International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Scalable Parallel Gene Selection Method Based on Hybrid Bio-Inspired Metaheuristic Algorithms with Shapley Value Analysis

Vijaya Lakshmi Alluri[*1], **Karteeka Pavan Kanadam**[2], **Hymavathi Thottathyl**[3]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: The development of microarray technology has made a significant contribution to the prediction of various cancer types and their subtypes through gene selection. Effective hybrid approaches are currently inadequate for the challenging problem of predicting highly discriminative genes in microarrays. Thus, this study proposes a novel approach to selecting parallel gene based on a hybrid bio inspired feature selection method. Initially, microarray data is augmented using the Synthetic Minority Oversampling Technique (SMOTE) to enhance dataset sizes. Then, the Cooperative based Kernel Shapley Values (CkSV) approach is employed to extract features and determine Shapley values. The Hybrid Genetic Dung beetle Optimization (HGDBO) approach was employed to identify the most valuable features. Also, the process is executed on the Apache Hadoop Distributed File System for storing large datasets and cost effectiveness. In addition, the features are classified using several machine learning methods such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and K-nearest neighbour. As a result, the proposed approach is compared to other machine learning based classification algorithms. Eleven datasets are used to assess the outcomes of the analysis of the proposed method, which is conducted using the Python tool. The results of the simulations show that the proposed strategy outperforms the existing methods. For SVM classifier, dataset 1 has an accuracy (0.97), dataset 2 (0.98), dataset3 (0.968), dataset 4 (0.975), dataset 5 (0.973), dataset 6 (0.979), dataset 7 (0.985), dataset 8(0.972), dataset 9 (0.973), dataset 10(0.980) and dataset 11(0.979), respectively.

Keywords: *Hybrid Genetic Dung beetle Optimization (HGDBO), Synthetic Minority Oversampling Technique (SMOTE), Cooperative based Kernel Shapley Values (CkSV).*

1. Introduction

Currently, the categorization of microarray datasets is known as extensive biological data analysis, which attracts significant interest from academics [1]. Molecular biology, specifically the diagnosis of cancer, depends heavily on the application of microarray technology, which makes it a compelling area of study [2]. In addition, Microarray data encompasses a wide range of genes that display varying degrees of expression within a restricted set of samples [3]. The procedure of gene selection from microarray data is of utmost significance in elucidating biological characteristics [4]. One potent tool for cancer categorization is the use of microarrays to obtain gene expression data [5]. The current study compares the levels of gene expression in malignant tissues to those found in normal tissues in an effort to find genes that exhibit either up or down regulation in cancerous cells [6].

Microarray gene expression data used for cancer classification often entails analyzing a large number of genes from both normal and malignant tissues [7]. The gene selection strategy is commonly preferred by researchers for cancer classification because of its superior performance in comparison to alternative approaches [8]. The aforementioned factor plays a pivotal role in effectively tackling the difficulties presented by the elevated dimensionality, limited sample size, and intrinsic noise present in microarray data [9]. In order to accurately categorize cancer, it is crucial to find the genes that provide the most meaningful data. The procedure of selecting microarrays greatly improves their classification performance [10]. Both deep learning and machine learning models are employed to analyze the chosen feature using meta-heuristic methods that derive inspiration from natural phenomena [11].

The Support Vector Machine (SVM) is widely employed by researchers for cancer classification tasks, primarily due to its superior performance [12]. In certain instances, the utilization of several classifiers for classification is observed [13]. Adaptive neuro-fuzzy inference system (ANFIS) techniques are employed for gene selection in large-scale data processing [14]. The process of selecting features requires a significant amount of time. Barnacles mating optimization is occasionally employed for gene selection, but it presents notable limitations when applied to microarray datasets that encompass a substantial number of genes [15]. Breast and colon cancer classification utilizes data mining approaches that operate on gene expression profiles [16]. Traditional data mining approaches may

[*1] *Research Scholar, Department of Computer Science and Engineering, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, 522510, India.*
[2] *Professor and Head, Department of Computer Applications,*
R.V.R & J.C College of Engineering, Chowdavaram, Andhra Pradesh, 522019, India. Email: kkp@rvrjc.ac.in
[3] *Asst. Professor, Department of Computer Applications, R.V.R & J.C College of Engineering, Chowdavaram, Andhra Pradesh, 522019, India. Email: thottathylhyma@gmail.com*
Corresponding Author Email: Vijayalakshmi9999@gmail.com

encounter difficulties in properly addressing the curse of dimensionality when dealing with high-dimensional data [17].

