	5

2.2. A hybrid of LEDF and embedded algorithms for feature selection

In our work, we developed a hybrid approach by integrating the EDF equation, providing an effective solution to address feature selection (FS) challenges and reduce variance among features. This approach also mitigates the overfitting problem, which often leads to complex and unreliable outcomes. The embedded algorithms further enhance classification accuracy. By combining the strengths of the EDF equation with embedded algorithms, we identified the most influential features. These features delivered outstanding results compared to other methods, highlighting genes and chromosomes significantly involved in cancer mutations.

2.2.1. Empirical distribution function (EDF) In our work, we applied a resampling method using the bootstrap method. The samples are drawn with replacements the same size as the original datasets. Bootstrap used EDF for drawing the samples. The EDF is also called an empirical cumulative distribution function (ECDF). It is based on the empirical measures [78]. The samples of bootstrap from EDF are denoted as follows:

Given $D = D_1, D_2, D_3, ...D_{\mathrm{n}}$, where D is the original dataset.

$$F_n(d) = \frac{1}{\mathrm{n}} \sum_{I=1}^{n} (1(X_i < d)) \tag{1}$$

Where 1 is the indicator function. The statistic is computed. The bootstrap dataset is denoted as $D^* = D_1^*, D_2^*, D_3^*, D_4^*, \ldots\ldots D_n^*$. Bootstrap samples are drawn with the same size as the original datasets. The LEDF is applied for many locations with RF, LRS, and RRS. The EDF was applied with bootstrap resampling to improve ML stability and minimize variance and overfitting.

2.2.2. Lasso regression for selection (LRS) The main objective of our work is to obtain the relevant features (genes) for the selection process. Hence, we do our best to implement the appropriate algorithms for this main task. In this direction, we utilized the advantages of LRS algorithm for FS. This algorithm decides which features were selected or not. It selected features based on coefficients correlation. The features with zero coefficients are cancelled and the others are selected. Therefore, this algorithm diminishes over-fitting between features and provides classification accuracy. The LRS equation is described as follows:-

$$L_(\beta_0, \beta) = \sum_{i=1}^{s} (y_i - \beta_0 - x_i^n \beta)^2 + \lambda \sum_{k=1}^{p} \mid \beta_k \mid \tag{2}$$

Suppose the original data D with S samples (y_i, x_i), where (y_i) is the class of the variables and (x_i) is the samples. The β is the correlation coefficient for the LRS, where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \ldots\ldots \beta_p)$ which represents the regularization term. λ is the regularization parameter. LRS is used to reduce errors and over-fitting. This process is called regularization (L1).

2.2.3. Ridge regression for selection (RRS) The RRS algorithm selects the high correlated values for features. It belongs to regularization called L2, which evaluates the square of the magnitude of the coefficients. It provides an important way to deal with missing values. The RRS equation is described as follows:-

$$L_(\beta_0, \beta) = \sum_{i=1}^{s} (y_i - \beta_0 - x_i^n \beta)^2 + \lambda \sum_{k=1}^{p} \beta_k^2 \tag{3}$$

Where (y_i) is the class of the variables, and (x_i) are the samples. The β is the correlation coefficient for the RRS. λ is the regularization parameter L2 to reduce the complexity of the model performance. For LRS and RRS, the EDF is applied in different locations. The first location is applied before LRS and RRS called EEDF. The second location is applied after the fitting process for the two algorithms called IEDF. The EDF is applied in both external and internal locations in the third location, called E/IEDF. The results are improved after the proposed methods.

2.2.4. Random forest for selection (RFS) Due to many issues in FS process, RFS is applied to fix these issues. RFS provides an effective and efficient way for selection by computing the features' importance. It reduces the over-fitting and variance. It is an embedded method that solves the time complexity problem found in wrapper methods. The EDF is applied in three locations as in the previous two algorithms. During applying the LEDF, the importance of features is computed using Gini importance scores, and the optimal subset features are selected. The proposed methods flowchart is illustrated in fig.1.

The proposed algorithm 1 using three different locations for LRS algorithm called LEDF-LRS is given below. This algorithm included external, internal and both (external & internal) locations of EDF for resampling. The Algorithm 1 steps are explained as follows in Table 2.

Table 2. Algorithm 1: Hybrid Proposed Method using LEDF-LRS

Algorithm 1 of the hybrid proposed method using LEDF-LRS
Input:
Datasets D = (d1, d2, ..., dn) // phenotype and genotype cancer datasets
Output:
Selected feature sets: $F_{\text{EEDF\_LRS}}$, $F_{\text{IEDF\_LRS}}$, $F_{\text{E/IEDF\_LRS}}$
#Begin:
1. External Bootstrapping for: $F_{\text{EEDF\_LRS}}$ and $F_{\text{E/IEDF\_LRS}}$
For each dataset di in D:
Generate D* using equation (1) for EEDF and E/IEDF
2. Feature Selection with LRS:
Compute correlation coefficient using equation (2)
3. Internal Bootstrapping for: $F_{\text{IEDF\_LRS}}$ and $F_{\text{E/IEDF\_LRS}}$
For each dataset di in D:
Generate D* using equation (1) for IEDF and E/IEDF based on correlation results
4. Feature Selection Steps:
For each location in LRS:
Select features where \$correlation ≥ 0.1
5. Return:
$F_{\text{EEDF\_LRS}}$, $F_{\text{IEDF\_LRS}}$, $F_{\text{E/IEDF\_LRS}}$
6. Evaluation:
Evaluate models using accuracy, precision, recall, F1-score, AUC, Variance
7. Set Operations:
GIntersection LRS = Intersection$F_{\text{EEDF_LRS}}$, $F_{\text{IEDF\_LRS}}$, $F_{\text{E/IEDF\_LRS}}$
GUnion LRS = Union$F_{\text{EEDF_LRS}}$, $F_{\text{IEDF\_LRS}}$, $F_{\text{E/IEDF\_LRS}}$
End of Algorithm 1

The proposed algorithm 2. called LEDF-RRS are explained below as follows in Table 3:-

Table 3. Algorithm 1: Hybrid Proposed Method using LEDF-LRS

Algorithm R of the hybrid proposed method using LEDF-RRS
Input:
Datasets D = (d1, d2, ..., dn) // phenotype and genotype cancer datasets
Output:
Selected feature sets: $F_{\text{EEDF\_RRS}}$, $F_{\text{IEDF\_RRS}}$, $F_{\text{E/IEDF\_RRS}}$
#Begin:
1. External Bootstrapping for: $F_{\text{EEDF\_RRS}}$ and $F_{\text{E/IEDF\_RRS}}$
For each dataset di in D:
Generate D* using equation (1) for EEDF and E/IEDF
2. Feature Selection with RRS:
Compute correlation coefficient using equation (3)
3. Internal Bootstrapping for: $F_{\text{IEDF\_RRS}}$ and $F_{\text{E/IEDF\_RRS}}$
For each dataset di in D:
Generate D* using equation (1) for IEDF and E/IEDF based on correlation results
4. Feature Selection Steps:
For each location in RRS:
Select features where \$correlation ≥ 0.1
5. Return:
$F_{\text{EEDF\_RRS}}$, $F_{\text{IEDF\_RRS}}$, $F_{\text{E/IEDF\_RRS}}$
6. Evaluation:
Evaluate models using accuracy, precision, recall, F1-score, AUC, Variance
7. Set Operations:
GIntersection RRS = Intersection$F_{\text{EEDF_RRS}}$, $F_{\text{IEDF\_RRS}}$, $F_{\text{E/IEDF\_RRS}}$
GUnion RRS = Union$F_{\text{EEDF_RRS}}$, $F_{\text{IEDF\_RRS}}$, $F_{\text{E/IEDF\_RRS}}$
End of Algorithm 2

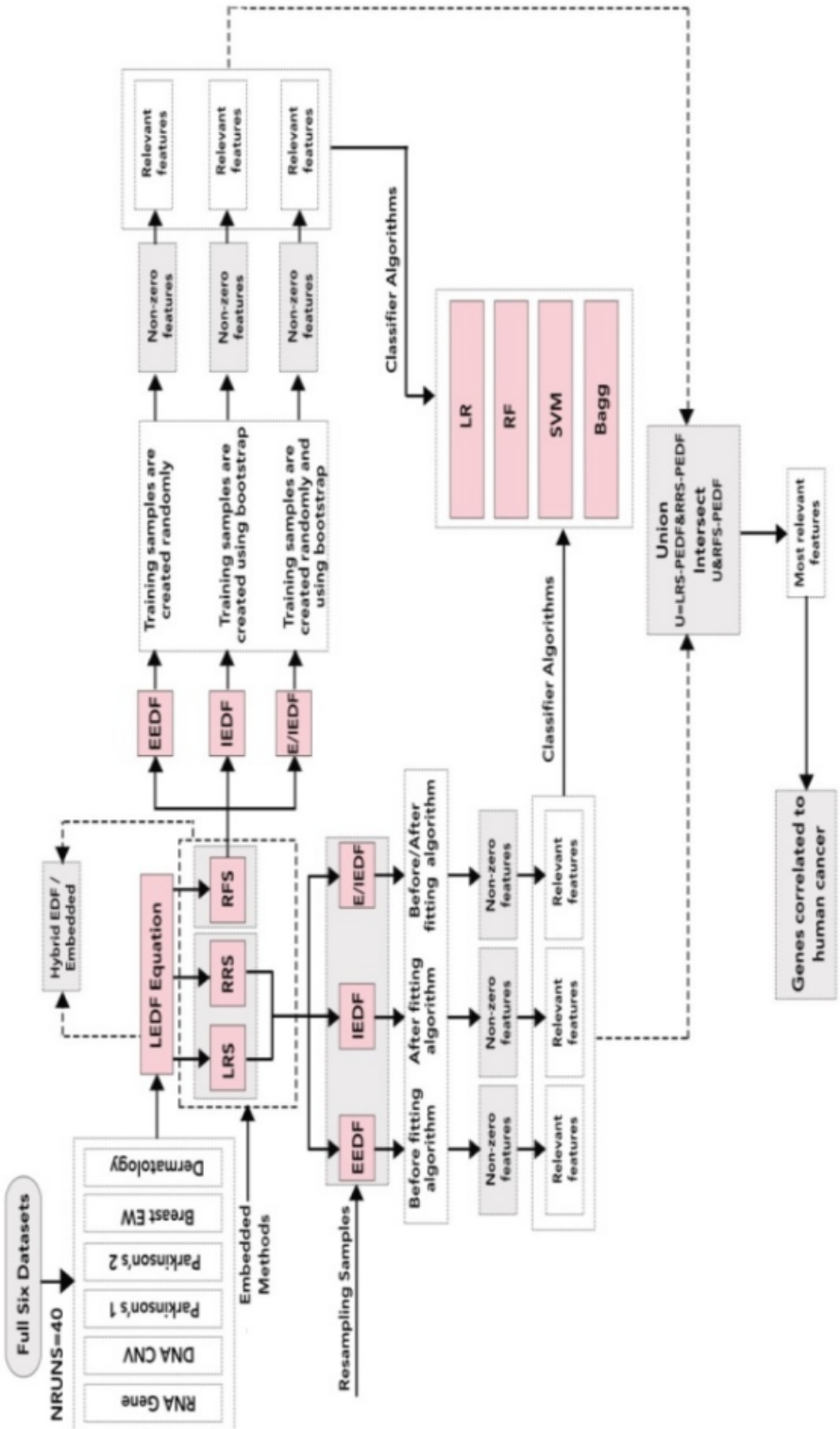

Figure 1. Diagram of the Proposed Methods

The proposed algorithm 3 using RFS algorithm with different bootstrap locations called LEDF-RFS is given as follows in Table 4:-

Table 4. Algorithm 3: Hybrid Proposed Method using LEDF-RFS

Algorithm 3 of the hybrid proposed method using LEDF-RFS
Input:
Datasets D = (d1, d2, ..., dn) // phenotype and genotype cancer datasets
Output:
Selected feature sets: $F_{\text{EEDF\_RFS}}$, $F_{\text{IEDF\_RFS}}$, $F_{\text{E/IEDF\_RFS}}$
#Begin:
1. External Bootstrapping for: $F_{\text{EEDF\_RFS}}$ and $F_{\text{E/IEDF\_RFS}}$
Generate D* using equation (1) for EEDF and E/IEDF
2. Feature Selection with RFS:
a. For $F_{\text{EEDF\_RFS}}$ and $F_{\text{IEDF\_RFS:}}$
Train using full dataset with M trees
b. For $F_{\text{EEDF\_RFS:}}$
Train using bootstrap samples with M trees
3. Repeat B times:
a. Build decision trees.
b. At each node:
- Select subset f $\subset$ F.
- Choose best feature from f.
- Rank features by importance.
- Remove features below threshold.
4. Return: $F_{\text{EEDF\_RFS}}$, $F_{\text{IEDF\_RFS}}$, $F_{\text{E/IEDF\_RFS}}$
5. Evaluation:
Assess with performance metrics
6. Set Operations: GIntersection RFS = Intersection$F_{\text{EEDF_RFS}}$, $F_{\text{IEDF\_RFS}}$, $F_{\text{E/IEDF\_RFS}}$
GUnion RFS = Union$F_{\text{EEDF_RFS}}$, $F_{\text{IEDF\_RFS}}$, $F_{\text{E/IEDF\_RFS}}$
End of Algorithm 3

3. Results

This section presents the extensive experiments conducted to validate our frame work and its algorithm. The experiments encompassed different dataset sizes to ensure the generalizability of our proposed method. We utilized six datasets comprising various human cancers and other diseases.

The proposed algorithms were designed to identify the most relevant features and genes from the datasets while eliminating those that would yield poor results if selected. The aim was to minimize the impact of high dimensionality, reduce processing time, prevent over-fitting, and maximize classification performance. We employed LR, SVM, RF, and Bagg classifiers to enhance the prediction model's performance.

Comparisons were made between our proposed methods and individual algorithms such as RF, RRS, and RLS. Additionally, we compared our methods with filter methods such as MIFS. Furthermore, we compared our proposed methods with other hybrid approaches that utilize RFS, RRS, and RLS. To evaluate the performance of our proposed methods, we employed stratified 30-fold cross-validation. The following metrics were used for performance evaluation.

3.1. Metrics

- Model Evaluation Metrics:- To assess the performance of the classification model, we use several metrics, including F1-score, Precision, Recall, variance, Area Under the Curve (AUC) and Receiver Operating Characteristic (ROC) area. These metrics help in quantifying the effectiveness of the model [1, 8, 35].

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} \tag{4}$$

$$\text{Recall (Sensitivity)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{Precision (PPV)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

- Processing time: - is the fitting time in second.

- Classification accuracy:- measures the proportion of correctly classified samples among the total number of samples evaluated by a classifier. It is ex pressed as the percentage of correct predictions out of all predictions made by the model.

Our proposed methods applied different classifiers included in:-

1. SVM to:-

- Handles High-Dimensional Data Well.
- SVM uses a margin-maximization principle which helps it generalize well to unseen data and avoid overfitting.
- You can apply different kernels (e.g., linear, RBF) to handle both linear and non-linear data patterns, which is useful in biological systems.

2. Logistic regression to:-

- Simple and Interpretable.
- Fast and Lightweight.
- Effective for Binary Classification.

3. Bagging to:-

- Improve the stability and accuracy.
- Reduce the risk of overfitting.
- Combine multiple weak learners.
- Handle noisy data more effectively.

4. RF to:-

- well-known robustness.
- Ability to handle high-dimensional datasets such as RNA-seq data.
- Built-in capability to estimate feature importance

3.2. Hyperparameter tuning

To improve model performance and ensure robust feature selection in high-dimensional datasets, we performed hyperparameter tuning for all applied classifiers. The tuning was performed using stratified k-fold cross-validation to maintain class balance during training and validation. The impact of hyperparameter tuning was evaluated by comparing model performance before and after tuning. The results showed a noticeable improvement in accuracy, precision, recall, and F1 score, confirming the importance of proper hyperparameter optimization in high-dimensional settings.

Additionally, we applied Ridge regression with 5-fold cross-validation (CV=5) to ensure that the model's performance was stable and reliable across different data subsets. The use of cross-validation helped us assess the generalizability of our model and further validate the effectiveness of the hyperparameter tuning.

The experiments were conducted using Python on a PC running Windows 10. The system is equipped with an 11th Gen Intel® Core™ i7-1165G7 processor, operating at 2.80 GHz, and is supported by 16 GB of RAM. The hyperparameters for the experiments were determined through an iterative trial-and-error process and set randomly. Details regarding the hyperparameter settings and their corresponding values are presented in Table 5.

3.3. Numerical Results

The proposed method aims to address the previous limitations and assist specialists in making informed decisions by selecting the most effective features. To achieve this, we employed a 30-fold stratified cross-validation technique, which ensures balanced label classes in each fold, similar to the entire dataset. Table 6 presents an ablation study for each algorithm individually and shows the results before each component of the proposed algorithms. Each combination was independently evaluated to assess how EDF interacts with different embedded methods. The goal was to analyze the performance and compatibility of EDF with LRS, RRS, and RFS separately. This design allowed us to examine the strength of each hybrid path rather than merging all techniques into a single ensemble. Table6 shows the results of each component in the algorithm, i.e. shows the results of the separated RFS, LRS and RRS without adding the EDF equation and they did not give the best results.

It is evident that the single algorithm required more time for the fitting process, resulting in suboptimal performance metrics across all datasets. We also observed various issues with the single algorithm.

To overcome these challenges, we developed our methods incorporating hybrid algorithms that take advantage of EDF with different locations and embedded algorithms. These methods were specifically designed to address and resolve the aforementioned issues.

Table 6. The Performance of Individual Algorithms for All Datasets

Metrics	RNA gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
RFS Algorithm						

Metrics	RNA gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
Train Data (%)	100.000	89.803	76.484	53.426	95.000	89.860
Test Data (%)	99.800	84.054	75.000	53.333	93.000	87.725
Over-Fitting Difference	0.200	5.749	1.484	0.509	93.000	2.135
Precision	0.999	0.819	0.629	0.509	0.938	0.837
Recall	0.998	0.764	0.557	0.533	0.928	0.821
F1-score	0.998	0.775	0.537	0.397	0.932	0.815
No of Features	374.000	1234.000	224.000	12.000	27.000	11.000
Fitting Time (S)	13.015	5.000	1.474	0.104	0.090	0.094
Classification Time (S)	0.275	5.085	0.158	0.0008	0.008	0.002
AUC	1.000	0.955	0.706	0.867	0.989	0.985
Variance	0.00002	0.000193	0.00108	0.002623	0.00583	0.002819
Standard deviation	0.00447	0.01389	0.03286	0.05122	0.07638	0.05308
Accuracy	99.800	84.054	75.000	53.333	93.000	87.725
LRS Algorithm						
Train Data (%)	100.000	97.241	76.382	74.722	94.345	97.936
Test Data (%)	99.377	85.288	74.856	69.167	93.850	97.545
Over-Fitting Difference	0.623	11.953	1.526	5.555	0.495	0.391
Precision	0.995	0.824	0.540	0.711	0.945	0.976
Recall	0.993	0.793	0.544	0.692	0.929	0.971
F1-score	0.993	0.800	0.506	0.672	0.933	0.972
No of Features	1486.000	11265.000	487.000	43.000	22.000	32.000
Fitting Time (S)	0.267	3.476	0.116	0.008	0.811	0.008
Classification Time (S)	0.386	15.052	0.206	0.0008	0.101	0.002
AUC	0.999	0.961	0.679	0.744	0.989	0.998
Variance	0.000043	0.000323	0.005194	0.029398	0.002116	0.000408
Standard deviation	0.00656	0.01797	0.07205	0.17143	0.04599	0.02020
Accuracy	99.377	85.288	74.856	69.167	93.850	97.545
RRS Algorithm						
Train Data (%)	100.000	96.929	91.196	74.306	92.072	97.966
Test Data (%)	99.875	86.077	80.419	66.667	91.570	97.545
Over-Fitting Difference	0.125	10.852	10.777	7.639	0.502	0.421
Precision	0.999	0.842	0.746	0.680	0.917	0.976
Recall	0.998	0.799	0.730	0.667	0.903	0.971
F1-score	0.999	0.810	0.733	0.650	0.908	0.972
No of Features	2195.000	7983.000	574.000	46.000	28.000	28.000
Fitting Time (S)	0.356	1.941	0.019	0.003	0.008	0.009
Classification Time (S)	0.611	10.223	0.548	0.018	0.004	0.002
AUC	1.000	0.965	0.797	0.729	0.964	0.998
Variance	0.000016	0.000328	0.001918	0.019676	0.000801	0.000408
Standard deviation	0.00400	0.01811	0.04380	0.14024	0.02828	0.02020
Accuracy	99.875	86.077	80.419	66.667	91.570	97.545

To strengthen our analysis, we can also include the performance results of the EDF method alone, without combining it with any embedded algorithms in a new table called Table 7. This will allow us to better isolate and evaluate the standalone contribution of EDF and to more clearly demonstrate the added value of each hybrid combination.

Table 7. The Performance of Individual EDF Equation for All Datasets

Metrics	RNA gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
EDF Equation						
Train Data (%)	99.987	99.272	79.379	75.509	94.933	99.856
Test Data (%)	99.630	90.532	78.309	73.750	94.520	98.085
Over-Fitting Difference	0.357	8.740	1.070	1.759	0.413	1.771
Precision	0.996	0.779	0.756	0.755	0.950	0.985
Recall	0.995	0.876	0.592	0.732	0.934	0.982
F1 Score	0.995	0.886	0.597	0.730	0.938	0.980
No. of Features	20531	16381	753	46	30	34
Classification Time (s)	15.131	131.430	0.279	0.004	0.004	0.016
AUC	0.999	0.974	0.732	0.806	0.994	0.997
Variance	0.000128	0.025951	0.000733	0.006578	0.001203	0.000718
Standard Deviation	0.01131	0.16108	0.02707	0.08110	0.03469	0.02679
Accuracy	99.630	90.532	78.309	73.750	94.520	98.085

Table 8 shows the results of our proposed methods after 40 runs. The proposed method used EDF with different locations when applying RFS. It is applied in external, internal, and both locations. The EEDF-RFS results showed that DNA CNV dataset gave the best results for

Table 5. Hyperparameter Settings and Definitions

Parameter	Definition	Value
NRuns	Number of runs	40
Problem Dimensions	No. of features F in the dataset.	Different size
X*	No. of data produced after the bootstrap resample method.	Different size
M	The number of trees used in the Random Forest algorithm.	100
Criterion	The method that measures the quality of split, Entropy, is applied.	–
min_samples_leaf	The minimum number of samples required to be at a leaf node.	100
λ	Alpha symbol in RLS and RRS algorithms used for regularization.	0.1
Tol	Tolerance to stop criteria in LRS and RRS algorithms.	0.0001
Max-iteration	Max iteration in LR classifier.	100
CV	No. of folds in cross-validation.	30

this location with classification accuracy value 94.477%, precision, recall, F1-score, and AUC values are 0.947, 0.925, 0.931 and 0.987, respectively, using Bagg classifier. RNA gene dataset achieved the best results with classification accuracy 100.000%, precision, recall, F1-score, and AUC values of 1.000 using RF, SVM and LR classifiers. The RF achieved the best accuracy value for Parkinson's disease1 and Parkinson's disease2 to become 96.030% and 94.167%.

Table 8. Average Results After Applying EEDF-RFS after 40 Runs on Various Datasets with Different Classifiers

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
EEDF-RFS Algorithm						
Bagg classifier						
Train Data %	99.987	99.047	99.808	99.138	99.642	100.000
Test Data %	99.630	94.477	94.711	91.250	97.719	99.730
Over-Fitting Difference	0.357	4.570	5.097	7.888	1.923	0.270
Precision	0.998	0.947	0.935	0.931	0.980	0.996
Recall	0.995	0.925	0.924	0.911	0.976	0.994
F1_score	0.996	0.931	0.929	0.909	0.976	0.994
No of features	364	1535	138	7	3	13
Fitting time (S)	0.288	3.284	0.817	0.063	0.088	0.017
Classification time (S)	0.333	3.361	0.269	0.019	0.011	0.011
AUC	0.999	0.987	0.971	0.926	0.990	1.000
Variance	0.000128	0.000451	0.000889	0.008782	0.001850	0.000073
Standard deviation &0.01131	0.02123	0.02983	0.09375	0.04301	0.00854	
Accuracy	99.630	94.477	94.711	91.250	97.719	99.730
RF classifier						
Train Data %	100.000	97.169	100.000	100.000	100.000	100.000
Test Data %	100.000	92.420	96.030	94.167	97.018	99.459
Over-Fitting Difference	0.0	4.749	3.970	5.833	2.982	0.541
Precision	1.000	0.930	0.962	0.953	0.972	0.994
Recall	1.000	0.887	0.932	0.940	0.970	0.989
F1_score	1.000	0.899	0.945	0.940	0.969	0.990
No of features	364	1535	138	7	3	13
Fitting time (S)	0.288	3.284	0.817	0.063	0.088	0.017
Classification time (S)	0.342	0.965	0.761	0.105	0.088	0.042
AUC	1.000	0.990	0.990	0.956	0.994	1.000
Variance	0.0	0.000576	0.000615	0.006178	0.001659	0.000130
Standard Deviation	0.0	0.0240	0.0248	0.0786	0.0407	0.0114
Accuracy	100.000	92.420	96.030	94.167	97.018	99.459
SVM classifier						

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
Train Data %	100.000	94.936	76.984	84.298	89.649	64.511
Test Data %	100.000	91.324	75.675	81.667	89.546	63.408
Over-Fitting Difference	0.0	3.612	1.309	2.631	0.103	1.103
Precision	1.000	0.927	0.593	0.834	0.903	0.424
Recall	1.000	0.878	0.547	0.817	0.887	0.577
F1_score	1.000	0.889	0.525	0.813	0.889	0.480
No of features	364	1535	138	7	3	13
Fitting time (S)	0.288	3.284	0.817	0.063	0.088	0.017
Classification time (S)	0.049	8.410	0.329	0.007	0.074	0.037
AUC	1.000	0.990	0.649	0.882	0.973	0.965
Variance	0.0	0.000844	0.001104	0.014327	0.006206	0.004007
Standard Deviation	0.0	0.0291	0.0332	0.1197	0.0788	0.0633
Accuracy	100.000	91.324	75.675	81.667	89.649	63.408
LR classifier						
Train Data %	100.000	94.217	75.926	52.268	87.880	99.909
Test Data %	100.000	90.774	75.405	52.083	87.875	99.459
Over-Fitting Difference	0.0	3.443	0.521	0.185	0.005	0.450
Precision	1.000	0.910	0.604	0.309	0.887	0.997
Recall	1.000	0.866	0.535	0.504	0.871	0.995
F1_score	1.000	0.876	0.509	0.349	0.872	0.995
No of features	364	1535	138	7	3	13
Fitting time (S)	0.288	3.284	0.817	0.063	0.088	0.017
Classification time (S)	0.102	1.751	0.081	0.0005	0.003	0.003
AUC	1.000	0.985	0.715	0.889	0.953	1.000
Variance	0.0	0.000795	0.000539	0.000482	0.0061367	0.000292
Standard Deviation	0.0	0.0282	0.0232	0.0220	0.0783	0.0171
Accuracy	100.000	90.774	75.405	52.083	87.875	99.459

Furthermore, Table 9 presents the outcomes obtained from employing the proposed methods with EEDF-LRS. Notably, the RNA gene dataset exhibited the most favorable results when utilizing this specific location, achieving a classification ac curacy value of 100.000%. Additionally, precision, recall, F1-score, and AUC values were all 1.000 when employing the LR classifier. For the DNA CNV dataset, EEDF LRS also yielded impressive results, with an accuracy of 94.850%. The precision, recall, F1-score, and AUC values were 0.950, 0.925, 0.934 and 0.988, respectively. In the case of the dermatology dataset, our proposed classifiers demonstrated superior performance across the board, resulting in the best outcomes when compared to other datasets.

Table 9. Average Results after Applying EEDF-LRS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
EEDF-LRS Algorithm						
Bagg classifier						
Train Data %	99.974	99.405	99.722	98.922	99.745	100.000
Test Data %	97.607	94.850	93.400	86.250	98.000	100.000
Over-Fitting Difference	2.367	4.555	6.322	12.672	1.745	0.0
Precision	0.983	0.950	0.914	0.888	0.982	1.000
Recall	0.973	0.925	0.908	0.860	0.976	1.000
F1-score	0.975	0.934	0.904	0.856	0.978	1.000
No. of features	836	1049	334	9	18	22
Fitting time (s)	0.278	1.964	0.247	0.0010	0.024	0.004
Classification time (s)	0.833	1.928	3.060	0.011	0.025	0.010
AUC	0.998	0.988	0.958	0.925	0.990	1.000
Variance	0.000642	0.000556	0.002308	0.012231	0.001452	0.0
Standard Deviation	0.0253	0.0236	0.0480	0.1106	0.0381	0.0000
Accuracy	97.607	94.850	93.400	86.250	98.000	100.000
RF classifier						
Train Data %	100.000	97.672	100.000	100.000	100.000	100.000
Test Data %	99.240	93.722	95.226	94.167	98.000	100.000
Over-Fitting Difference	0.760	3.950	94.774	5.833	2.000	0.0
Precision	0.994	0.951	0.957	0.953	0.981	1.000
Recall	0.990	0.907	0.908	0.940	0.974	1.000
F1-score	0.991	0.919	0.923	0.940	0.976	1.000
No. of features	836	1049.000	334	9.000	18.000	22.000

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
Fitting time (s)	0.278	1.964	0.247	0.0010	0.024	0.004
Classification time (s)	0.608	0.732	1.99	0.094	0.116	0.040
AUC	1.000	0.993	0.981	0.956	0.998	1.000
Variance	0.000341	0.000561	0.002137	0.006178	0.001850	0.0
Standard Deviation	0.0185	0.0237	0.0462	0.0786	0.0430	0.0000
Accuracy	99.240	93.722	95.226	94.167	98.000	100.000
SVM classifier						
Train Data %	98.510	97.173	76.589	61.272	94.509	100.000
Test Data %	98.247	93.380	76.587	60.833	93.509	100.000
Over-Fitting Difference	0.263	3.793	0.002	0.439	1.000	0.000
Precision	0.990	0.947	0.383	0.599	0.946	1.000
Recall	0.973	0.901	0.500	0.597	0.926	1.000
F1 Score	0.978	0.916	0.434	0.535	0.930	1.000
No. of Features	836	1049	334	9	18	22
Fitting Time (S)	0.278	1.964	0.247	0.0010	0.024	0.004
Classification Time (S)	0.521	5.695	0.706	0.012	0.021	0.012
AUC	1.000	0.995	0.831	0.615	0.998	1.000
Variance	0.000560	0.000307	0.000027	0.022003	0.002996	0.000
Standard Deviation	0.0237	0.0175	0.0052	0.1483	0.0547	0.0000
Accuracy	98.247	93.380	76.587	60.833	93.509	100.000
LR Classifier						
Train Data %	100.000	96.383	76.665	51.667	81.716	100.000
Test Data %	100.000	93.759	76.451	51.667	81.559	100.000
Over-Fitting Difference	0.000	2.624	0.214	0.000	0.157	0.000
Precision	1.000	0.946	0.401	0.258	0.876	1.000
Recall	1.000	0.898	0.505	0.500	0.778	1.000
F1 Score	1.000	0.913	0.444	0.340	0.781	1.000
No. of Features	836	1049	334	9	18	22
Fitting Time (S)	0.278	1.964	0.247	0.0010	0.024	0.004
Classification Time (S)	0.219	0.735	0.136	0.0004	0.0003	0.003
AUC	1.000	0.991	0.803	0.603	0.954	1.000
Variance	0.000	0.000444	0.000990	0.001868	0.006799	0.000
Standard Deviation	0.0000	0.0211	0.0315	0.0432	0.0825	0.0000
Accuracy	100.000	93.759	76.451	51.667	81.559	100.000

The results in Table 10 of the proposed methods used EEDF-RRS. It showed that the LR and RF classifiers gave the best results for RNA gene datasets with 100.000% classification accuracy. In addition, the Bagg classifier achieved the best accuracy results for DNA CNV and BreastEW datasets to become 94.683% and 98.596%, respectively. The RF classifier gave the best accuracy results for Parkinson's disease1 and Parkinson's disease2 datasets to become 96.426% and 94.16%, respectively. All classifiers gave the full percentage results for the dermatology erythemato-squamous diseases dataset.

Table 10. Average Results after Applying EEDF-RRS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
EEDF-RRS Algorithm						
Bagg Classifier						
Train Data (%)	99.927	99.386	99.790	99.253	99.806	100.000
Test Data (%)	97.745	94.683	94.713	91.250	98.596	100.000
Over-Fitting Difference	2.182	4.703	5.077	8.003	1.210	0.0
Precision	0.984	0.945	0.941	0.928	0.988	1.000
Recall	0.978	0.925	0.927	0.913	0.984	1.000
F1 Score	0.979	0.932	0.928	0.9103	0.985	1.000
No of Features	1739	562	581	9	22	25
Fitting Time (s)	1.000	0.953	1.381	0.010	0.111	0.003
Classification Time (s)	1.162	1.033	1.361	0.012	0.024	0.010
AUC	0.998	0.948	0.979	0.957	0.996	1.000
Variance	0.001232	0.000425	0.002750	0.009860	0.000560	0.0
Standard deviation	0.0351	0.0206	0.0524	0.0993	0.0237	0.0000
Accuracy	97.745	94.683	94.713	91.250	98.596	100.000
RF Classifier						
Train Data (%)	100.000	97.421	100.000	100.000	100.000	100.000

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
Test Data (%)	100.000	93.380	96.426	94.167	98.070	100.000
Over-Fitting Difference	0.0	4.041	3.574	5.833	1.930	0.0
Precision	1.000	0.949	0.972	0.953	0.984	1.000
Recall	1.000	0.896	0.933	0.940	0.979	1.000
F1 Score	1.000	0.911	0.949	0.940	0.980	1.000
No of Features	1739	562	581	9	22	25
Fitting Time (s)	1.000	0.953	1.381	0.010	0.111	0.003
Classification Time (s)	0.677	0.900	0.598	0.090	0.106	0.042
AUC	1.000	0.993	0.992	0.956	0.999	1.000
Variance	0.0	0.000576	0.001353	0.006178	0.001048	0.0
Standard deviation	0.0000	0.0240	0.0368	0.0786	0.0324	0.0000
Accuracy	100.000	93.380	96.426	94.167	98.070	100.000
SVM Classifier						
Train Data %	69.839	96.700	78.288	63.157	89.801	100.000
Test Data %	68.272	93.069	77.795	62.083	89.464	100.000
Over-Fitting Diff.	1.567	3.631	0.493	1.074	0.337	0.000
Precision	0.740	0.944	0.670	0.657	0.909	1.000
Recall	0.620	0.893	0.553	0.615	0.881	1.000
F1 Score	0.622	0.910	0.524	0.573	0.886	1.000
No. of Features	1739	562	581	9	22	25
Fitting Time (s)	1.000	0.953	1.381	0.010	0.111	0.003
Classification Time (s)	1.461	3.192	0.231	0.011	0.026	0.011
AUC	1.000	0.994	0.815	0.689	0.972	1.000
Variance	0.003086	0.000413	0.001037	0.018622	0.004957	0.000
Standard deviation	0.05555	0.02032	0.03220	0.13644	0.07043	0.00000
Accuracy	68.272	93.069	77.795	62.083	89.463	100.000
LR Classifier						
Train Data %	100.000	94.161	94.650	62.744	90.702	100.000
Test Data %	100.000	90.878	89.825	62.500	90.700	100.000
Over-Fitting Diff.	0.000	3.283	4.825	0.244	0.002	0.000
Precision	1.000	0.909	0.874	0.640	0.920	1.000
Recall	1.000	0.855	0.846	0.621	0.894	1.000
F1 Score	1.000	0.869	0.857	0.597	0.900	1.000
No. of Features	1739	562	581	9	22	25
Fitting Time (s)	1.000	0.953	1.381	0.010	0.111	0.003
Classification Time (s)	0.293	0.302	0.514	0.0003	0.00054	0.003
AUC	1.000	0.986	0.926	0.685	0.924	1.000
Variance	0.000	0.000605	0.001292	0.029095	0.004907	0.000
Standard deviation	0.00000	0.02460	0.03595	0.17060	0.07005	0.00000
Accuracy	100.000	90.878	89.825	62.500	90.700	100.000

The results in Table 11 of the proposed methods used IEDF-RFS for the second location. It showed the EDF in an internal location. The SVM and LR classifiers achieved 100.000% accuracy results for RNA gene dataset. The RF classifier gave the best accuracy results for the BreastEW dataset to become 98.421%. The IEDF-RFS did not give the best results for other datasets.

Table 11. Average Results after Applying IEDF-RFS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
IEDF-RFS Algorithm						
Bagg Classifier						
Train Data %	42.219	97.676	99.539	98.218	99.794	98.519
Test Data %	39.834	78.637	80.703	75.000	98.070	81.725
Over-Fitting Difference	2.385	19.039	18.836	23.218	1.724	16.794
Precision	0.210	0.730	0.769	0.788	0.984	0.774
Recall	0.229	0.698	0.727	0.750	0.978	0.792
F1 Score	0.162	0.703	0.726	0.728	0.980	0.757
No of Features	247	1001	141	25	16	8
Fitting Time (S)	0.218	5.133	1.025	0.123	0.152	0.005
Classification Time (S)	0.010	10.600	1.108	0.113	0.089	0.060
AUC	0.533	0.909	0.814	0.814	0.993	0.950
Variance	0.000746	0.001668	0.010533	0.031250	0.001239	0.009160
Standard deviation	0.02732	0.04083	0.10262	0.17678	0.03521	0.09573

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
Accuracy	39.834	78.637	80.703	75.000	98.070	81.725
RF Classifier						
Train Data %	100.000	89.154	100.000	100.000	100.000	99.871
Test Data %	99.748	79.668	83.354	77.917	98.421	83.655
Over-Fitting Difference	0.252	9.486	16.646	22.083	1.579	16.216
Precision	0.998	0.749	0.792	0.830	0.987	0.815
Recall	0.997	0.706	0.734	0.779	0.982	0.819
F1_score	0.998	0.709	0.741	0.762	0.983	0.795
No of features	247.000	1001.000	141.000	25.000	16.000	8.000
Fitting time (S)	0.218	5.133	1.025	0.123	0.152	0.005
Classification time (S)	0.378	3.980	0.969	0.392	0.515	0.272
AUC	1.000	0.940	0.863	0.873	0.997	0.974
Variance	0.000092	0.001426	0.009885	0.026598	0.000984	0.008810
Standard Deviation	0.00959	0.03776	0.09942	0.16306	0.03137	0.09387
Accuracy	99.748	79.668	83.354	77.917	98.421	83.654
SVM Classifier						
Train Data %	100.000	88.817	75.661	72.778	89.746	52.445
Test Data %	100.000	81.961	72.246	71.667	89.649	51.140
Over-Fitting Difference	0.0	6.856	3.415	1.111	0.097	1.305
Precision	1.000	0.797	0.418	0.724	0.911	0.355
Recall	1.000	0.737	0.495	0.717	0.883	0.458
F1_score	1.000	0.747	0.438	0.708	0.888	0.373
No of features	247.000	1001.000	141.000	25.000	16.000	8.000
Fitting time (S)	0.218	5.133	1.025	0.123	0.152	0.005
Classification time (S)	0.053	36.823	0.411	0.046	0.117	0.208
AUC	1.000	0.959	0.637	0.827	0.972	0.869
Variance	0.0	0.001280	0.004331	0.01844	0.005441	0.004889
Standard Deviation	0.00000	0.03578	0.06583	0.13581	0.07378	0.06993
Accuracy	100.000	81.961	72.246	71.667	89.649	51.140
LR Classifier						
Train Data %	100.000	88.777	79.39	65.618	94.782	85.548
Test Data %	100.000	84.000	78.047	61.667	93.674	81.404
Over-Fitting Difference	0.0	4.777	0.343	3.951	1.108	4.144
Precision	1.000	0.824	0.719	0.620	0.944	0.767
Recall	1.000	0.764	0.624	0.617	0.931	0.768
F1_score	1.000	0.774	0.637	0.586	0.933	0.746
No of features	247.000	1001.000	141.000	25.000	16.000	8.000
Fitting time (S)	0.218	5.133	1.025	0.123	0.152	0.005
Classification time (S)	0.105	5.238	0.103	0.0043	0.014	0.016
AUC	1.000	0.958	0.745	0.671	0.985	0.969
Variance	0.0	0.000765	0.003703	0.047342	0.002749	0.010965
Standard Deviation	0.00000	0.02766	0.06085	0.21758	0.05243	0.10470
Accuracy	100.000	84.000	78.047	61.667	93.674	81.404

The results in Table 12 of the proposed methods used IEDF-LRS. The LR classifier achieved 100% accuracy results for the RNA gene dataset. The Bagg classifier gave the best accuracy results for Parkinson's disease1 dataset to become 95.231%. The RF classifier achieved the best accuracy results for Parkinson's disease2 and BreastEW datasets to become 92.500% and 98.246%, respectively. The Bagg, RF, and SVM classifiers achieved the best classification accuracy results for the Dermatology erythemato-squamous diseases dataset to become 100.000%.