Various machine learning algorithms, including logistic regression (LR), Ada boost classifier, linear discriminant analysis (LDA), random forest (RF), gradient boosting, and k-nearest neighbor (KNN) classifier, are utilized to process these gene expression data [18]. Numerous studies employ genetic algorithms (GAs) to select genes in the context of cancer detection. It takes a long time to apply GA because there are a lot of genes in the microarray [19]. Each current technique has several drawbacks, necessitating the development of a new algorithm for processing Microarray data [20].

1.1 Motivation

The microarray dataset has a substantial quantity of genes exhibiting diverse expression levels, accompanied by a imperfect quantity of samples. The process of identifying genes using microarray data is an essential component in the analysis of biological features. The utilization of microarray gene expression data has demonstrated its efficacy as a technique for the categorization of cancer. An approach to identifying genes that are up-regulated or down-regulated in cancer cells involves comparing the levels of gene expression between normal and malignant tissues. In the context of cancer classification, microarray gene expression data frequently encompasses an extensive array of genes originating from both benign and malignant entities. Bioinformatics relies heavily on high-dimensional data. Duplicate and redundant attributes add complexity to high-dimensional categorization and may reduce classification accuracy. The major contributions of the first research objective are given below:

- To leverage Apache Hadoop for parallel execution of a hybrid bio-inspired algorithm for optimal gene selection
- To integrate Kernel Shapley value evaluation for assessing gene importance within each candidate subset during the selection process
- To achieve a significant reduction in feature selection time on large gene expression datasets using the parallel computing implementation
- To attain high classification accuracy, the gene subset identified by the parallel hybrid bio-inspired feature selection was used with Kernel Shapley value evaluation.

The organization of the paper is structured as follows: section 2 presents the existing research based on machine learning approaches for parallel gene selection. Section 3 explains the proposed methodology in detail. Section 4 presents the findings and analysis. The conclusion and potential future directions were presented in Section 5.

2. Related works

Some recent works on gene selection using different approaches are described as follows.

To tackle the problem of high-dimensional microarray datasets, Ali et al. [21] introduced a hybrid filter-genetic feature selection method; this method enhances the accuracy of cancer classification. The most relevant features were extracted from the cancer microarray datasets using filter feature selection processes such information gain, information gain ratio, and Chi-squared. To further improve the selected features' potential for cancer classification, a genetic algorithm was employed. Using four carcinogenic microarray datasets, primarily pertaining to breast, lung, brain, and central nervous system cancers, the effectiveness of the suggested approach was evaluated. The Central Nervous System (CNS) cancer dataset was used in the evaluation to obtain 93.81% accuracy, 93.8% recall, precision, and F-measure via random forest (RF) classifier. The class and feature properties of the model had a problem, which led to the wrong feature selection.

Akhavan et al. [22] established an innovative method for selecting genes in microarray data, which involves two distinct phases. In the first stage of this process, which included both healthy and malignant samples, the genes that made up the microarray were used as training samples. Subsequently, the number of genes was reduced through the application of anomaly detection. In order to identify the ultimate functional genes, a targeted genetic algorithm was used for the genes obtained from the previous phase in the second step. Using this method, the experimental findings showed that the gene count on all datasets could be reduced by at least 99%. The enormous amount of genes present on the microarray is the primary cause of the significant time expenditure associated with gene selection using metaheuristic algorithms.