Table 12. Average Results after Applying IEDF-LRS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
IEDF-LRS Algorithm						
Bagg classifier						
Train Data %	99.897	98.860	99.685	99.325	99.752	100.000
Test Data %	98.006	90.090	95.231	88.750	97.544	100.000
Over-Fitting Difference	1.891	8.770	4.454	10.575	2.208	0.0
Precision	0.981	0.901	0.949	0.907	0.976	1.000
Recall	0.977	0.866	0.924	0.889	0.972	1.000
F1-score	0.976	0.876	0.932	0.884	0.973	1.000

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
No of features	1645	674	211	12	18	20
Fitting time (S)	0.389	4.193	0.139	0.150	0.027	0.223
Classification time (S)	1.548	4.186	1.195	0.041	0.136	0.054
AUC	0.998	0.969	0.966	0.957	0.990	1.000
Variance	0.001597	0.000521	0.001807	0.011153	0.001286	0.0
Standard deviation	0.03997	0.02283	0.04250	0.10560	0.03586	0.00000
Accuracy	98.006	90.090	95.231	88.750	97.544	100.000
RF classifier						
Train Data %	100.000	96.021	100.000	100.000	100.000	100.000
Test Data %	99.753	89.437	94.328	92.500	98.246	100.000
Over-Fitting Difference	0.247	6.584	5.672	7.500	1.754	0.000
Precision	0.999	0.913	0.955	0.935	0.983	1.000
Recall	0.997	0.839	0.892	0.925	0.980	1.000
F1 Score	0.998	0.856	0.913	0.923	0.981	1.000
No of Features	1645.000	674.000	211.000	12.000	18.000	20.000
Fitting Time (S)	0.389	4.193	0.139	0.150	0.027	0.223
Classification Time (S)	0.903	2.075	0.896	0.310	0.647	0.432
AUC	1.000	0.983	0.988	0.970	0.997	1.000
Variance	0.000088	0.000855	0.002442	0.008190	0.001210	0.000
Standard deviation	0.00938	0.02924	0.04942	0.09050	0.03478	0.00000
Accuracy	99.753	89.437	94.328	92.500	98.246	100.000
SVM classifier						
Train Data %	99.251	95.755	80.732	77.632	91.603	100.000
Test Data%	99.245	91.000	80.691	76.667	91.563	100.000
Over-Fitting Difference	0.006	4.755	0.041	0.965	0.040	0.000
Precision	0.996	0.925	0.831	0.788	0.929	1.000
Recall	0.989	0.857	0.609	0.765	0.890	1.000
F1 Score	0.992	0.876	0.614	0.761	0.904	1.000
No of Features	1645	674	211	12	18	20
Fitting Time (S)	0.389	4.193	0.139	0.150	0.027	0.223
Classification Time (S)	0.922	32.072	0.281	0.077	0.115	0.042
AUC	1.000	0.985	0.808	0.863	0.972	1.000
Variance	0.000236	0.000643	0.001115	0.012135	0.000809	0.000
Standard deviation	0.01536	0.02536	0.03339	0.11016	0.02844	0.00000
Accuracy	99.245	91.000	80.691	76.667	91.563	100.000
LR classifier						
Train Data %	100.000	93.404	92.445	74.167	90.002	99.396
Test Data %	100.000	89.919	89.939	74.028	89.637	98.070
Over-Fitting Difference	0.0	3.485	2.506	0.139	0.365	1.326
Precision	1.000	0.902	0.888	0.762	0.931	0.940
Recall	1.000	0.857	0.834	0.745	0.885	0.945
F1 Score	1.000	0.869	0.853	0.736	0.877	0.938
No. of Features	1645	674	211	12	18	20
Fitting Time (s)	0.389	4.193	0.139	0.150	0.027	0.223
Classification Time (s)	0.334	1.299	0.908	0.002	0.006	0.022
AUC	1.000	0.978	0.935	0.820	0.984	1.000
Variance	0.0	0.001132	0.000610	0.007639	0.001321	0.000728
Standard deviation	0.0	0.03363	0.02470	0.08742	0.03634	0.02698
Accuracy	100.000	89.919	89.939	74.028	89.637	98.070

On the other hand, the results in Table 13 of the proposed methods used IEDF-RRS. RNA gene dataset gave the best accuracy results using LR classifier to become 100.000%. Parkinson's disease1, Parkinson's disease2, and BreastEW datasets achieved 95.918%, 92.500%, and 98.947% accuracy results using RF classifier, respectively. All classifiers gave the best accuracy results for the dermatology erythemato-squamous diseases dataset to become 100.000%.

Table 13. Average Results after Applying IEDF-RRS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
IEDF-RRS Algorithm						
Bagg Classifier						
Train Data %	99.892	98.599	99.758	99.195	99.849	100.000
Test Data %	97.626	89.468	94.856	90.000	98.762	100.000
Over-Fitting Difference	2.266	9.131	4.902	9.195	1.087	0.000

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
Precision	0.982	0.890	0.940	0.916	0.989	1.000
Recall	0.976	0.851	0.927	0.898	0.986	1.000
F1-score	0.977	0.862	0.930	0.896	0.986	1.000
No. of Features	1653	1170	291	22	21	25
Fitting Time (s)	0.297	2.601	1.025	0.150	0.175	0.110
Classification Time (s)	1.075	9.431	1.499	0.110	0.121	0.059
AUC	0.997	0.863	0.971	0.942	0.994	1.000
Variance	0.000527	0.000630	0.002519	0.012284	0.000521	0.000
Standard deviation	0.02295	0.02510	0.05019	0.11088	0.02283	0.00000
Accuracy	97.626	89.468	94.856	90.000	98.762	100.000
RF Classifier						
Train Data %	100.000	95.890	100.000	99.986	100.000	100.000
Test Data %	99.753	88.819	95.918	92.500	98.947	100.000
Over-Fitting Difference	0.247	7.071	4.082	7.486	1.053	0.0
Precision	0.998	0.900	0.964	0.932	0.991	1.000
Recall	0.997	0.828	0.928	0.923	0.988	1.000
F1 Score	0.998	0.844	0.941	0.923	0.989	1.000
No. of Features	1653	1170	291.000	22.000	21.000	25.000
Fitting Time (s)	0.297	2.601	1.025	0.150	0.175	0.110
Classification Time (s)	0.707	3.202	1.080	0.536	0.554	0.200
AUC	0.999	0.977	0.988	0.964	1.000	1.000
Variance	0.000088	0.000725	0.002008	0.012499	0.000458	0.0
Standard deviation	0.00938	0.02693	0.04460	0.11177	0.02140	0.00000
Accuracy	99.753	88.819	95.918	92.500	98.947	100.000
SVM Classifier						
Train Data (%)	81.510	95.187	74.208	87.412	93.013	100.000
Test Data (%)	80.783	91.000	74.206	87.083	92.963	100.000
Over-Fitting Difference	0.727	4.187	0.002	0.329	0.050	0.000
Precision	0.733	0.926	0.371	0.881	0.945	1.000
Recall	0.788	0.854	0.500	0.870	0.910	1.000
F1-score	0.752	0.875	0.425	0.869	0.920	1.000
No. of Features	1653	1170	291	22	21	250
Fitting Time (s)	0.297	2.601	1.025	0.150	0.175	0.110
Classification Time (s)	1.438	35.067	0.728	0.045	0.105	0.072
AUC	1.000	0.983	0.828	0.926	0.973	1.000
Variance	0.000510	0.000735	0.000032	0.011312	0.002755	0.000
Standard Deviation	0.02258	0.02711	0.00566	0.10633	0.05249	0.00000
Accuracy (%)	80.783	91.000	74.206	87.083	92.963	100.000
LR Classifier						
Train Data (%)	100.000	94.487	75.147	59.444	88.267	100.000
Test Data (%)	100.000	90.462	75.002	59.167	88.051	100.000
Over-Fitting Difference	0.000	4.025	0.145	0.277	0.216	0.000
Precision	1.000	0.915	0.688	0.634	0.895	1.000
Recall	1.000	0.854	0.522	0.562	0.856	1.000
F1-score	1.000	0.870	0.473	0.464	0.863	1.000
No. of Features	1653	1170	291	22	21	25
Fitting Time (s)	0.297	2.601	1.025	0.150	0.175	0.110
Classification Time (s)	0.303	2.992	0.049	0.005	0.007	0.037
AUC	1.000	0.978	0.802	0.925	0.926	1.000
Variance	0.000	0.000752	0.000159	0.002623	0.004568	0.000
Standard Deviation	0.00000	0.02743	0.01261	0.05122	0.06760	0.0
Accuracy (%)	100.000	90.462	75.002	59.167	88.051	100.000

In contrast; Table 14 presents the outcomes obtained when employing the proposed methods with E/IEDF-RFS. Notably, the RNA gene dataset achieved the highest accuracy result of 100% when utilizing the LR classifier. For the Parkinson's Disease1 dataset, Parkinson's Disease2 dataset, and BreastEW dataset, the RF classifier yielded accuracy results of 96.000%, 94.583%, and 98.246%, respectively. In the case of the dermatology erythemato-squamous diseases dataset, both the Bagg and RF classifiers demonstrated exceptional accuracy results, with a perfect score of 100.000%.

Table 14. Average Results after Applying E/IEDF-RFS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
E/IEDF-RFS Algorithm						
Bagg Classifier						
Train Data (%)	99.966	98.255	99.826	99.425	99.788	100.000
Test Data (%)	99.373	90.570	95.103	91.667	97.193	100.000
Over-Fitting Difference	0.593	7.685	4.723	7.758	2.595	0.000
Precision	0.996	0.910	0.947	0.929	0.975	1.000
Recall	0.993	0.880	0.939	0.916	0.971	1.000
F1-score	0.993	0.889	0.939	0.914	0.971	1.000
No. of Features	227	1021	151	26	13	7
Fitting Time (s)	14.520	2.501	0.210	0.251	0.250	0.254
Classification Time (s)	1.057	6.043	0.913	0.056	0.101	0.066
AUC	0.999	0.973	0.978	0.946	0.994	1.000
Variance	0.000203	0.000840	0.002019	0.011135	0.001668	0.000
Standard deviation	0.01425	0.02898	0.04492	0.10551	0.04084	0.00000
Accuracy (%)	99.373	90.570	95.103	91.667	97.193	100.000
RF Classifier						
Train Data (%)	100.000	93.767	100.000	99.943	100.000	100.000
Test Data (%)	99.744	88.168	96.000	94.583	98.246	100.000
Over-Fitting Difference	0.256	5.599	4.000	5.360	1.754	0.000
Precision	0.999	0.888	0.961	0.953	0.985	1.000
Recall	0.997	0.835	0.940	0.946	0.981	1.000
F1-score	0.998	0.850	0.948	0.945	0.982	1.000
No. of Features	227	1021	151	26	13	7
Fitting Time (s)	14.520	2.501	0.210	0.251	0.250	0.254
Classification Time (s)	1.907	2.206	0.829	0.329	0.465	0.232
AUC	1.000	0.976	0.991	0.971	0.997	1.000
Variance	0.000095	0.000936	0.001583	0.008279	0.000828	0.000
Standard deviation	0.00975	0.03060	0.03979	0.09101	0.02877	0.00000
Accuracy (%)	99.744	88.168	96.000	94.583	98.246	100.000
SVM Classifier						
Train Data (%)	100.000	92.146	73.104	72.917	89.716	60.995
Test Data (%)	99.872	87.586	71.958	72.851	89.474	60.045
Over-Fitting Difference	0.128	4.560	1.146	0.066	0.242	0.950
Precision	0.999	0.899	0.615	0.758	0.908	0.456
Recall	0.999	0.823	0.550	0.726	0.881	0.498
F1-score	0.999	0.843	0.531	0.718	0.886	0.452
No. of Features	227	1021	151	26	13	7
Fitting Time (s)	14.520	2.501	0.210	0.251	0.250	0.254
Classification Time (s)	0.252	20.005	0.461	0.041	0.126	0.174
AUC	1.000	0.980	0.657	0.814	0.973	0.934
Variance	0.000049	0.001036	0.000260	0.016721	0.005349	0.010545
Standard deviation	0.00700	0.03218	0.01612	0.12927	0.07313	0.10267
Accuracy (%)	99.872	87.586	71.958	72.851	89.474	60.045
LR Classifier						
Train Data (%)	100.000	90.796	72.076	74.676	92.564	95.630
Test Data (%)	100.000	88.029	71.944	73.750	91.929	95.628
Over-Fitting Difference	0.000	2.767	0.132	0.926	0.635	0.002
Precision	1.000	0.891	0.586	0.780	0.924	0.971
Recall	1.000	0.834	0.542	0.733	0.912	0.948
F1-score	1.000	0.849	0.512	0.721	0.915	0.952
No. of Features	227	1021	151	26	13	7
Fitting Time (s)	14.520	2.501	0.210	0.251	0.250	0.254
Classification Time (s)	0.990	3.216	0.072	0.003	0.009	0.015
AUC	1.000	0.973	0.686	0.790	0.981	0.997
Variance	0.000	0.001441	0.001993	0.010050	0.004343	0.000521
Standard deviation	0.00000	0.03795	0.04465	0.10025	0.06589	0.02283
Accuracy (%)	100.000	88.029	71.944	73.750	91.929	95.628

On the other hand, the results in Table 15 of the proposed methods used E/IEDF RLS. RNA gene dataset gave the best accuracy results using LR classifier to be come 100.000%. Parkinson's disease1, Parkinson's disease2, and BreastEW datasets achieved 95.000%, 94.167% and 98.421% accuracy results using RF classifier, respectively. All classifiers gave the best accuracy results for the dermatology erythemato squamous diseases dataset to become 100.000%.

Table 15. Average Results after Applying E/IEDF-LRS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
E/IEDF-LRS Algorithm						
Bagg Classifier						
Train Data (%)	99.901	99.102	99.754	99.080	99.739	100.000
Test Data (%)	97.142	89.953	94.585	87.083	97.193	100.000
Over-Fitting Difference	2.759	9.149	5.169	11.997	2.546	0.000
Precision	0.976	0.895	0.945	0.884	0.973	1.000
Recall	0.968	0.861	0.897	0.869	0.967	1.000
F1-score	0.969	0.871	0.910	0.867	0.968	1.000
No. of Features	788	666	101	10	19	20
Fitting Time (s)	13.520	12.400	0.114	0.200	0.180	0.130
Classification Time (s)	3.510	13.844	0.829	0.040	0.118	0.093
AUC	0.998	0.970	0.968	0.908	0.988	1.000
Variance	0.001116	0.000895	0.002295	0.022073	0.001095	0.000
Standard deviation	0.03341	0.02991	0.04791	0.14857	0.03309	0.00000
Accuracy (%)	97.142	89.953	94.585	87.083	97.193	100.000
RF Classifier						
Train Data (%)	100.000	99.078	100.000	100.000	100.000	100.000
Test Data (%)	99.373	88.237	95.000	94.167	98.421	100.000
Over-Fitting Difference	0.627	10.841	5.000	5.833	1.579	0.000
Precision	0.999	0.901	0.953	0.956	0.985	1.000
Recall	0.991	0.822	0.896	0.939	0.981	1.000
F1-score	0.993	0.842	0.916	0.936	0.982	1.000
No. of Features	788	666	101	10	19	20
Fitting Time (s)	13.520	12.400	0.114	0.200	0.180	0.130
Classification Time (s)	5.622	6.582	0.887	0.285	0.487	0.366
AUC	0.999	0.980	0.978	0.976	0.994	1.000
Variance	0.000203	0.001001	0.001676	0.009410	0.000793	0.000
Standard deviation	0.01425	0.03163	0.04094	0.09701	0.02817	0.00000
Accuracy (%)	99.373	88.237	95.000	94.167	98.421	100.000
SVM Classifier						
Train Data (%)	99.247	95.542	74.206	63.443	94.012	100.000
Test Data (%)	98.880	90.022	74.077	62.500	93.850	100.000
Over-Fitting Difference	0.367	5.520	0.129	0.943	0.162	0.000
Precision	0.995	0.920	0.370	0.641	0.952	1.000
Recall	0.986	0.846	0.500	0.618	0.918	1.000
F1-score	0.989	0.866	0.426	0.597	0.929	1.000
No. of Features	788	666	101	10	19	20
Fitting Time (s)	13.520	12.400	0.114	0.200	0.180	0.130
Classification Time (s)	2.758	34.523	0.341	0.042	0.112	0.046
AUC	0.999	0.985	0.770	0.706	0.980	1.000
Variance	0.000587	0.000869	0.000025	0.017178	0.002498	0.000
Standard deviation	0.02423	0.02948	0.00500	0.13106	0.04998	0.00000
Accuracy (%)	98.880	90.022	74.077	62.500	93.850	100.000
LR Classifier						
Train Data (%)	100.000	92.567	83.470	63.009	93.219	100.000
Test Data (%)	100.000	88.961	82.800	62.917	93.138	100.000
Over-Fitting Difference	0.000	3.606	0.670	0.092	0.081	0.000
Precision	1.000	0.899	0.761	0.679	0.949	1.000
Recall	1.000	0.835	0.638	0.621	0.907	1.000
F1-score	1.000	0.850	0.650	0.592	0.921	1.000
No. of Features	788	666	101	10	19	20
Fitting Time (s)	13.520	12.400	0.114	0.200	0.180	0.130
Classification Time (s)	1.455	1.955	0.051	0.002	0.006	0.014
AUC	1.000	0.980	0.770	0.690	0.982	1.000
Variance	0.000	0.000693	0.002135	0.012905	0.001955	0.000
Standard deviation	0.0000	0.0263	0.0462	0.1136	0.0442	0.0000
Accuracy (%)	100.000	88.961	82.800	62.916	93.138	100.000

On the other hand, the results in Table 16 of the proposed methods used E/IEDF-RRS. RNA gene dataset gave the best accuracy results using LR classifier to become 100.000%. Parkinson's Disease2 and BreastEW datasets achieved 94.167% and 99.288% accuracy using RF classifier, respectively. All classifiers gave the best accuracy results for the dermatology erythemato-squamous diseases dataset to become 100.000%.