To find the most significant genes, Alomari et al. [23] used a novel hybrid filter-wrapper method. For filtering purposes, this approach employs resilient Minimum Redundancy Maximum Relevancy (rMRMR). One method that summarizes the procedure for finding smaller groupings of genes is the Modified Gray Wolf Optimizer (MGWO). The incorporation of novel optimization operators from the Teoriya Resheniya Izobretatelskikh Zadatch (TRIZ) innovative solution into the existing GWO algorithm was done to enhance the diversity of the population. The proposed technique is evaluated on nine widely used microarray datasets to assess its efficacy. Support Vector Machines (SVMs) were employed to carry out classification tasks and achieved an accuracy of 0.9586. However, it should be noted that SVMs are also faced with the drawback of having a high computational cost when estimating inventive solutions.

It was created by Deng et al. [24] to sort microarray datasets into cancer groups. The method is made up of two steps: extreme gradient boosting (XGBoost) and a multi-objective optimization genetic algorithm (XGBoost-MOGA). XGBoost is used for ensemble-based feature selection to rank the genes in the first step. At this point, it is possible to get rid of genes that aren't linked to the class. This leaves only the most important genes to be picked. The second part of XGBoost-MOGA uses a genetic algorithm with multiple goals to find the best group of genes, focusing on the most important gene group. Advanced feature selection methods, including the XGBoost-MOGA algorithm, conducted thorough testing on thirteen publically accessible microarray expression datasets against two popular learning classifiers. Based on the CNS dataset, the experiment's findings show that XGBoost–MOGA achieves an accuracy of 83.33%. The challenge of overfitting could affect the model.

Azadifar et al. [25] introduced a gene selection methodology for cancer diagnostics that relies on graph theory. The Maximum Clique and Edge Centrality (MCEC) technique was integrated with the provided graph network. Genes are evaluated and ordered using established and efficient social network methods, such as edge centrality and the maximum weighted clique criterion, in supervised and unsupervised modes. The suggested method aims to make the chosen genes more relevant to the target class and less redundant within themselves. This method picks a maximum weighted clique over and over again for each run. The genes that are important are found from the traits that are already present in this maximal clique using edge centrality and gene relevance. The study uses a variety of datasets with various features, including lung cancer, leukemia, SRBCT, prostate tumors, and colon. The results demonstrate categorization accuracy rates of approximately 88.32%, 92.09%, 83.19%, 83.67%, and 87.09% for each dataset. The model exhibits lower performance and suffers from generalizability issues. The overview of issues that arise in the existing system is depicted in Table 1.

Table 1: Existing techniques with their drawbacks

Author name and Reference	**Technique used**	**Performance**	**Disadvantage**
Ali et al. [21]	Hybrid filter-genetic feature selection strategy	Attain 93.81% accuracy, 93.8% recall, precision, and F-measure by RF	The model possesses an issue in connection with class and feature characteristics, resulting in inaccurate feature selection.
Akhavan et al. [22]	Two-phase microarray data gene selection technique	Obtain at least 99% accuracy	Utilizing GA incurs significant time costs, primarily because of the extensive number of genes present in the microarray.
Alomari et al. [23]	MGWO	Attain an accuracy of 0.9586	Possess the computational burden in the estimation of inventive solutions.
Deng et al. [24]	XGBoost-MOGA	Produces accuracy of 83.33% in CNS dataset	Prone to overfitting challenge
Azadifar et al. [25]	SMCEC	Attain accuracy of about 92.09% in the leukemia dataset	The model possesses fewer performance and generalizability issues.

Problem statement: Due to the fact that problems get exponentially more difficult as the number of dimensions increases, experts have used metaheuristic-based optimization methods to solve the gene selection problem a lot. A problem that metaheuristic-based algorithms may encounter is that, when selecting features, some features with little individual influence may be overlooked, even though when combined with other features, they may improve performance as a whole. Nevertheless, pertinent characteristics strongly linked to previously chosen characteristics may be ignored. This means that some feature selection strategies might not always produce the best outcomes. This study attempt aims to create a scalable parallel gene selection method using Kernel Shapley Values for feature importance assessment and Hybrid Bio-Inspired Feature Selection.