Table 16. Average Results after Applying E/IEDF-RRS after 40 Runs

Metrics	RNA Gene	DNA CNV	Parkinson's1	Parkinson's2	BreastEW	Dermatology
E/IEDF-RRS Algorithm						
Bagg Classifier						
Train Data (%)	99.931	98.737	99.662	99.109	99.927	100.000
Test Data (%)	98.376	91.117	93.256	89.583	98.060	100.000
Over-Fitting Difference	1.555	7.620	6.406	9.526	1.867	0.000
Precision	0.986	0.913	0.927	0.919	0.983	1.000
Recall	0.981	0.881	0.910	0.889	0.979	1.000
F1-score	0.982	0.889	0.912	0.890	0.980	1.000
No. of Features	1573	1179	294	21	23	25
Fitting Time (s)	14.300	14.250	0.110	0.004	0.009	0.014
Classification Time (s)	4.507	23.423	1.979	0.054	0.162	0.086
AUC	0.998	0.973	0.967	0.970	0.996	1.000
Variance	0.000452	0.001160	0.001291	0.009788	0.001055	0.000
Standard deviation	0.0213	0.0341	0.0359	0.0989	0.0325	0.0000
Accuracy (%)	98.376	91.117	93.256	89.583	98.060	100.000
RF Classifier						
Train Data (%)	100.000	95.289	100.000	100.000	100.000	100.000
Test Data (%)	99.877	89.162	94.841	94.167	99.288	100.000
Over-Fitting Difference	0.123	6.127	5.159	5.833	0.712	0.000
Precision	0.999	0.901	0.957	0.956	0.993	1.000
Recall	0.999	0.840	0.914	0.936	0.993	1.000
F1-score	0.999	0.857	0.929	0.938	0.993	1.000
No. of Features	1573	1179	294	21	23	25
Fitting Time (s)	14.300	14.250	0.110	0.004	0.009	0.014
Classification Time (s)	2.994	8.280	2.641	0.337	0.713	0.226
AUC	1.000	0.980	0.988	0.985	1.000	1.000
Variance	0.000046	0.001115	0.001530	0.007256	0.000341	0.000
Standard deviation	0.0068	0.0334	0.0391	0.0852	0.0185	0.0000
Accuracy (%)	99.877	89.162	94.841	94.167	99.288	100.000
SVM Classifier						
Train Data (%)	80.481	95.065	73.149	82.829	92.452	100.000
Test Data (%)	79.520	90.121	73.148	82.500	92.438	100.000
Over-Fitting Difference	0.961	4.944	0.001	0.329	0.014	0.000
Precision	0.773	0.926	0.366	0.848	0.939	1.000
Recall	0.781	0.855	0.500	0.824	0.908	1.000
F1-score	0.758	0.876	0.422	0.819	0.917	1.000
No. of Features	1573	1179	294	21	23	25
Fitting Time (s)	14.300	14.250	0.110	0.004	0.009	0.014
Classification Time (s)	6.987	97.609	0.631	0.044	0.172	0.035
AUC	1.000	0.983	0.805	0.888	0.979	1.000
Variance	0.001164	0.001075	0.000031	0.010161	0.002248	0.000
Standard deviation	0.0341	0.0328	0.0056	0.1008	0.0474	0.0000
Accuracy (%)	79.520	90.121	73.148	82.500	92.438	100.000
LR Classifier						
Train Data (%)	100.000	94.047	73.148	56.435	87.025	100.000
Test Data (%)	100.000	90.190	73.148	55.833	86.813	100.000
Over-Fitting Difference	0.000	3.857	0.000	0.602	0.212	0.000
Precision	1.000	0.911	0.366	0.573	0.883	1.000
Recall	1.000	0.859	0.500	0.531	0.849	1.000
F1-score	1.000	0.873	0.422	0.409	0.855	1.000
No. of Features	1573	1179	294	21	23	25
Fitting Time (s)	14.300	14.250	0.110	0.004	0.009	0.014
Classification Time (s)	2.031	5.916	0.121	0.002	0.009	0.014
AUC	1.000	0.982	0.788	0.876	0.936	1.000
Variance	0.000	0.001125	0.000274	0.001235	0.003983	0.000
Standard deviation	0.0000	0.0335	0.0166	0.0351	0.0631	0.0000
Accuracy (%)	100.000	90.190	73.148	55.833	86.813	100.000

Table 17 presented the summary of recent related work which compared with our proposed methods. Various previous researches are showed which used the same datasets. Our proposed methods achieved the best results using genotype and phenotype datasets.

Table 17. Summary of recent cancer prediction studies for genotype and phenotype datasets with different FS methods and ML models

Ref.	Dataset	Cases	Genes	GS Method	Selected Genes	ML Model	Performance Metrics%
[30]	RNA Gene	801	20531	GGA	49	Voting System	ACC: 98.810
[7]	DNA CNV	2916	16381	mRMR & IFS	19	Dagging	ACC: 75.000 AUC: 0.973
[8]				PSO & GA	2050	RF, SVM, J48, LR, Bagg	ACC:84.600 AUC: 0.961
[12]				IG	16381	RF, SVM, J48, LR, Dagging, Bagg, Neural Network	ACC:85.900 AUC:0.965
[31]	Park1	756	753	mRMR	50	KNN	ACC: 85.000
[79]	Park2	240	46	Correlation Ranking	8	Stratified CV	ACC: 88.000 AUC: 0.951
[9]	BC	569	30	CSSA	5.200	Voting&stacking	ACC: 97.080
[1]	DNA CNV	2916	16381	PFBS-RFS-RFE	675.0	RL, SVM, RF, BAGG	ACC: 92.762, AUC: 0.981
	RNA Gene	801	20531		119.2		ACC: 99.994, AUC: 1.000
	Park1	756	753		113.85		ACC: 95.000, AUC: 0.985
	BC	569	30		13.3		ACC: 98.000, AUC: 0.997
	Derma	366	34		10.0		ACC: 100.000, AUC: 1.000

4. Biological Interpretation of Key Features

4.1. Functional Enrichment Analysis Using GO and KEGG Pathways

To further investigate the biological significance of genes identified by our model, we performed functional enrichment analysis using the KEGG pathway and Gene Ontology (GO) databases. This analysis aimed to uncover the involvement of these genes in known biological processes and disease-related pathways. The results revealed several significantly enriched pathways, highlighting the potential roles of selected genes in critical cellular mechanisms and disease progression. A summary of the enriched pathways and associated genes is presented in Table 18 , providing valuable biological context and supporting the relevance of our predictive gene set.

Table 18. Functional enrichment analysis of the identified genes using GO and KEGG pathways

Pathway	Overlap	P-value	Adjusted P-value	Odds Ratio	Genes
RIG-I-like receptor signaling pathway	3/70	0.006	0.393	8.454	IFNA1; IFNE; IL12B
Cell adhesion molecules	4/148	0.008	0.393	5.274	IGSF11; SDC2; NCAM1; ITGA9
Cellular senescence	4/156	0.010	0.393	4.994	CDKN2B; CDKN2A; MYC; ETS1
JAK-STAT signaling pathway	4/162	0.011	0.393	4.803	IFNA1; MYC; IFNE; IL12B
Influenza A	4/172	0.014	0.393	4.515	IL33; IFNA1; KPNA6; IL12B
Bladder cancer	2/41	0.020	0.436	9.604	CDKN2A; MYC
Cell cycle	3/124	0.029	0.479	4.668	CDKN2B; CDKN2A; MYC
Human T-cell leukemia virus 1 infection	4/219	0.031	0.479	3.520	CDKN2B; CDKN2A; MYC; ETS1

Functional enrichment analysis revealed that the identified genes are significantly involved in various biological pathways, particularly those related to the immune response, cancer, and cell cycle regulation. In particular, immune-related pathways such as the RIG-I-like receptor signaling pathway, the JAK-STAT signaling pathway, Influenza A, and human T cell leukemia virus 1 infection were enriched, indicating the potential role of these genes in immune signaling mechanisms. In addition, several cancer-associated pathways, including cellular senescence, bladder cancer, and the cell cycle, were also enriched, suggesting the relevance of these genes in tumorigenesis and cellular proliferation. Among the results, the bladder cancer pathway showed the highest odds ratio (9.604), indicating strong enrichment. Key genes such as CDKN2A, MYC, and CDKN2B appeared in multiple cancer-related pathways, while IFNA1, IL12B, and IFNE were prominent in immune pathways. Although adjusted p-values were relatively moderate (0.393), the biological significance of these enriched pathways supports the relevance of the identified genes and their potential roles in disease development and progression.

4.2. Validation of Gene-Disease Associations via NCBI

To further validate the key genes identified in our study, we used the NCBI database to verify their known associations with diseases and biological functions in Table 19. By cross-referencing our gene list with existing entries and published studies in NCBI, we confirmed the relevance of these genes in various disease pathways and biological processes. This validation strengthens the credibility of our findings and supports the potential roles of the identified genes in disease mechanisms, providing a solid basis for future experimental research.

Table 19. Gene validation using biological databases (NCBI, ProteinAtlas) which associated with Human Cancer

Dataset	Gene Name	Gene Description	Associated Cancer/Disease	Reference
DNA CNV	PLCH2	Belongs to the PLC-eta subgroup, catalyzes cleavage of PtdIns(4,5)P_2 to produce second messengers inositol 1,4,5-trisphosphate and diacylglycerol [80].	Head and Neck Squamous Cell Carcinoma, Gallbladder Cancer	[81]
-	NPPA-AS1	Predicted to bind mRNA and regulate gene expression .	-	-
	MYC	Proto-oncogene; regulates cell cycle, apoptosis, and transformation. Frequently amplified in many cancers [80].	Breast, Bladder, lung, colon, lymphoma, leukemia	[82]
	CDKN2A	Produces multiple transcripts via alternative splicing, encoding different proteins [80].	Bladder, Breast	[83, 84]
	CDKN2B	Located near CDKN2A, often mutated or deleted in cancers [80].	Bladder	[85]
	CDKN2B-AS1	Part of CDKN2B-CDKN2A cluster; interacts with PRC1 and PRC2 to suppress gene expression epigenetically [80].	Bladder, breast	[86, 87]
	MTAP	Enzyme essential in polyamine metabolism and salvage pathways of adenine and methionine [80].	Bladder Cancer	[88]
	LAPTM4B	Binds ceramide, enzymes, and phosphatidylinositol bisphosphate [80].	Bladder Cancer	[89]
	SFRP1	Member of SFRP family; contains Wnt-binding domain similar to Frizzled proteins [80].	Bladder, Breast	[90, 91]
	CHMP2B	Component of ESCRT-III complex, involved in receptor sorting and degradation [80].	Parkinson's Disease	[92]
	NDUFS4	Helper protein in mitochondrial Complex I, essential for respiratory chain [80].	Parkinson's Disease	[93]

The analysis of genes extracted from the union and intersection of multiple datasets reveals several genes strongly associated with human cancers. Notably, CDKN2A, CDKN2B, CDKN2B-AS1, MTAP, LAPTM4B, and SFRP1 are consistently linked to bladder cancer, suggesting a potential genetic signature for this cancer type. Additionally, CDKN2A, CDKN2B-AS1, and SFRP1 also show associations with breast cancer, indicating their broader relevance across multiple cancer types. The well-known oncogene MYC is confirmed to be

involved in a range of cancers, including breast, lung, colon, lymphoma, and leukemia, reinforcing its critical role in tumorigenesis. PLCH2 is associated with head and neck squamous cell carcinoma and gallbladder cancer, while NPPA-AS1 has no reported cancer association. Meanwhile, CHMP2B and NDUFS4 are linked to Parkinson's disease, rather than cancer, highlighting their neurological relevance. These findings underscore the importance of specific genes—particularly those clustered around the 9p21 locus—in the diagnosis and potential treatment of bladder and breast cancers.

4.3. Comparison with other studies

Our proposed methods are compared with other studies as a state of the art. The filters one's algorithms included in MIFS, IG, mRMR, and chi-square for all datasets are presented in Table 20, Table 21, Table 22 and Table 23. The results are compared with the proposed method. The performance using MIFS, IG, mRMR, and chi-square doesn't achieve better results than our proposed method. In Table 24, we implemented the related work method in [41] as the state of the art to validate our proposed methods. The comparison using MIFS, CBF, and FCBF algorithms is implemented. These methods don't achieve better results than our proposed methods. Conversely, Table 25, our proposed methods were compared with wrapper algorithms included in GA as a state of the art. The results proved that our proposed methods achieved the best performance. The GA performance results didn't give a better result than our proposed methods.

Table 20. The proposed methods compared with the MIFS method

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
			LR Classifier			
Train Data %	100.000	96.597	77.058	97.845	94.396	65.556
Test Data %	99.875	84.978	75.525	96.989	93.678	60.417
Over-fitting Diff. %	0.125	11.619	1.533	0.856	0.718	5.139
Precision	0.999	0.817	0.620	0.971	0.938	0.619
Recall	0.998	0.782	0.556	0.965	0.928	0.604
F1-Score	0.988	0.788	0.538	0.966	0.932	0.591
No. of Features	10000	9000	300	25	20	20
F-Time (sec)	192.552	173.955	0.377	0.203	0.067	0.299
C-Time (sec)	2.896	25.195	0.037	0.003	0.002	0.009
AUC	1.000	0.954	0.682	0.997	0.988	0.650
Variance	0.000016	0.000416	0.001001	0.000585	0.000694	0.034819
Standard deviation	0.0040	0.0204	0.0316	0.0242	0.0263	0.1867
Accuracy %	99.875	84.978	75.525	96.989	93.678	60.417
			SVM Classifier			
Train Data %	100.000	91.606	75.676	98.421	92.013	71.996
Test Data %	99.750	84.122	72.228	97.523	91.563	71.667
Over-fitting Diff. %	0.250	7.484	3.448	0.898	0.450	0.329
Precision	0.998	0.860	0.472	0.976	0.929	0.722
Recall	0.997	0.756	0.498	0.967	0.895	0.717
F1-Score	0.997	0.775	0.448	0.969	0.906	0.700
No. of Features	10000	9000	300	25	20	20
F-Time (sec)	192.552	173.955	0.203	0.203	0.067	0.299
C-Time (sec)	2.534	75.394	0.138	0.028	0.017	0.091
AUC	1.000	0.949	0.627	0.998	0.976	0.818
Variance	0.000028	0.000668	0.000814	0.000924	0.001014	0.033918
Standard deviation	0.0053	0.0258	0.0285	0.0304	0.0318	0.1842
Accuracy %	99.750	84.122	72.228	97.523	91.563	71.667
			RF Classifier			
Train Data %	100.000	92.962	100.000	100.000	100.000	100.000
Test Data %	99.627	80.623	84.782	96.456	96.140	78.333
Over-fitting Diff. %	0.373	12.339	15.218	3.544	3.860	21.667
Precision	0.998	0.771	0.827	0.972	0.963	0.796
Recall	0.996	0.719	0.748	0.950	0.956	0.783
F1-Score	0.997	0.718	0.773	0.955	0.958	0.762
No. of Features	10000	9000	300	25	20	20
F-Time (sec)	192.552	173.955	0.377	0.203	0.067	0.299
C-Time (sec)	1.252	3.528	0.376	0.148	0.110	1.046
AUC	1.000	0.942	0.876	0.999	0.990	0.871
Variance	0.000036	0.000614	0.002303	0.001473	0.000944	0.035489
Standard deviation	0.0060	0.0248	0.0480	0.0384	0.0307	0.1883
Accuracy %	99.627	80.623	84.782	96.456	96.140	78.333
			Bagg Classifier			
Train Data %	99.847	98.960	99.574	99.696	99.492	98.261

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Test Data %	98.628	78.806	79.239	95.105	95.435	75.833
Over-fitting Diff. %	1.219	20.154	20.335	4.591	4.057	22.428
Precision	0.989	0.733	0.729	0.955	0.957	0.786
Recall	0.985	0.699	0.729	0.940	0.947	0.758
F1-Score	0.987	0.707	0.727	0.939	0.950	0.736
No. of Features	10000	9000	300	25	20	20
F-Time (sec)	192.552	173.955	0.377	0.203	0.067	0.299
C-Time (sec)	7.322	25.309	0.673	0.021	0.022	0.173
AUC	0.999	0.912	0.794	0.995	0.986	0.811
Variance	0.000036	0.000613	0.002002	0.001473	0.000901	0.033333
Standard deviation	0.0060	0.0248	0.0447	0.0384	0.0300	0.1825
Accuracy %	98.628	78.806	79.239	95.105	95.435	75.833

Table 21. The Proposed Methods Compared with the IGF Method

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
			LR Classifier			
Train Data %	100.000	93.115	77.822	97.784	94.170	71.343
Test Data %	99.875	81.310	76.984	97.260	93.674	70.833
Over-fitting Diff. %	0.125	11.805	0.838	0.524	0.496	0.510
Precision	0.999	0.782	0.680	0.973	0.942	0.715
Recall	0.999	0.706	0.576	0.968	0.928	0.708
F1-Score	0.998	0.705	0.566	0.969	0.931	0.677
No. of Features	3576	3315	396	25.000	22.000	12.000
F-Time (sec)	1.182	5.651	0.093	0.032	0.064	0.049
C-Time (sec)	2.121	0.595	0.057	0.0009	0.001	0.010
AUC	1.000	0.951	0.710	0.998	0.989	0.782
Variance	0.000016	0.000576	0.001445	0.000677	0.002176	0.020448
Standard deviation	0.004000	0.024000	0.038013	0.026019	0.046648	0.142997
Accuracy %	99.875	81.310	76.984	97.260	93.674	70.833
			SVM Classifier			
Train Data %	100.000	94.273	75.666	98.269	92.007	73.089
Test Data %	99.750	85.872	72.379	97.530	91.569	72.917
Over-fitting Diff. %	0.250	8.401	3.287	0.739	0.438	0.172
Precision	0.999	0.873	0.434	0.975	0.930	0.737
Recall	0.997	0.780	0.497	0.972	0.895	0.729
F1-Score	0.998	0.801	0.443	0.972	0.904	0.701
No. of Features	3576	3315	396	25.000	22.000	12.000
F-Time (sec)	1.182	5.651	0.093	0.032	0.064	0.049
C-Time (sec)	2.272	3.142	0.204	0.014	0.021	0.055
AUC	1.000	0.969	0.640	0.999	0.979	0.833
Variance	0.000028	0.000486	0.004378	0.000752	0.004502	0.041038
Standard deviation	0.000028	0.000486	0.004378	0.000752	0.004502	0.041038
Accuracy %	99.750	85.872	72.379	97.530	91.569	72.917
			RF Classifier			
Train Data %	100.000	92.558	100.000	100.000	99.982	100.000
Test Data %	99.502	81.139	83.733	96.997	96.140	77.917
Over-fitting Diff. %	0.498	11.419	16.267	3.003	3.842	22.083
Precision	0.997	0.773	0.793	0.973	0.961	0.830
Recall	0.994	0.714	0.726	0.962	0.959	0.779
F1-Score	0.996	0.721	0.734	0.964	0.958	0.762
NO. F	3576	3315	396	25.000	22.000	12.000
F-Time (sec)	1.182	5.651	0.093	0.032	0.064	0.049
C-Time (sec)	0.826	1.584	0.719	0.098	0.118	0.619
AUC	0.999	0.944	0.860	0.999	0.986	0.873
Variance	0.000410	0.000531	0.009057	0.000567	0.002280	0.026598
Standard deviation	0.000410	0.000531	0.009057	0.000567	0.002280	0.026598
ACC %	99.502	81.139	83.733	96.997	96.140	77.917
			Bagg Classifier			
Train Data %	99.940	98.701	99.653	99.696	99.636	98.635