3. Proposed methodology

Cancer is widely recognized as a significant threat to human health, ranking as the second most common cause of death worldwide. Despite advancements in detection methods, late-stage diagnosis often proves ineffective in avoiding patient fatalities. Thus, it is imperative to develop a robust framework capable of reliably predicting early-stage cancer diagnoses. Most researchers use the gene selection approach in cancer classification due to its appropriate results compared to other approaches. It is important for dealing with the problems that come up because microarray data has a lot of dimensions, a small sample size, and noise. The current system has certain limitations, such as challenges related to the correlation between class and feature attributes, leading to imprecise feature selection. The microarray is time-consuming and expensive due to the large number of genes it contains. Certain individuals face the computational load and the issue of overfitting, leading to a decrease in overall performance. This paper introduces a scalable parallel gene selection method that utilizes Hybrid Bio-Inspired Feature Selection with Kernel Shapley Values based Feature Importance evaluation to address these issues. The suggested method's block diagram is displayed in Figure 1.

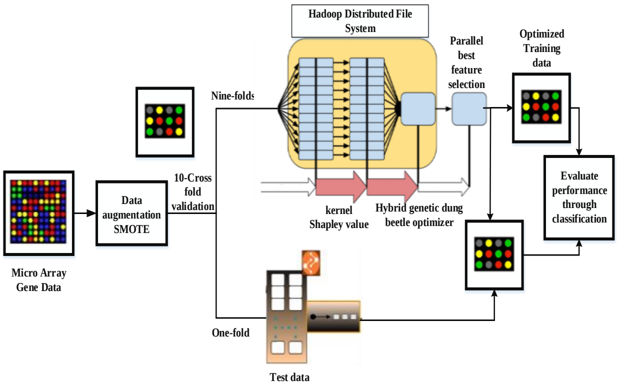

Hadoop Distributed File System
Parallel best feature selection
Optimized Training data
Nine-folds
10-Cross fold validation
Data augmentation SMOTE
Micro Array Gene Data
kernel Shapley value
Hybrid genetic dung beetle optimizer
Evaluate performance through classification
One-fold
Test data

Fig 1: Block diagram of the proposed method

The given dataset will undergo k-fold validation for both training and testing purposes. In the context of 10-fold cross-validation, it is common practice to allocate 9 folds for training purposes and 1 fold for the testing phase. The input data augmentation is initially conducted through the utilization of the Synthetic Minority Oversampling Technique (SMOTE). The "Cooperative based Kernel Shapley Values (CkSV)" technique is used to extract features from the Microarray data. To estimate Shapley values, this study uses a cooperative game-theoretic feature extraction approach. The approach known as "Hybrid Genetic Dungeon Beetle Optimization (HGDBO)" is employed to enhance the quality of the most efficient feature retrieved by CkSV. In order to create parallel processing and shorten computation times, the two processes mentioned above are executed within the "Apache Hadoop distributed file system." This parallel operation concurrently enhances scalability. The classifiers, such as Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and K-nearest neighbors (KNN), utilize the final characteristics for classification. In the testing phase, the performance of the proposed model is assessed using the 1 folded data. Ultimately, the evaluation of the accuracy achieved by each classifier across various datasets is conducted.

3.1 Data augmentation

Data augmentation is employed to enhance the diversity and magnitude of the dataset. Specifically, when the initial dataset is limited or lacks diversity, it is crucial for the development of efficient machine learning models. Models can acquire a broader range of examples by employing data augmentation techniques, thereby enhancing their performance and enabling them to expand their knowledge of previously unseen data. Furthermore, machine learning algorithms possess the capability to successfully train in order to identify connections and generate accurate selections through the efficient enhancement of data. It can eventually contribute to better gene selection and management. This approach that has been proposed analyzes data augmentation methods such as SMOTE.