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Test Data %	99.126	79.045	82.297	95.375	95.253	79.167
Over-fitting Diff. %	0.814	19.656	17.356	4.321	4.383	19.468
Precision	0.996	0.741	0.790	0.958	0.957	0.827
Recall	0.990	0.698	0.752	0.950	0.946	0.792
F1-Score	0.992	0.708	0.754	0.949	0.948	0.781
NO. F	3576	3315	396	25.000	22.000	12.000
F-Time (sec)	1.182	5.651	0.093	0.032	0.064	0.049
C-Time (sec)	3.040	1.095	1.079	0.012	0.029	0.135
AUC	0.999	0.911	0.830	0.993	0.987	0.849
Variance	0.000260	0.000553	0.011270	0.001466	0.001788	0.027299
Standard deviation	0.000260	0.000553	0.011270	0.001466	0.001788	0.027299
ACC %	99.126	79.045	82.297	95.375	95.253	79.167

Table 22. The Proposed Methods Compared with the mRMR Method

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
LR Classifier						
Train Data %	100.000	91.819	74.617	95.508	93.085	68.426
Test Data %	99.750	79.699	73.011	95.075	92.620	67.917
Over-fitting Diff. %	0.250	12.120	1.606	0.433	0.465	0.509
Precision	0.999	0.746	0.500	0.950	0.936	0.734
Recall	0.997	0.688	0.515	0.908	0.910	0.679
F1-Score	0.998	0.689	0.479	0.919	0.917	0.632
NO.F	650.000	505.000	145.000	15.000	19.000	4.000
F-Time (sec)	1200.011	2296.409	61.005	3.996	4.181	0.009
C-Time (sec)	0.251	0.686	0.017	0.002	0.002	0.056
AUC	1.000	0.940	0.659	0.995	0.981	0.809
Variance	0.000028	0.000529	0.002502	0.000796	0.003358	0.015451
Standard deviation	0.000028	0.000529	0.002502	0.000796	0.003358	0.015451
ACC %	99.750	79.699	73.011	95.075	92.620	67.917
SVM Classifier						
Train Data %	100.000	92.486	75.661	52.793	89.049	73.333
Test Data %	99.748	83.848	72.379	52.185	88.938	72.870
Over-fitting Diff. %	0.252	8.638	3.282	0.608	0.111	0.463
Precision	0.999	0.845	0.435	0.325	0.915	0.738
Recall	0.997	0.747	0.497	0.463	0.860	0.733
F1-Score	0.998	0.766	0.443	0.363	0.870	0.726
NO. F	650.000	505.000	145.000	15.000	19.000	4.000
F-Time (sec)	1200.011	2296.409	61.005	3.996	4.181	0.009
C-Time (sec)	0.382	3.609	0.142	0.053	0.044	0.048
AUC	1.000	0.961	0.639	0.948	0.945	0.833
Variance	0.000028	0.000559	0.004378	0.002302	0.005405	0.017052
Standard deviation	0.000028	0.000559	0.004378	0.002302	0.005405	0.017052
ACC %	99.748	83.848	72.379	52.185	88.938	72.917
RF Classifier						
Train Data %	100.000	90.959	100.000	100.000	100.000	100.000
Test Data %	99.627	79.935	81.918	97.553	95.604	79.583
Over-fitting Diff. %	0.373	11.024	18.082	2.447	4.396	20.417
Precision	0.998	0.727	0.767	0.981	0.960	0.805
Recall	0.996	0.690	0.703	0.968	0.950	0.796
F1-Score	0.997	0.689	0.709	0.972	0.952	0.793
NO. F	650.000	505.000	145.000	15.000	19.000	4.000
F-Time (sec)	1200.011	2296.409	61.005	3.996	4.181	0.009
C-Time (sec)	0.398	0.534	0.467	0.1000	0.183	0.554
AUC	1.000	0.942	0.833	0.999	0.991	0.833
Variance	0.000036	0.001249	0.011138	0.000561	0.002693	0.017535
Standard deviation	0.000036	0.001249	0.011138	0.000561	0.002693	0.017535
ACC %	99.627	79.935	81.918	97.553	95.604	79.583
Bagg Classifier						
Train Data %	99.961	97.817	99.498	99.545	99.642	98.017

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Test Data %	98.746	77.468	79.369	94.017	93.684	77.083
Over-fitting Diff. %	1.215	20.349	20.129	5.528	5.958	20.934
Precision	0.991	0.731	0.725	0.951	0.943	0.794
Recall	0.984	0.682	0.712	0.936	0.928	0.771
F1-Score	0.986	0.687	0.705	0.937	0.931	0.757
NO. F	650.000	505.000	145.000	15.000	19.000	4.000
F-Time (sec)	1200.011	2296.409	61.005	3.996	4.181	0.009
C-Time (sec)	1.680	1.135	0.561	0.009	0.042	0.117
AUC	0.999	0.910	0.799	0.982	0.980	0.834
Variance	0.000430	0.000937	0.010620	0.002412	0.002369	0.035650
Standard deviation	0.020736	0.030604	0.103029	0.049116	0.048671	0.189011
ACC %	98.746	77.468	79.369	94.017	93.684	77.083

Table 23. The Proposed Methods Compared with the Chi-Square Method

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
			LR Classifier			
Train Data %	82.339	81.411	76.496	98.118	94.334	94.278
Test Data %	69.889	69.100	74.456	97.815	93.674	93.853
Overfitting Diff. %	12.450	12.311	2.040	0.303	0.660	0.425
Precision	0.653	0.652	0.523	0.976	0.943	0.939
Recall	0.616	0.609	0.540	0.974	0.928	0.930
F1-Score	0.619	0.614	0.502	0.973	0.931	0.934
NO. F	7555	5555	398	24.000	21.000	15.000
F-Time (sec)	0.080	0.528	0.016	0.094	0.016	0.007
C-Time (sec)	24.086	24.033	0.094	0.427	0.137	0.211
AUC	0.885	0.882	0.677	0.973	0.990	0.988
Variance	0.001689	0.001423	0.004373	0.000619	0.002176	0.000629
Standard deviation	0.0411	0.0377	0.0661	0.0249	0.0466	0.0251
ACC %	69.889	69.100	74.456	97.815	93.674	93.853
			SVM Classifier			
Train Data %	100.000	79.862	75.661	71.220	91.994	92.013
Test Data %	99.625	70.130	72.228	70.488	91.563	91.739
Over-fitting Diff. %	0.375	9.732	3.433	0.732	0.431	0.274
Precision	0.997	0.592	0.471	0.556	0.929	0.930
Recall	0.995	0.586	0.497	0.653	0.895	0.898
F1-Score	0.996	0.584	0.448	0.565	0.906	0.908
NO. F	7555	5555	398	24.000	21.000	15.000
F-Time (sec)	0.0801	0.528	0.016	0.094	0.016	0.007
C-Time (sec)	2.379	3.050	0.210	0.093	0.016	0.098
AUC	1.000	0.901	0.628	0.653	0.976	0.976
Variance	0.000036	0.000369	0.000814	0.001305	0.001014	0.001099
Standard deviation	0.0060	0.0192	0.0285	0.0361	0.0318	0.0332
ACC %	99.625	70.130	72.228	70.488	91.563	91.739
			RF Classifier			
Train Data %	100.000	86.934	100.000	100.000	100.000	100.000
Test Data %	99.502	68.552	81.087	98.355	96.832	95.789
Over-fitting Diff. %	0.498	18.382	18.913	1.645	3.168	4.211
Precision	0.997	0.585	0.755	0.984	0.973	0.959
Recall	0.995	0.572	0.701	0.981	0.962	0.954
F1-Score	0.996	0.570	0.704	0.982	0.965	0.954
NO. F	7555	5555	398	24.000	21.000	15.000
F-Time (sec)	0.0801	0.528	0.016	0.094	0.016	0.007
C-Time (sec)	1.009	2.817	0.471	0.229	0.104	0.490
AUC	1.000	0.891	0.836	0.998	0.990	0.989
Variance	0.000041	0.000240	0.008783	0.000363	0.001265	0.002178
Standard deviation	0.0064	0.0155	0.0937	0.0191	0.0356	0.0467
ACC %	99.502	68.552	81.087	98.355	96.832	95.789
			Bagg Classifier			
Train Data %	96.979	96.757	99.452	99.696	99.642	99.630

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Test Data %	66.015	65.598	78.713	95.375	95.429	95.789
Over-fitting Diff. %	30.964	31.159	20.739	4.321	4.213	3.841
Precision	0.568	0.592	0.745	0.960	0.957	0.961
Recall	0.565	0.566	0.708	0.943	0.949	0.952
F1-Score	0.559	0.566	0.710	0.946	0.950	0.954
NO. F	7555	5555	398	24.000	21.000	15.000
F-Time (sec)	0.0801	0.528	0.016	0.094	0.016	0.007
C-Time (sec)	35.022	68.571	7.492	0.063	0.183	0.106
AUC	0.853	0.846	0.803	0.992	0.988	0.986
Variance	0.001600	0.001928	0.018132	0.000654	0.001861	0.002942
Standard deviation	0.04000	0.04391	0.13468	0.02557	0.04313	0.05424
ACC %	66.015	65.598	78.713	95.375	95.429	95.789

Table 24. The Proposed Methods are Compared with the MIFS, CBF, and FCBF Methods

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
MIFS						
KNN Classifier						
Train Data %	99.736	82.686	80.879	97.966	94.435	83.935
Test Data %	99.627	76.097	72.479	97.267	92.628	72.083
Over-fitting Diff. %	0.109	6.589	8.400	0.699	1.807	11.852
Precision	0.998	0.745	0.610	0.975	0.927	0.728
Recall	0.996	0.663	0.568	0.969	0.917	0.721
F1-Score	0.997	0.667	0.572	0.969	0.920	0.717
NO. F	10000	9000	300.000	25.000	20.000	20.000
F-Time (sec)	258.902	180.314	2.121	0.351	0.083	0.467
C-Time (sec)	0.008	0.011	0.0001	0.002	0.00002	0.001
AUC	1.000	0.854	0.624	0.963	0.958	0.785
Variance	0.000036	0.000368	0.002344	0.000839	0.001419	0.017766
Standard deviation	0.00600	0.01919	0.04842	0.02896	0.03766	0.13327
ACC %	99.627	76.097	72.479	97.267	92.628	72.083
CBF						
KNN Classifier						
Train Data %	99.867	52.831	81.158	94.171	94.747	83.161
Test Data %	99.748	49.073	72.612	90.953	92.976	72.083
Over-fitting Diff. %	0.119	3.758	8.546	3.218	1.771	11.078
Precision	0.999	0.447	0.612	0.871	0.931	0.735
Recall	0.997	0.402	0.571	0.855	0.920	0.721
F1-Score	0.998	0.369	0.575	0.846	0.924	0.705
NO. F	900.000	750.000	320.000	20.000	17.000	23.000
F-Time (sec)	2.600	1.850	0.255	0.202	0.105	0.020
C-Time (sec)	0.003	0.003	0.002	0.002	0.002	0.003
AUC	1.000	0.669	0.627	0.947	0.961	0.781
Variance	0.000092	0.000490	0.002308	0.002243	0.000953	0.051383
Standard deviation	0.00959	0.02214	0.04804	0.04737	0.03088	0.22667
ACC %	99.748	49.073	72.612	90.953	92.976	72.083
FCBF Classifier						
KNN Classifier						
Train Data %	99.742	81.390	82.657	97.936	95.333	80.560
Test Data %	99.625	76.236	73.270	97.005	95.078	71.667
Over-fitting Diff. %	0.117	5.154	9.387	0.931	0.255	8.893
Precision	0.998	0.721	0.585	0.970	0.953	0.716
Recall	0.996	0.671	0.587	0.967	0.945	0.717
F1-Score	0.997	0.676	0.575	0.966	0.947	0.697
NO. F	400.000	13.000	16.000	14.000	7.000	4.000
F-Time (sec)	1.750	0.800	1.500	0.101	0.006	0.050
C-Time (sec)	0.001	0.007	0.002	0.002	0.002	0.002
AUC	1.000	0.905	0.675	0.961	0.953	0.750
Variance	0.000131	0.001131	0.001767	0.001217	0.000261	0.048420

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Standard deviation	0.000131	0.001131	0.001767	0.001217	0.000261	0.048420
ACC %	99.625	76.236	73.270	97.005	95.078	71.667

Table 25. The Proposed Methods are Compared with the GA Method

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
SVM Classifier						
Train Data %	99.861	92.433	75.882	87.734	94.064	72.778
Test Data %	99.625	84.260	72.884	85.000	93.675	71.667
Over-fitting Diff. %	0.236	8.173	2.998	2.734	0.389	0.236
Precision	0.997	0.857	0.443	0.867	0.480	0.730
Recall	0.995	0.759	0.500	0.848	0.435	0.717
F1-Score	0.996	0.779	0.447	0.837	0.455	0.706
NO. F	6247	5881	299.000	11.000	11.000	17.000
F-Time (sec)	32.000	1.900	1.125	0.200	0.180	0.120
C-Time (sec)	0.006038	0.021496	0.019769	0.067715	0.028850	0.126990
Variance	0.006038	0.021496	0.019769	0.067715	0.028850	0.126990
Standard deviation	0.006038	0.021496	0.019769	0.067715	0.028850	0.126990
ACC %	99.625	84.260	72.884	85.000	93.675	71.667
KNN Classifier						
Train Data %	99.736	80.733	86.742	86.764	95.157	85.185
Test Data %	99.625	74.275	71.293	81.156	93.678	72.917
Over-fitting Diff. %	0.111	6.458	15.449	5.608	1.479	12.268
Precision	0.997	0.697	0.608	0.817	0.469	0.733
Recall	0.995	0.648	0.588	0.804	0.446	0.729
F1-Score	0.996	0.650	0.590	0.797	0.456	0.725
NO. F	6247	5881	299.000	11.000	11.000	17.000
F-Time (sec)	32.000	1.900	1.125	0.200	0.180	0.120
C-Time (sec)	0.024	0.009	0.003	0.001	0.008	0.003
Variance	0.000036	0.001627	0.003576	0.004006	0.000489	0.021316
Standard deviatino	0.006000	0.040345	0.059799	0.063311	0.022113	0.146028
ACC %	99.625	74.275	71.293	81.156	93.678	72.917
XG-Boost Classifier						
Train Data %	100.000	92.723	100.000	90.619	95.081	99.861
Test Data %	98.750	55.250	85.051	50.736	95.079	80.000
Over-fitting Diff. %	1.250	37.473	14.949	39.883	0.002	19.861
Precision	0.993	0.650	0.830	0.054	0.953	0.810
Recall	0.976	0.300	0.761	0.042	0.944	0.800
F1-Score	0.983	0.452	0.780	0.042	0.474	0.791
NO. F	6247	5881	299.000	11.000	11.000	17.000
F-Time (sec)	32.000	1.900	1.125	0.200	0.180	0.120
C-Time (sec)	490.170	1378.432	8.310	1.363	0.543	0.613
Variance	0.000025	0.001652	0.002210	0.002217	0.000475	0.021914
Standard devistion	0.005000	0.040637	0.047025	0.047093	0.021794	0.147999
ACC %	1.000	0.966	0.749	0.964	0.996	0.979

In addition, in Table 26, we compare our proposed methods with LRS using the Naïve Bayes classifier. In Table 27, the methods are compared with RASGD and Lasso-ASGD. Table 28 presents a comparison between our proposed method and the LRS, RRS, and RFE algorithms.

Table 26. The Proposed Methods Compared with the LRS With Naïve Bayes Classifier

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Naïve-Bayes Classifier						
Train Data %	85.785	66.388	84.142	88.707	92.736	81.759
Test Data %	64.107	65.335	79.763	87.703	92.271	79.583
Overfi.Diff.%	21.678	1.053	4.379	1.004	0.465	2.176
Pre	0.568	0.655	0.743	0.857	0.931	0.814
Rec	0.537	0.656	0.677	0.877	0.908	0.796

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
F1-Score	0.527	0.624	0.691	0.839	0.915	0.789
No. F	1486.000	11265.000	487.000	32.000	22.000	43.000
F-Time (sec)	0.267	3.476	0.116	0.008	0.811	0.008
C-Time (sec)	0.052	1.700	0.194	0.004	0.043	0.108
AUC	0.807	0.842	0.793	0.982	0.979	0.866
Var.	0.007429	0.002555	0.004344	0.001546	0.000898	0.015219
Standard deviation	0.0862	0.0505	0.0659	0.0393	0.0299	0.1234
ACC %	64.107	65.335	79.763	87.703	92.271	79.583

Table 27. The Proposed Methods Compared with the RASGD, Lasso ASGD

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
RASGD Algorithm						
Train Data %	100.000	99.249	99.985	99.362	96.270	87.824
Test Data %	99.875	83.299	65.882	96.719	95.429	68.333
Overfit.Diff.%	0.125	15.950	34.103	2.643	0.841	19.491
Pre.	0.999	0.791	0.604	0.965	0.964	0.706
Rec.	0.999	0.779	0.625	0.963	0.940	0.683
F1-Score	0.999	0.780	0.601	0.963	0.950	0.672
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	2.090	29.943	0.077	0.239	0.003	0.008
AUC	0.956	0.890	0.805	0.923	0.921	0.852
Var.	0.000016	0.000243	0.002603	0.000965	0.000768	0.011265
Standardd Deviation	0.0040	0.0156	0.0510	0.0311	0.0277	0.1061
ACC %	99.875	83.299	65.882	96.719	95.429	68.333
Lasso ASGD Algorithm						
Train Data %	100.000	99.276	99.573	99.454	95.567	87.407
Test Data %	99.502	82.852	70.491	96.997	95.075	67.917
Overfi.Diff.%	0.498	16.424	29.082	2.457	0.492	19.490
Pre	0.996	0.782	0.643	0.969	0.961	0.703
Rec.	0.994	0.769	0.665	0.966	0.936	0.679
F1-Score	0.995	0.771	0.643	0.966	0.946	0.668
No. F	1486.000	11265.000	487.000	32.000	22.000	43.000
F-Time (sec)	0.267	3.476	0.116	0.008	0.811	0.008
C-Time (sec)	2.091	43.313	0.072	0.263	0.004	0.028
AUC	0.961	0.895	0.801	0.987	0.932	0.821
Var.	0.000041	0.000216	0.001165	0.001072	0.000540	0.011207
Standard deviation	0.0064	0.0147	0.0341	0.0327	0.0232	0.1058
ACC %	99.502	82.852	70.491	96.996	95.075	67.917
Ridge with SVM Classifier						
Train Data %	88.639	95.560	79.218	98.755	93.556	72.638
Test Data %	86.894	81.520	77.647	96.734	93.327	71.667
Overfit.Diff.%	1.745	14.040	1.571	2.021	0.229	0.971
Pre.	0.948	0.830	0.793	0.967	0.944	0.724
Rec.	0.851	0.829	0.570	0.962	0.917	0.717
F1-Score	0.850	0.831	0.556	0.963	0.926	0.708
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	34.490	1.831	3.746	0.066	0.132	0.046
AUC	1.000	0.962	0.774	0.991	0.996	0.825
Var.	0.000349	0.000285	0.000469	0.000948	0.000666	0.018441
Standard deviation	0.01868	0.01688	0.02165	0.03078	0.02581	0.13577
ACC %	86.894	81.520	77.647	96.734	93.327	71.667
Ridge with LR Classifier						
Train Data %	100.000	96.929	91.196	97.966	92.072	74.306
Test Data %	99.875	86.077	80.419	97.545	91.570	66.667
Overfit.Diff.%	0.125	10.852	10.777	0.421	0.502	7.639
Pre.	0.999	0.842	0.746	0.976	0.917	0.680
Rec.	0.998	0.799	0.730	0.971	0.903	0.667

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
F1-Score	0.999	0.810	0.733	0.972	0.908	0.650
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	0.611	10.223	0.548	0.002	0.004	0.018
AUC	1.000	0.965	0.797	0.998	0.964	0.729
Var.	0.000016	0.000328	0.001918	0.000408	0.000801	0.019676
Standard deviation	0.01265	0.01811	0.04380	0.02020	0.02829	0.14028
ACC %	99.875	86.077	80.419	97.545	91.570	66.667

Table 28. The Proposed Methods Compared with LRS, RRS and RFE Algorithms

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Lasso (LRS) Algorithm with RF Classifier						
Train Data %	100.000	93.355	100.000	100.000	100.000	100.000
Test Data %	98.500	81.138	84.525	97.564	96.140	77.500
Overfit.Diff.%	1.500	12.217	15.475	2.436	3.860	22.500
Pre	0.991	0.775	0.843	0.964	0.962	0.803
Rec	0.978	0.719	0.729	0.963	0.956	0.775
F1-Score	0.983	0.715	0.755	0.960	0.958	0.766
No. F	1486.000	11265.000	487.000	32.000	22.000	43.000
F-Time (sec)	0.267	3.476	0.116	0.008	0.811	0.008
C-Time (sec)	1.672	8.114	1.875	0.653	0.961	0.371
AUC	1.000	0.941	0.855	1.000	0.989	0.839
Var.	0.000444	0.000647	0.003380	0.002322	0.001149	0.012809
Standard deviation	0.02107	0.02545	0.05814	0.04819	0.03389	0.11316
ACC %	98.500	81.138	84.525	97.564	96.140	77.500
Lasso (LRS) Algorithm with LR Classifier						
Train Data %	100.000	97.241	76.382	97.936	94.345	74.722
Test Data %	99.377	85.288	74.856	97.545	93.850	69.167
Overfit.Diff. %	0.623	11.953	1.526	0.391	0.495	5.555
Pre.	0.995	0.824	0.540	0.976	0.945	0.711
Rec.	0.993	0.793	0.544	0.971	0.929	0.692
F1-Score	0.993	0.800	0.506	0.972	0.933	0.672
No. F	1486.000	11265.000	487.000	32.000	22.000	43.000
F-Time (sec)	0.267	3.476	0.116	0.008	0.811	0.008
C-Time (sec)	0.386	15.052	0.206	0.002	0.101	0.0008
AUC	0.999	0.961	0.679	0.998	0.989	0.744
Var.	0.000043	0.000323	0.005194	0.000408	0.002116	0.029398
Standard deviation	0.00656	0.01797	0.07206	0.02020	0.04599	0.17141
ACC %	99.377	85.288	74.856	97.545	93.850	69.167
Lasso (LRS) Algorithm with KNN Classifier						
Train Data %	99.875	81.402	81.158	92.410	94.728	83.935
Test Data %	99.875	74.177	72.612	87.447	92.976	72.083
Overfit.Diff. %	0.0	7.225	8.546	4.963	1.752	11.852
Pre	0.999	0.668	0.612	0.867	0.930	0.728
Rec	0.999	0.636	0.571	0.864	0.921	0.720
F1-Score	0.999	0.633	0.575	0.846	0.924	0.717
No. F	1486.000	11265.000	487.000	32.000	22.000	43.000
F-Time (sec)	0.267	3.476	0.116	0.008	0.811	0.008
C-Time (sec)	9.964	331.211	0.303	0.015	0.010	0.008
AUC	1.000	0.867	0.627	0.015	0.961	0.906
Var.	0.000016	0.000231	0.002308	0.002653	0.000953	0.017766
Standard deviation	0.00400	0.01520	0.04804	0.05151	0.03087	0.13330
ACC %	99.875	74.177	72.612	87.447	92.976	72.083
Lasso (LRS) Algorithm with DT Classifier						
Train Data %	99.154	65.626	86.508	88.494	96.466	85.185
Test Data %	95.250	64.574	75.660	85.015	94.029	68.333
Overfit.Diff. %	3.904	1.052	10.848	3.479	2.437	16.852
Pre.	0.974	0.553	0.693	0.725	0.940	0.692
Rec.	0.976	0.559	0.638	0.745	0.936	0.683

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
F1-Score	0.974	0.551	0.649	0.721	0.935	0.663
No. F	1486.000	11265.000	487.000	32.000	22.000	43.000
F-Time (sec)	0.267	3.476	0.116	0.008	0.811	0.008
C-Time (sec)	17.565	5.686	0.911	0.009	0.022	0.019
AUC	0.983	0.804	0.708	0.934	0.968	0.802
Var.	0.000410	0.000900	0.005609	0.003209	0.001310	0.021682
Standard deviation	0.02025	0.03000	0.07486	0.05661	0.03621	0.14725
ACC %	97.250	64.574	75.660	85.015	94.029	68.333
Ridge (RRS) Algorithm with RF Classifier						
Train Data %	100.000	93.336	100.000	100.000	100.000	100.000
Test Data %	99.753	81.379	84.525	97.564	96.140	77.500
Overfit.Diff.%	0.247	11.957	15.475	2.436	3.860	22.500
Pre.	0.999	0.786	0.843	0.910	0.962	0.803
Rec.	0.998	0.713	0.729	0.911	0.956	0.775
F1-Score	0.998	0.718	0.755	0.908	0.958	0.765
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	2.944	8.033	6.925	0.589	1.412	1.840
AUC	1.000	0.943	0.855	1.000	0.989	0.839
Var.	0.000088	0.000674	0.003380	0.001914	0.001149	0.012809
Standard deviation	0.00938	0.02596	0.05814	0.04375	0.03389	0.11317
ACC %	99.753	81.379	84.525	97.564	96.140	77.500
Ridge (RRS) Algorithm with LR Classifier						
Train Data %	100.000	96.929	91.196	97.966	92.072	74.306
Test Data %	99.875	86.077	80.419	97.545	91.570	66.667
Overfit.Diff.%	0.125	10.852	10.777	0.421	0.502	7.639
Pre.	0.999	0.842	0.746	0.976	0.917	0.680
Rec.	0.998	0.799	0.730	0.971	0.903	0.667
F1-Score	0.999	0.810	0.733	0.972	0.908	0.650
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	0.611	10.223	0.548	0.002	0.004	0.018
AUC	1.000	0.965	0.797	0.998	0.964	0.729
Var.	0.000016	0.000328	0.001918	0.000408	0.000801	0.019676
Standard deviation	0.0040	0.0181	0.0438	0.0202	0.0283	0.1402
ACC %	99.875	86.077	80.419	97.545	91.570	66.667
Ridge (RRS) Algorithm with KNN Classifier						
Train Data %	99.877	81.402	81.158	92.698	94.728	83.935
Test Data %	99.875	74.177	72.612	87.137	92.976	72.083
Overfit.Diff. %	0.002	7.225	8.546	5.561	1.752	11.852
Precision	0.999	0.668	0.612	0.846	0.930	0.727
Recall	0.999	0.636	0.571	0.847	0.921	0.721
F1-Score	0.999	0.633	0.575	0.830	0.924	0.717
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	5.679	18.533	0.480	0.011	0.013	0.014
AUC	1.000	0.867	0.627	1.000	0.961	0.785
Var.	0.000046	0.000450	0.002308	0.01152	0.000952	0.017766
Standard deviation	0.00678	0.02121	0.04804	0.10735	0.03086	0.13326
ACC %	99.875	74.177	72.612	87.137	92.977	72.083
Ridge (RRS) Algorithm with DT Classifier						
Train Data %	99.153	65.626	86.508	88.494	96.466	85.185
Test Data %	97.250	64.574	75.660	85.015	94.029	68.333
Overfit.Diff.%	1.903	1.052	10.848	3.479	2.437	16.852
Precision	0.974	0.552	0.693	0.725	0.939	0.692
Recall	0.976	0.559	0.638	0.745	0.936	0.683
F1-Score	0.974	0.551	0.649	0.721	0.935	0.663
No. F	2195.000	7983.000	574.000	28.000	28.000	46.000
F-Time (sec)	0.356	1.941	0.019	0.009	0.008	0.003
C-Time (sec)	17.109	12.661	0.782	0.006	0.052	0.047
AUC	0.983	0.804	0.708	0.970	0.968	0.710
Var.	0.000410	0.000810	0.005609	0.003209	0.001310	0.021682

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
Standard deviation	0.02025	0.02846	0.07492	0.05663	0.03620	0.14726
ACC %	97.250	64.574	75.660	85.015	94.029	68.333
RFE based on Gradient Boosting						
Train Data %	100.000	80.500	100.000	92.563	99.805	100.000
Test Data %	89.250	73.256	78.696	90.991	96.485	70.417
Overfit.Diff.%	10.750	7.244	21.304	1.572	3.320	29.583
Pre.	0.982	0.823	0.723	0.774	0.964	0.712
Rec.	0.984	0.826	0.706	0.807	0.961	0.704
F1-Score	0.981	0.822	0.709	0.785	0.962	0.699
No. F	10265	8190.000	376.000	17.000	15.000	23.000
F-Time (sec)	190000	182295	144.783	0.062	0.142	0.314
C-Time (sec)	25250	20500	12.951	0.870	0.648	0.310
AUC	0.975	0.758	0.705	0.825	0.996	0.722
Var.	0.022258	0.002345	0.002090	0.000805	0.000684	0.011748
Standard deviation	0.14919	0.04843	0.04572	0.02837	0.02615	0.10837
ACC %	89.250	73.256	78.696	90.991	96.485	70.417

Moreover, Table 29 summarizes the performance of the embedded and filter algorithms used to evaluate our proposed framework against other hybrid models that integrate both filtering and embedding techniques. The following hybrid combinations were implemented for comparison included in mRMR–RFS, Chi-square-LRS, IG-RRS. Our proposed methods outperformed the hybrid filter-embedded algorithms in all evaluated metrics.

Table 29. The comparison of the proposed methods with hybrid models combining filter and embedded algorithmss

Datasets	**RNA Gene**	**DNA CNV**	**Parkinson's1**	**Dermatology**	**BreastEW**	**Parkinson's2**
mRMR-RFS						
Train Data %	99.972	86.199	99.515	98.518	80.206	89.085
Test Data %	99.127	82.751	77.647	74.583	77.057	88.990
Overfit.Diff.%	0.846	3.447	21.868	23.935	3.149	0.095
Pre	0.425	0.812	0.711	0.523	0.728	0.913
Rec	0.325	0.742	0.687	0.485	0.705	0.862
F1-Score	0.450	0.755	0.694	0.498	0.691	0.872
No. F	223	285	72	11	6	17
F-Time (sec)	0.839	0.127	0.448	0.115	0.139	0.099
C-Time (sec)	0.185	0.133	0.303	0.0414	0.169	0.017
AUC	0.999	0.951	0.792	0.806	0.958	0.947
Var.	0.000584	0.021728	0.005887	0.017149	0.003908	0.005181
Standard deviation	0.02417	0.14742	0.07674	0.13095	0.06253	0.07199
ACC %	99.127	82.751	77.647	74.583	77.057	88.990
Chi-square-LRS						
Train Data %	99.931	59.983	83.347	80.462	89.739	94.806
Test Data %	98.503	58.367	82.405	79.583	89.632	94.552
Overfit.Diff.%	1.428	1.616	0.943	0.879	0.107	0.254
Pre	0.989	0.433	0.798	0.820	0.877	0.950
Rec	0.982	0.342	0.700	0.796	0.843	0.937
F1-Score	0.985	0.344	0.722	0.76	0.833	0.941
No. F	152	38	20	16	18	8
F-Time (sec)	0.033	0.142	0.120	0.003	0.005	0.003
C-Time (sec)	0.473	0.821	0.068	0.028	0.035	0.005
AUC	0.998	0.818	0.680	0.785	0.985	0.995
Var.	0.000981	0.025677	0.069225	0.120201	0.039887	0.027980
Standard deviation	0.03133	0.03133	0.263107	0.34670	0.19972	0.16728
ACC %	98.503	58.367	82.405	79.583	89.632	94.552
IG-RRS						
Train Data %	99.376	90.890	80.688	78.148	78.840	94.942
Test Data %	99.375	85.494	79.107	77.916	78.446	94.731
Overfit.Diff.%	0.001	5.395	1.580	0.231	0.395	0.212
Pre	0.994	0.705	0.757	0.391	0.655	0.949
Rec	0.993	0.643	0.661	0.400	0.674	0.940
F1-Score	0.993	0.658	0.673	0.392	0.639	0.943
No. F	8	758	40	20	5	12

Datasets	RNA Gene	DNA CNV	Parkinson's1	Dermatology	BreastEW	Parkinson's2
F-Time (sec)	0.001	0.193	0.002	0.0009	0.0009	0.0009
C-Time (sec)	0.031	0.563	0.027	0.005	0.009	0.009
AUC	0.999	0.963	0.743	0.951	0.943	0.995
Var.	0.008839	0.013382	0.05885	0.012154	0.047460	0.02476
Standard deviation	0.09401	0.11566	0.24258	0.11023	0.21787	0.15734
ACC %	99.375	85.494	79.107	77.916	78.446	94.731

In Table 30, we use Keras, we implemented our proposed deep learning model and compared its performance to other methods in terms of accuracy, precision, recall, and F1-score. The results demonstrate that our approach outperforms the baseline models, which is attributed to the advanced feature extraction and optimization strategies employed within the Keras framework. While Keras is certainly a powerful tool, we found that the proposed method, designed with a focus on task-specific optimizations and advanced feature selection, provided a more efficient and accurate solution for the problem we were addressing.

We utilized the Keras deep learning framework to analyze the RNA gene dataset. In our experimental comparisons, the proposed method consistently outperformed the Keras-based models during both the training and testing phases. Additionally, the Keras approach required a significantly longer runtime compared to our method. Table 30 presents a detailed comparison between our proposed methods and the Keras models in terms of performance and execution time.

Table 30. Comparison of the proposed method and Keras-based method across evaluation metrics

Metrics	Proposed Method	Keras-based Method
Accuracy	100.000%	99.120%
Precision	1.000	0.973
Recall	1.000	1.000
F1-score	1.000	0.986
No. Features	277.000	30.000
Time (s)	0.218	1,746,469,523.30
AUC	1.000	0.998

To further evaluate the robustness of our proposed methods, we conducted validation using an external diabetic disease dataset. The results, presented in Table 31, show that our methods outperformed existing approaches, highlighting their superior accuracy and reliability. To improve model performance, we employed the early stopping technique during training

Table 31. Performance Comparison of Proposed Feature Selection Algorithms on an External Diabetic Dataset with Early Stopping

Metrics	All features before our proposed methods	RFS Algorithm	LRS Algorithm	RRS Algorithm
Train Data (%)	75.190	99.742	99.000	99.707
Test Data (%)	74.370	94.969	94.589	94.560
Over-Fitting Difference	0.820	4.773	4.411	5.147
Precision	0.750	0.950	0.946	0.946
Recall	0.746	0.950	0.947	0.946
F1-score	0.743	0.950	0.946	0.946
Number of Features	20	16	12	13
Fitting Time (s)	0.450	0.385	0.210	0.185
Classification Time (s)	0.055	0.320	0.217	0.218
AUC	0.818	0.990	0.987	0.988
Variance	0.037436	0.000058	0.000295	0.000228
Standard Deviation	0.1934	0.0076	0.0172	0.0151
Accuracy (%)	74.370	94.969	94.589	94.560

Table 32, presents a summary of the characteristics of six biomedical datasets used in the study, covering aspects such as sample size, class distribution, and preprocessing steps.

Table 32. Summary of datasets characteristics including sample size, class distribution, and preprocessing steps applied prior to model training

Datasets	Sample Size	Class Distribution	Missing Values
RNA gene	801	Class BRCA = 300 Class KIRC = 146	

		Class LUAD = 141 Class PRAD = 136 Class COAD = 78	No
DNA CNV	366	Class 0 = 112 Class 1 = 61 Class 2 = 72 Class 4 = 52 Class 3 = 49 Class 5 = 20	No
Parkinson's Disease 1	756	Class 1 = 564 Class 0 = 192	No
Parkinson's Disease 2	240	Class 0 = 120 Class 1 = 120	No
Dermatology diseases	366	Class 0 = 112 Class 2 = 72 Class 1 = 61 Class 4 = 52 Class 3 = 49 Class 5 = 20	Yes. We applied mode imputation for features to handle the issue of missing values.
BreastEW	569	Class 0 = 357 Class 1 = 212	No

In Table 33,we applied the comparison between the LEDF (RFS, RLS, RRS) and state of art.

Table 33. The comparison between the LEDF (RFS, RLS, RRS) and state of art

NO. F	F1- score	AUC	Var.	ACC	NO. F	F1- score	AUC	Var.	ACC
MIFS					**IGF**				
10000	0.988	1	0.000016	99.875	3576	0.998	1	0.000016	99.875
mRMR					**Chi-square**				
650	0.998	1.000	0.000028	99.750	75550	0.996	1	0.000036	99.625
MIFS, CBF and FCBF					**GA**				
900	0.998	1	0.000092	99.748	6247	0.996	1	0.006038	99.625
LRS-Naïve Bayes					**RASGD**				
22	0.915	0.979	0.000898	92.271	2195	0.999	0.956	0.000016	99.875
Lasso ASGD					**LRS, RRS, RFE (LRS-KNN)**				
1486	0.995	0.961	0.000041	99.502	1486	0.999	1.000	0.000016	99.875
LRS, RRS, RFE (RRS-LR)					**LRS, RRS, RFE (RFE-Gradient boosting)**				
2195	0.999	1.000	0.000016	99.875	15	0.962	0.996	0.000684	96.485
Proposed method (E/IEDF-RFS) for RNA Gene					**Proposed method (E/IEDF-RFS) for Parkinson's Disease2**				
277	1	1	0.0	100	26	0.945	0.973	0.008279	94.583
Proposed method (E/IEDF-RFS) for Dermatology					**Proposed method (E/IEDF-RRS) for BreastEW**				
7	1	1	0.0	100.000	23	0.993	1	0.000341	99.288
Proposed method (EEDF-RLS) for DNA CNV					**Proposed method (EEDF-RRS) for Parkinson's Disease1**				
1049	0.934	0.988	0.000556	94.850	581	0.949	0.992	0.001353	96.426

In fig. 2 the selected features are presented for all datasets using different locations of EDF. We can see that the smallest number of selected features were for BreastEW and Parkinson's disease2 datasets which achieved by EEDF-RFS algorithm. The E/IEDF-RLS achieved the smallest number of features for Parkinson's disease2. On the other side, the E/IEDF-RFS algorithm achieved smallest number

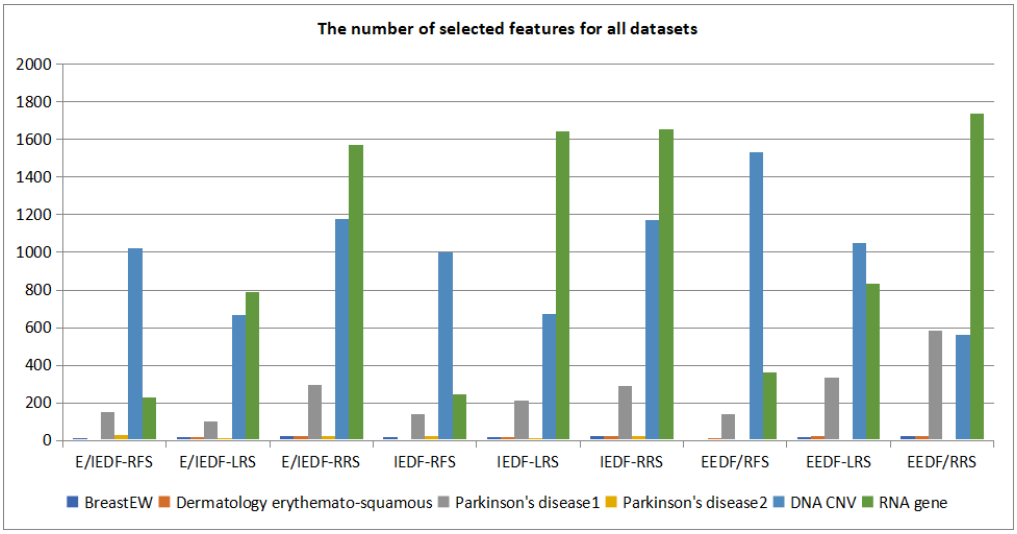

Figure 2. The Number of Selected Features for All Datasets using Different EDF Locations

of selected features for Dermatology erythemato-squamous and RNA gene datasets. In addition, the EEDF-RRS achieved the smallest number of features for DNA CNV dataset.

In fig. 3, many proposed methods gave the best variance. All proposed methods gave zero variance for RNA gene dataset using LR classifier. On the other hand, the EEDF/RFS algorithm achieved the best variance for RNA gene dataset using SVM and RF classifiers. The E/IEDF-RFS algorithm achieved the best variance for Dermatology erythemato-squamous dataset using RF and Bagg classifiers, while the E/IEDF-RLS, E/IEDF-RRS, E/IEDF-RRS, EEDF-RLS and EEDF/RRS algorithms achieved the best variance for Dermatology erythemato-squamous dataset using all classifiers.

The IEDF-RFS achieved the best variance with RNA gene SVM. The IEDF-RLS achieved the best variance for Dermatology erythemato-squamous using SVM, RF and Bagg classifiers, while E/IEDF-RFS algorithm achieved the best results for the same dataset using RF and Bagg classifiers.

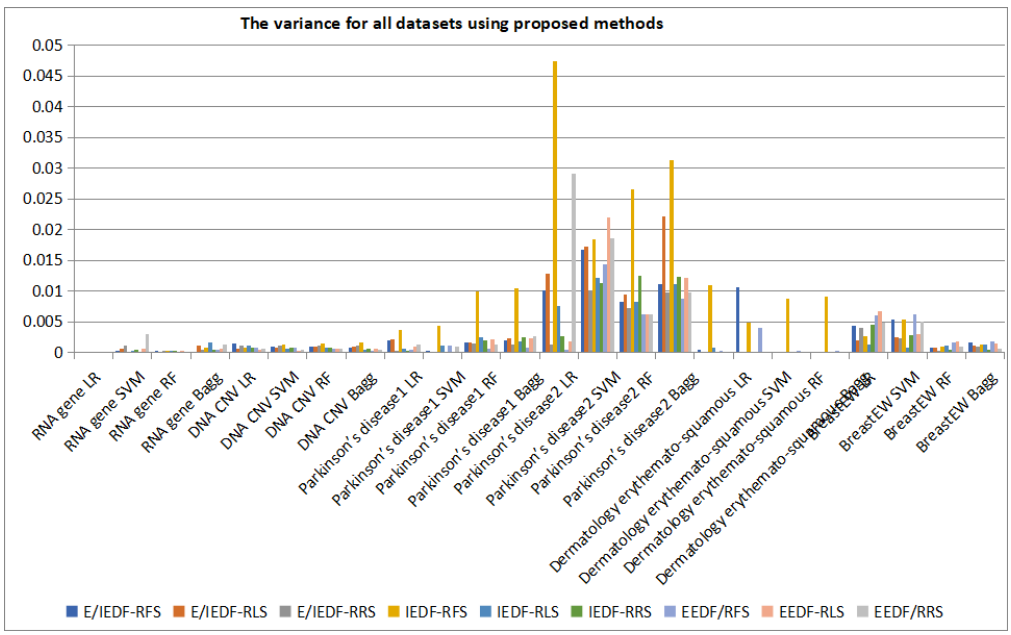

Figure 3. The Variance for All Datasets using Different EDF Locations

In fig. 4, the accuracy results were presented and showed the best one. The E/IEDF-RFS algorithm achieved the best results for RNA gene LR, Dermatology erythemato-squamous RF and Bagg classifiers. In addition, the E/IEDF-LRS, E/IEDF-RRS, IEDF-RRS, EEDF-LRS and EEDF-RRS algorithms gave the superior results for RNA gene LR and Dermatology erythemato-squamous dataset using all classifiers.

The IEDF-RFS algorithm obtained the best results for RNA gene LR and SVM classifiers. Furthermore, IEDF-LRS algorithm achieved the best results for RNA gene LR and Dermatology erythemato-squamous SVM, RF and Bagg classifiers. The EEDF-RFS algorithm obtained the best results for RNA gene LR, SVM and RF classifiers.

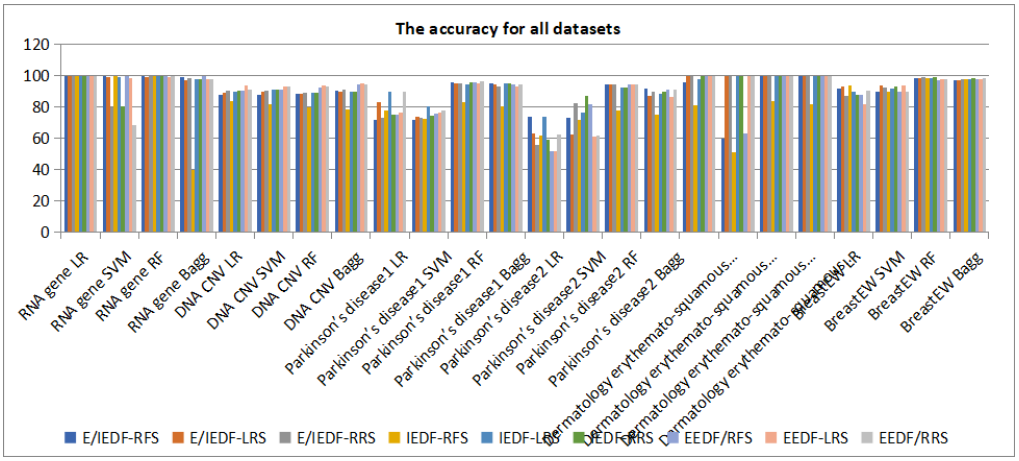

Figure 4. The Accuracy for All Datasets using Different EDF Locations

In Fig. 5, we applied sensitivity analyses, which provided valuable insights into the behavior of our algorithms under different settings. The results demonstrated that certain parameter configurations can significantly enhance model performance. Overall, the sensitivity analyses highlight the importance of careful parameter tuning to achieve optimal outcomes and confirm the reliability and adaptability of our proposed approaches across diverse datasets.

We apply F1 score vs. alpha values using for RNA Gene dataset using different values of alpha are shown as follow:-
alpha-values = [0.01, 0.001, 0.003, 0.0001]
F1-scores = [1.0000, 0.998, 0.960, 0.997]

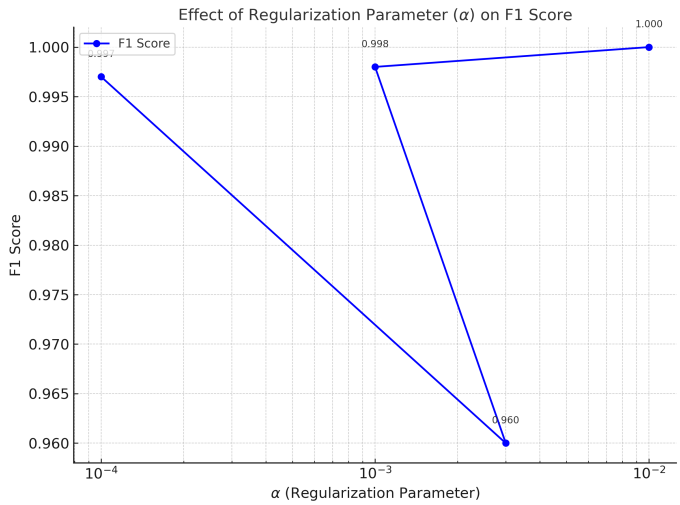

Figure 5. F1 Score vs. Alpha Values plot showing the performance of a model at different regularization strengths

In fig. 6, We add heatmap for RNA gene dataset with 227 features. The figure is shown below with 80 features. We employed heatmap visualizations to highlight the importance of selected features across different classes. These heatmaps provide an intuitive overview of the contribution of each feature to the model's predictions, facilitating the interpretation of feature relevance.

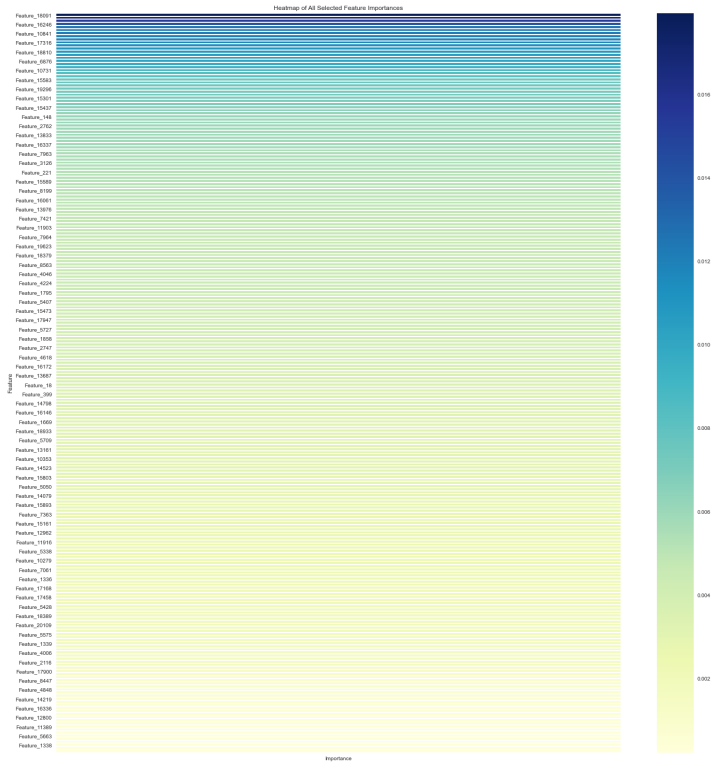

Figure 6. Heatmap showing the importance of selected features from the RNA dataset

In fig.7, we use a SHAP (SHapley Additive explanations) to explain model decisions. It Provides detailed insights into how each feature influences the model predictions by computing Shapley values, which fairly attribute the contribution among the features. We applied SHAP to a Ridge regression model (RRS) trained on the dataset. This helped identify the most impactful features and understand the direction and magnitude of their effects on predictions, enhancing the interpretability of the otherwise linear model.

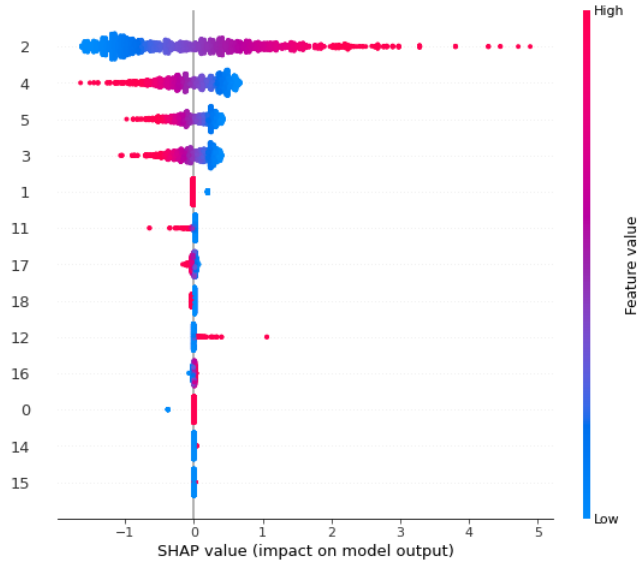

Figure 7. F1 Score vs. Alpha Values plot showing the performance of a model at different regularization strengths

5. Discussion

LEDF, which stands for Location and Embedded algorithm-based Feature Selection, was proposed as a solution for the limitations faced in the FS (Feature Selection) process. It addresses these challenges by utilizing an equation-based bootstrapping resampling method. The LEDF approach is implemented in different locations for three embedded methods: LRS, RRS, and RFS, resulting in a total of nine proposed algorithms.

In the case of LRS and RRS, the bootstrapping equation is applied before the FS process, after the fitting process, and in both locations. On the other hand, RFS uses the EDF bootstrap resampling equation during the training samples, where all data is randomly utilized in the training process in both locations.

Among these algorithms, RFS achieved the most favorable outcomes when employing the E/IEDF method for RNA, Parkinson's disease 2, and dermatology datasets. Specifically, the LR classifier demonstrated the best accuracy results for RNA, achieving 100% accuracy with a variance of 0.0 and selecting relevant features. For Parkinson's disease2, the RF classifier achieved the highest classification accuracy of 94.583%, with 26 features and a variance of 0.008279. In the case of the dermatology datasets, both the Bagg and RF classifiers achieved a perfect accuracy of 100% with 7 features and a variance of 0.0. For the BreastEW dataset, the IEDF method yielded the best results, with a classification accuracy of 98.421%, 16 features, and a variance of 0.000984. Additionally, the EEDF method achieved the best results for the DNA CNV and Parkinson's disease 1 datasets, with classification accuracies of 94.477% (1535 features, variance of 0.000128) and 96.030% (138 features, variance of 0.000615), respectively.

In summary, LEDF is a set of proposed methods for embedded algorithms within LRS, RRS, and RFS. The methods address the limitations of the FS process by employing a bootstrapping resampling equation. Using the E/IEDF method, RFS demonstrated the best performance for RNA, Parkinson's disease2, and dermatology datasets. The LR classifier achieved 100% accuracy for RNA, while the RF classifier achieved the highest accuracy for Parkinson's disease2. The Bagg and RF classifiers both achieved perfect accuracy for the dermatology datasets. The IEDF method yielded the best results for the BreastEW dataset, and the EEDF method achieved the best performance for the DNA CNV and Parkinson's disease 1 dataset.

The LRS algorithm demonstrated the highest performance for the RNA and BreastEW datasets when using the E/IEDF method. Specifically, the LR classifier achieved the best results for RNA, while the RF classifier yielded the highest accuracy for the BreastEW dataset. The classification accuracy, number of features, and variance for RNA were 100%, 788 features, and 0.0 variance, respectively. For the BreastEW dataset, the corresponding values were 98.421% accuracy, 19 features, and a variance of 0.000793. Furthermore, the dermatology dataset achieved the best outcomes with the IEDF method, utilizing the RF, SVM, and Bagg classifiers. In contrast, the Bagg classifier produced the best results for the Parkinson's disease1 dataset. The classification accuracy, number of features, and variance for the dermatology dataset were 100%, 20 features, and 0.0 variance, while for the Parkinson's disease1 dataset, they were 95.231% accuracy, 211 features, and a variance of 0.001807. The EEDF method achieved the top performance for the DNA CNV and Parkinson's disease2 datasets when using the Bagg and RF classifiers, respectively. The classification accuracy, number of features, and variance for the DNA CNV dataset were 94.850%, 1049 features and 0.000556 variances. For the Parkinson's disease2 dataset, they were 94.167% accuracy, 9 features, and a variance of 0.006178.

Similarly, the RRS algorithm yielded the best results for the RNA, BreastEW, and dermatology datasets using the E/IEDF method. The LR classifier achieved the highest accuracy for RNA, while the RF classifier performed the best for the BreastEW dataset. The classification accuracy, number of features, and variance for RNA and BreastEW datasets were 100%, 1573, 0.0, 99.288%, 23, and 0.000341, respectively. In the case of the dermatology dataset, all classifiers achieved perfect accuracy of 100%, with 25 features and 0.0 variance. Additionally, the EEDF method produced the best results for the DNA CNV, Parkinson's disease1, and Parkinson's disease2 datasets, with the RF classifier achieving the highest accuracy for Parkinson's disease1 and Parkinson's disease2, while the Bagg classifier per formed the best for the DNA CNV dataset. The classification accuracy, number of features, and variance for the DNA CNV dataset were 94.683%, 562 and 0.000425. For the Parkinson's disease1 dataset, the classification accuracy, number of features and variance were 96.426%, 581 and 0.001353, respectively. The same for the Parkinson's disease2 dataset, the results were 94.167%, 9 and 0.006178, respectively.

Based on the testing results, our proposed methods proved to be effective for all the different datasets. The E/IEDF-RFS, I/EEDF-RLS, and E/IEDF-RRS algorithms achieved the best results for RNA, all with a classification accuracy of 100.000%. The E/IEDF-RFS algorithm attained the highest accuracy of 94.583% for Parkinson's disease2. For the dermatology dataset, the E/IEDF-RFS, IEDF-RLS, and E/IEDF-RRS algorithms all yielded perfect classification accuracy of 100.000%, while the E/IEDF-RRS algorithm achieved the best accuracy of 99.288% for the BreastEW dataset. Furthermore, the EEDF-RLS algorithm achieved the highest classification accuracy of 94.850% for the DNA CNV dataset, and the EEDF RRS algorithm achieved the best accuracy of 96.426% for the Parkinson's disease1 dataset.

In addition to the technical contributions, practical challenges such as data privacy concerns and the interpretability of models for clinical practitioners must be addressed. To overcome these challenges, we propose deployment strategies that include the use of secure, cloud-based APIs, which can facilitate scalable access to the model while preserving data confidentiality. Furthermore, enhancing model transparency through interpretable outputs will support clinical decision-making and promote trust in AI-driven systems. The practical challenges are as follow:-

1. Ensuring Data Privacy and Security: One of the main challenges in clinical settings is ensuring patient data remains private and secure. To address this, our model complies with key data protection regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation). This means that patient data will be anonymized, encrypted, and stored in a secure manner, ensurin g that only authorized personnel can access it. Additionally, we ensure that data handling follows best practices to protect patient confidentiality while allowing the model to be used effectively.

2. Making the Model Understandable for Doctors (Explainable AI): To make the model more understandable and trustworthy for doctors, we have incorporated explainable AI techniques. These techniques help provide transparent insights into the model's decision-making process, allowing clinicians to understand how and why specific predictions are made. By offering this level of transparency, we aim to increase the model's acceptance and confidence among healthcare professionals.

3. Deploying Through Secure Cloud-Based APIs: To make our model easy to use in clinical settings without the need for complex installations, we propose deploying it through secure cloud-based APIs. This way, hospitals and clinics can access the model remotely over the internet. The model will run on cloud servers, meaning there is no need for expensive or complicated hardware installations on-site. Data will be securely transmitted between the hospital's system and the cloud, ensuring privacy and security, while the model's results can be used in real-time for clinical decision-making.

6. Conclusions

The research introduced a novel resampling hybrid method called LEDF, which utilizes EDF in different locations. We applied LEDF to diverse datasets, including multi-class datasets consisting of phenotype and genotype information. The EDF was implemented within embedded algorithms such as LRS, RRS, and RFS. LEDF was specifically designed to address various challenges encountered in the FS process. A comparative analysis of the proposed methods against existing state-of-the art approaches in recent studies and the results demonstrated their effectiveness.

In the future, it would be beneficial to implement a new EDF bootstrapping hybrid method with GA, PSO, and evolutionary algorithms to enhance the prediction performance in the FS process. Furthermore, we will apply new predictors. Additionally, incorporating various types of datasets, not limited to healthcare datasets, and employing different classifiers will further expand the scope of analysis.

7. Limitations

Despite the promising results achieved by our proposed methods, several limitations should be acknowledged:-

- Datasets biases:- The data source is limited or from a single source only. In future work, we plan to use datasets from different sources and collect real clinical data from hospitals to validate and enhance the model's performance in practical settings.
- Hyperparameter tunig:- In future work, we plan to apply different systematic hyperparameter tuning techniques to identify the most suitable parameter configurations and improve overall model accuracy and robustness.

Acknowledgements

We'd like to express our gratitude to Dr. Mohamed for his invaluable help and support and to Dr. Ghada for her guidance and encouragement.

Author contribution

To emphasize the focus on cancer, particularly bladder cancer, as well as Parkinson's disease. On the other hand to resolve the issues arising from FS algorithms those impede the accuracy of the forecasting process. EDF equation is applied with embedded algorithms with many locations to fix the previous issues.

Availability of data and materials

All datasets and details can be obtained upon request from the corresponding author.

REFERENCES

1. Abdelwahed, N.M., El-Tawel, G.S., Makhlouf, M.A., *Effective hybrid feature selection using different bootstrap enhances cancers classification performance*, BioData Mining, vol. 15, no. 24, pp. 555, 2022.
2. Atran, K.A., et al., *Deep learning in cancer diagnosis, prognosis and treatment selection*, Genome Med., vol.13, no. 2, pp. 152, 2021.
3. Bi, W.L., et al., *Artificial intelligence in cancer imaging: Clinical challenges and applications*, G CA Cancer J Clin., vol.69, no. 2, pp. 127-157, 2019.
4. Liew, X.Y., Hameed, N., Clos, J., *A review of computer-aided expert systems for breast cancer diagnosis*, Cancers (Basel), vol.13, no. 11, pp. 2764, 2021.
5. Saini, A., Kumar, M., Bhatt, S., Saini, V., Malik, A., *Cancer causes and treatments*, Int J Pharm Sci Res., vol.11, no. 17, pp. 3121–3134, 2020.
6. Saba, T.:, *Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges*, Journal of Infection and Public Health, vol.13, no. 9, pp. 1274–1289, 2020.

7. Hughes, G.L., et al., *Machine learning discriminates a movement disorder in a zebrafish model of parkinson's disease*, Disease Models Mechanisms, vol.13, no. 10, 2020.
8. Lamba, R., Gulati, T., Alharbi, H.F., Jain, A., *A hybrid system for parkinson's disease diagnosis using machine learning techniques*, International Journal of Speech Technology, vol.25, pp. 583–593, 2022.
9. Sakar, C.O., et al.:, *A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform*, Applied Soft Computing, vol.74, pp. 255–263, 2019.
10. Arora, P., Mishra, A., Malhi, A., *N-semble-based method for identifying parkinson's disease genes*, Neural Computing and Applications, vol.74, pp. 74–255263, 2021.
11. Peng, J., Guan, J., Shang, X., *Predicting parkinson's disease genes based on node2vec and autoencoder, based on node2vec and autoencoder*, Front. Genet., vol.10, no. 226, 2019.
12. Nahar, N., et al., *Feature selection-based machine learning to improve prediction of parkinson disease*, In: In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N. (eds.) Brain Informatics, PP. 496–508. Springer, Cham., 2021
13. Cerri, S., Mus, L., Blandini, F., *Parkinson's disease in women and men: What's the difference?*, J Parkinsons Dis., vol.9, no.3, PP. 501–515, 2019.
14. Cort´es, R.L., G´omez, B.B., Est´evez, S.V., Fentes, D.P., N´u˜nez, C.:, *Blood-based protein biomarkers in bladder urothelial tumors*, Journal of Proteomics, vol.247, pp. 104329, 2021.
15. Antoni, S., et al., *Bladder cancer incidence and mortality: A global overview and recent trends*, Eur Urol, vol.71, no.1, pp. 96–108, 2017.
16. Cho, H., et al., *Prediction of the immune phenotypes of bladder cancer patients for precision oncology*, IEEE Open J Eng Med Biol., vol.15, no.3, pp. 47–57, 2022.
17. Mi, H., et al., *Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture*, Cell Reports Medicine, vol.2, no.9, 2021.
18. Bladder Cancer: *Definition and Causes*, https://www.mayoclinic.org/diseases-conditions/bladder-cancer/symptoms-causes/syc20356104 Accessed Visited in 10/10/20223.
19. Burger, M., et al., *Epidemiology and risk factors of urothelial bladder cancer*, Eur Urol, vol.63, no.2, PP. 234–241, 2013.
20. Halaseh, S.A., et al., *A review of the etiology and epidemiology of bladder cancer: All you need to know*, Cureus, vol.14, no.7, 2022.
21. Sung, H., et al., *Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries*, CA Cancer J Clin, vol.71, no.3, PP. 209–249, 2021.
22. Gu, J., Wu, X., *Genetic susceptibility to bladder cancer risk and outcome*, Per Med., vol.8, no.3, PP. 365–374, 2011.
23. Dianatinasab, M., et al., *Dietary patterns and risk of bladder cancer: a systematic review and meta-analysis*, BMC Public Health, vol.22, no.1, PP. 73, 2022.
24. Ho, C.H., et al., *Chronic indwelling urinary catheter increase the risk of bladder cancer, even in patients without spinal cord injury*, Medicine (Baltimore), vol.94, no.43, 2015.
25. Tang, H., et al., *Pioglitazone and bladder cancer risk: a systematic review and meta-analysis*, Cancer Med., vol.7, no.4, PP. 1070-1080, 2018.
26. Choi, J.B., et al., *Estimating the impact of body mass index on bladder cancer risk: Stratification by smoking status*, Sci Rep., vol.8, no.1, PP. 947, 2018.
27. Kurahashi, N., et al., *Passive smoking and lung cancer in japanese non-smoking women: a prospective study*, Int J Cancer, vol.122, no.3, PP. 653–657, 2008.
28. Keimling, M., Behrens, G., Schmid, D., Jochem, C., Leitzmann, M.F., *The association between physical activity and bladder cancer: systematic review and meta-analysis*, Br J Cancer, vol.110, no.7, PP. 1862–1870, 2014.
29. Yamaguchi, N., Tazaki, H., Okubo, T., Toyama, T., *Periodic urine cytology surveillance of bladder tumor incidence in dyestuff workers*, Am J Ind Med., vol.3, no.2, PP. 139–148 , 1982.
30. Chouser, K., Leibovich, B., Bergstralh, E., Blute, M., Zincke, H., *Bladder cancer risk following primary and adjuvant external beam radiation for prostate cancer*, J Urol., vol.174, PP. 107–110, 2005.
31. Salmanpour, M.R., et al., Bergstralh, E., Blute, M., Zincke, H., *Boptimized machine learning methods for prediction of cognitive outcome in parkinson's disease*, Computers in Biology and Medicine, vol.111, PP. 103347, 2019.
32. Abdelwahed, N.M.A., Eltoukhy, M.M., Wahed, M.E., *Computer aided system for breast cancer diagnosis in ultrasound images*, Heal. Env., vol.3, no.3 PP. 71–76, 2015.
33. Zhang, N., Wang, M., Zhang, P., Huang, T., *Classification of cancers based on copy number variation landscapes*, Biochim Biophys Acta (BBA)-General Subjects, vol.1860, no.11 PP. 2750–2755 , 2016.
34. Elsadek, S.F.A., Makhlouf, M.A.A., El-Sayed, B.B.S.T., Mohamed, H.N.E., *Hybrid feature selection using swarm and genetic optimization for dna copy number variation*, International Journal of Engineering Research and Technology, vol.12, no.7 PP. 1110–1116 , 2019.
35. Hegazy, A.h.E., Makhlouf, M.A., El-Tawel, G.h.S., *selection using chaotic salp swarm algorithm for data classification*, Arabian Journal for Science and Engineering, vol.44, no.4 PP. 3801–3816 , 2019.
36. Elsadek, S.F.A., Makhlouf, M.A.A., Aldeen, M.A., *Supervised classification of cancers based on copy number variation*, In: Hassanien A., Tolba M., Shaalan K., Azar A. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018.AISI 2018. Advances in Intelligent Systems and Computing. Springer, Cham., pp. 198–207, 2019.
37. Huljanah, M., Rustam, Z., Utama1, S., Siswantining, T., *Feature selection-based machine learning to improve prediction of Parkinson disease*, In: Feature Selection Using Random Forest Classifier for Predicting Prostate Cancer. In: IOP Conference Series Materials Science and Engineering, 2019.
38. Neelaveni, J., Devasana, M.S.G., *Feature selection-based machine learning to improve prediction of Parkinson disease*, In: Alzheimer Disease Prediction Using Machine Learning Algorithms. 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020.
39. Alam, S., Kwon, G.R., *Alzheimer disease classification using kpca, lda, and multikernel learning svm*, Int. J. Imaging Syst. Technol., vol.27, no.2, PP. 133–143, 2017.

40. Chaising, S., Temdee, P., Prasad, R., *Individual attribute selection using information gain based distance for group classification of elderly people with hypertension*, IEEE Access, vol. 9, PP. 82713–82725, 2021.
41. Kuswanto, N.R.Y.H., Ohwada, H., *Feature selection-based machine learning to improve prediction of parkinson disease. In: Comparison of Feature Selection Methods to Classify Inhibitors in Dud-e Database*, In: 3rd International Neural Network Society Conference on Big Data and Deep Learning, INNS BDDL 2018- Sanur, Bali, 25 Indonesia, 2018.
42. Saxena, R., Sharma, S.K., Gupta, M., Sampada, G.C., *A novel approach for feature selection and classification of diabetes mellitus: machine learning methods*, Comput. Intell. Neurosci., vol.2022, no.1, PP. 1-11, 2022.
43. C ¸alı¸skan A., *Diagnosis of malaria disease by integrating chi-square feature selection algorithm with convolutional neural networks and autoencoder network*, Transactions of the Institute of Measurement and Control, vol.45, no.5, PP. 975–985, 2023.
44. Ullah, I., Mahmoud, Q.H.A., *filter-based feature selection model for anomaly-based intrusion detection systemstional neural networks and autoencoder network*, ieee international conference on big data (big data), Boston, MA, USA, 2017, 2020–62055207620914777, 2017.
45. Njoku, U., et al., *Feature selection-based machine learning to improve prediction of parkinson disease*, In: Impact of Filter Feature Selection on Classification: an Empirical Study. A: International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data. Proceedings of the 24rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP): Co-located with the 24th International Conference on Extending Database Technology and the 24th International Conference on Database Theory (EDBT/ICDT 2022): Regne Unit, March 29, 2022. CEUR-WS.org, 2022.
46. Patel, N.K.M.a., Ashoka, K.a., Park, P. Choonkil, Shanmukha, M.C.c., Muhammadd, A., *Disease categorization with clinical data using optimized bat algorithm and fuzzy value*, Journal of Intelligent Fuzzy Systems, vol.44, no.3, PP. 5467–5479, 2023.
47. Parlak, B., Uysal, A.K., *A novel filter feature selection method for text classification: Extensive feature selector*, Journal of Information Science, vol.49,no.1, PP. 59–78, 2021.
48. Rajab, M., Wang, D., *Practical challenges and recommendations of filter methods for feature selection*, Journal of Information Knowledge Management, vol. 19, no.1. 2020.
49. Mao, Y., Yang, Y., *A wrapper feature subset selection method based on randomized search and multilayer structure*, Biomed Res Int., vol.2019, 2019.
50. Halim, Z., et al., *An effective genetic algorithm-based feature selection method for intrusion detection systems*, Computers Security, vol. 110, 2019.
51. Hegazy, M.A. A.h.E.and Makhlouf, El-Tawel, G.h.S., *Improved salp swarm algorithm for feature selection*, Journal of King Saud University-Computer and Information Sciences, vol.32, no3, PP. 335-344, 2020.
52. Nguyen, T.T., Huang, J.Z., Nguyen, T.T., *Unbiased feature selection in learning random forests for high-dimensional data*, The Scientific World Journal, vol.2015, PP. 471371–471389, 2015.
53. Xu, C., Wang, J., Zhen, T., Cao, Y., Ye, F., *Prediction of prognosis and survival of patients with gastric cancer by a weighted improved random forest model: an application of machine learning in medicine*, Arch Med Sci., vol.18, no5, PP. 1208–1220, 2021.
54. Nitta, G.R., Rao, B.Y., Sravani, T., Ramakrishiah, N., B., *Lasso-based feature selection and na¨ ıve bayes classifier for crime prediction and its type*, Service Oriented Computing and Applications, vol.13, PP. 187–197, 2019.
55. Bose, E., Maganti, S., Bowles, K.H., Brueshoff, B.L., Monsen, K.A., *Machine learning methods for identifying critical data elements in nursing documentation*, Nurs Res., vol.68, no. 1, PP. 65–72, 2019.
56. Wang, K., An, Y., Zhou, J., Long, Y., Chen, X., *A novel multi-level feature selection method for radiomics*, Alexandria Engineering Journal, vol.66, PP. 993–999, 2023.
57. Yu, S.H., et al., *Lasso and bioinformatics analysis in the identification of key genes for prognostic genes of gynecologic cancer*, J. Pers. Med., vol.11, no. 11, pp. 1177, 2021.
58. Sethi, J.K., Mittal, M., *An efficient correlation based adaptive lasso regression method for air quality index prediction*, Earth Science Informatics, vol.14, PP. 1777–1786, 2021.
59. Kumarage, P.M., Yogarajah, B., Ratnarajah, N., *Feature selection-based machine learning to improve prediction of parkinson disease*, In: Efficient Feature Selection for Prediction of Diabetic Using LASSO. 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019.
60. Ranstam, J., Cook, J.A., *Feature selection-based machine learning to improve prediction of parkinson disease*, Lasso regression. British Journal of Surgery, vol.105, no.10, PP. 1348, 2018.
61. Paul, S., Drineas, P, *Feature selection for ridge regression with provable guarantees*, Neural Computation, vol.28, PP. 1–27, 2016.
62. Xu, W., Liu, X., Leng, F., Li, W., *Blood-based multi-tissue gene expression inference with bayesian ridge regression*, Bioinformatics, vol.36, no.12, PP. 3788–3794, 2020.
63. Deepa, N., et al., *An ai based intelligent system for healthcare analysis using ridge adaline stochastic gradient descent classifer*, The Journal of Supercomputing, vol. 77, PP. 1998–2017, 2021.
64. Mufassirin, M.M., Ragel, R.G., *A novel filter-wrapper based feature selection approach for cancer data classification*, In: IEEE International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 2018.
65. Qasem, S.N., Saeed, F., *Hybrid feature selection and ensemble learning methods for gene selection and cancer classification*, International Journal of Advanced Computer Science and Applications, vol.12, no.2, 2021.
66. Huang, D.H., Tsai, C.H., Chueh, H.E., Wei, L.Y., *A hybrid model based on emd-feature selection and random forest method for medical data forecasting*, International Journal of Academic Research in Accounting, Finance and Management Sciences, vol.9, no.4, PP. 241–252, 2019.
67. Ali, M.A.S., et al., *A novel method for survival prediction of hepatocellular carcinoma using feature-selection techniques*, Appl. Sci., vol.12, no.13, PP. 6427, 2022.
68. Singh, Y., Tiwari, M., *A novel hybrid approach for detection of type-2 diabetes in women using lasso regression and artificial neural network*, Journal of Intelligent Systems and Applications (IJISA), vol.14, no.4, PP. 11-20, 2022.
69. Jomthanachai, S., Wong, W.P., Khaw, K.W., *An application of machine learning to logistics performance prediction: An economics attribute-based of collective instance*, Comput. Econ., vol.63, no.2, PP. 741–792, 2023.

70. Al-Rajab, M., Lu, J., Xu, Q., *A framework model using multifilter feature selection to enhance colon cancer classification*, PLoS ONE, vol.16, no.4, 2021.
71. Panda, D., Ray, R., Abdullah, A.A., Dash, S.R., *Predictive systems: Role of feature selection in prediction of heart disease*, Journal of Physics Conference Series, vol. 1372, no.1, PP. 012074, 2019.
72. Muthukrishnan, R., Rohini, R., *Lasso: A feature selection technique in predictive modeling for machine learning*, In: IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016.
73. Dissanayake, K., Johar, M.d., *Comparative study on heart disease prediction using feature selection techniques on classification algorithms*, Applied Computational Intelligence and Soft Computing, vol.2021, no.1, 2021.
74. UCI Machine Learning Repository, *UCI Machine Learning Repository: Data Sets*, http: //archive.ics.uci.edu/ml/index.php. Accessed 30 Apr 2021.
75. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., et al., *The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data*, Cancer Discov., vol.2, PP.401-404, 2012.
76. Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., Sander, C., *Emerging landscape of oncogenic signatures across human cancers*, Nat. Genet., vol.45, PP. 1127–1133, 2013.
77. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal*, Science Signaling, vol.6, no.269, PP.1, 2013.
78. Empirical distribution function, $https://en.wikipedia.org/wiki/Empirical_distribution_function$, Visited on 25/8/2021.
79. Karabayir, I., Goldman, S.M., Pappu, S., Akbilgic, O., *Gradient boosting for parkinson's disease diagnosis from voice recordings*, BMC Med. Inform. Decis. Mak., vol. 20, no. 1, PP. 228, 2020.
80. Gene validation , *https://www.ncbi.nlm.nih.gov/gene/4146*, Visited on 20/05/2025.
81. The Human Protein Atlas, *https://www.proteinatlas.org*, Visited on 15/05/2025.
82. Yu, J., Chen, L., Wang, G., Qian, K., Weng, H., Yang, Z. et al., *RBPMS inhibits bladder cancer metastasis by downregulating MYC pathway through alternative splicing of ANKRD10*, Commun Biol. vol. 8, no. 367, 2025.
83. Cheng,T., Wu, Y., Liu, Z., Yu, Y., Sun, S., Guo, M. et al., *CDKN2A-mediated molecular subtypes characterize the hallmarks of tumor microenvironment and guide precision medicine in triple-negative breast cancer*, Front Immunol.vol. 13, PP. :970950, 2022.
84. Worst,T., Weis,C., Stöhr, R., Bertz, S., Eckstein, M, Otto, W. et al., *CDKN2A as transcriptomic marker for muscle-invasive bladder cancer risk stratification and therapy decision-making*, Sci Rep. vol. 8, no. 1 PP. :14383, 2018.
85. Mo, Q., Li, R., Adeegbe, D.O., Peng, G. & Chan, K.S, *Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy*, Commun Biol. vol. 3, no. 784, 2020.
86. Kattan, SW., Hobani, YH., Shaheen, S., Mokhtar, SH., Hussein, MH., Toraih, EA., et al. , *Association of cyclin-dependent kinase inhibitor 2B antisense RNA 1 gene expression and rs2383207 variant with breast cancer risk and survival*, Cell Mol Biol Lett. vol. 26, no. 1, pp. 14, 2021.
87. Oskuie, AM., Jahankhani, K., Rostamlou, A., Arabi, S.,Razavi, ZS., Mardi, A. et al., *Molecular landscape of LncRNAs in bladder cancer: From drug resistance to novel LncRNA-based therapeutic strategies*, Biomedicine & Pharmacotherapy. vol. 165, pp. 115242, 2023.
88. AL, DS., Mega, AE., Douglass, J., Olszewski, AJ.,ED, GZ., Uzun, A.,et al., *features of patients with MTAP-deleted bladder cancer*, Am J Cancer Res. vol. 13, no. 1, pp. 326-339, 2023.
89. Yin, Y,. Fan, Y., Yu, G., Du, Y., *LAPTM4B promotes the progression of bladder cancer by stimulating cell proliferation and invasion*, Oncol Lett. vol. 22, no. 5, pp. 765, 2021.
90. Wang, X., Wang, H., Bu, R., Fei, X., Zhao, C., Song, Y., *Methylation and aberrant expression of the Wnt antagonist secreted Frizzled-related protein 1 in bladder cancer*, Oncol Lett., vol. 4, no. 2, pp. 334-338, 2012.
91. Clemenceau, A.,acouture, A., Bherer, J., Ouellette, G., Michaud, A., Walsh, EA., et al., *Role of Secreted Frizzled-Related Protein 1 in Early Breast Carcinogenesis and Breast Cancer Aggressiveness*, Cancers (Basel)., vol. 15, no. 8, pp. 2251, 2023.
92. Tanikawa, S., Mori, F., Tanji, K., Kakita, A., Takahashi, H., Wakabayashi, K., et al., *Endosomal sorting related protein CHMP2B is localized in Lewy bodies and glial cytoplasmic inclusions in α-synucleinopathy*, Neurosci Lett., vol. 527, no. 1, pp. 16-21, 2012.
93. Shil, S.K., Kagawa, Y., Umaru, B.A., Hara, FN., Miyazaki, H., Yamamoto, Y., Kobayashi, SH., et al., *Ndufs4 ablation decreases synaptophysin expression in hippocampus*, Sci Rep., vol. 11, pp. 10969, 2021.