3.1.1 Synthetic minority oversampling technique (SMOTE)

The conventional oversampling strategy suggested is the SMOTE technique, which is commonly used to address data imbalance problems [26]. SMOTE efficiency surpasses the random oversampling technique in reducing overfitting by including negative data to achieve balanced distributions alongside positive data. The fundamental concept is linear interpolating using the present negative data and neighboring values. The specific stages of SMOTE are as follows:

- Every sample y_i in the minority sample class Y, determines the Euclidean distance amongst the trial in the collection and provides the sample's k nearest neighbor represented as $x_i(i=1,2,....,k)$
- The sampling scale is determined by setting the sampling rate based on the data imbalance ratio. In order to create new data for y_i, n integers are chosen at random by their K-nearest neighbors. This can be done in the following ways:

$$y_{new} = y_j + rand(0,1) \times (x_i - y_j) \qquad (1)$$

Hence $y_i = 1,2,...,n$, and $rand(0,1)$ denotes a arbitrary amount between 0 and 1.

3.2 Apache Hadoop distributed file system

HDFS is a powerful tool employed in cancer classification analysis to organize and analyze extensive gene activity information. The information regarding gene selection can be extensive and is commonly obtained through microarrays. Analysis requires the processing and storage of gigabytes of data. Normal file systems face difficulties when dealing with large datasets. HDFS effectively manages to expand datasets by straightforwardly adding supplementary data points to the system. Hadoop's architecture utilizes distributed storage to analyze data in parallel across the group, resulting in a substantial improvement in analysis performance. In HDFS, microarray data from several cancer patients is stored. HGDBA processes can be utilized to clean, normalize, and feature select the data stored in HDFS. By examining patterns of gene activity, machine learning algorithms like SVM, RF, NB, and KNN have been used to categorize various cancer kinds. The HDFS has the potential to serve as a valuable tool for the categorization of cancer on a wide scale, utilizing gene data. It offers enhanced analytical speed, cost efficiency, and scalability.

3.3 Feature extraction

Feature extraction is a method that finds a specific set of important qualities in order to reduce the number of dimensions in the input. Furthermore, minimizing overfitting and cutting down on processing expenses can greatly increase the effectiveness of machine learning models. The proposed research utilizes a specific method known as cooperative based Kernel Shapley Values (CkSV) to compute KSVs for feature extraction. It addresses several limitations of the fundamental KSV method, such as instability and computational complexity. Cooperative interactions among features are taken into consideration when performing CkSVs, as opposed to focusing just on the inputs of individual features.

3.3.1 Cooperative based Kernel Shapley Values (CkSV)

In a cooperative game, the Shapley value is used to calculate how participants' power is distributed. The application of this technique extends to the process of feature extraction, whereby the relevance of each attribute is assessed [27]. The feature groups are viewed in this technique as essential subsets that may contribute to the ideal subset. To ascertain the relative relevance of traits while simultaneously accounting for complex interactions between components, the Shapley value is employed in a methodical and efficient manner. The Shapley value indicates by $\phi(v)$ and ϕ denotes the coefficients for i^{th} player, where $\phi(v) \in \Re^n$. The Shapley value can be calculated using the following formula:

$$\phi(v) = \sum_{S \subset N} \Delta_i(S) y \frac{|S|!(n-|S|-1)!}{N!} \qquad (2)$$
$$\Delta_i(S) = v(S \cup i) - v(S)$$

In order to make simpler the procedure, the measure of interconnection is represented as $\psi(i, j)$ is determined, and then the evaluation $S \ of \ N$ does not comprise the player i. Thus, N denotes the total number of players.

$$\psi(i, j) = \begin{cases} 1, & t(f_j; class \mid f_i) > t(f_j : class) \\ 0, & else \end{cases} \qquad (3)$$

Where, $f_i \ and \ f_j$ denotes the total amount of feature classes. The function $\Delta_i(S)$ for feature extraction was rewritten using equation (3), which is expressed as follows.

$$\Delta_i(S) = \begin{cases} 1, & t(S, class; f_i) \geq 0 \, and \sum_{f_j \in K} \psi(i, j) \geq \frac{|S|}{2} \\ 0, & else \end{cases} \qquad (4)$$

The challenge is in the evaluation of Shapley values for databases characterized by an average number of features, denoted as n. Due to the need for two assessments of the model, the conventional method of obtaining Shapley values is not feasible for models with a reasonable number of features. Additionally, a KernelSHAP algorithm, which employs a linear regression methodology to calculate the Shapley values, will be utilized to address this issue.

Shapley values of features importance: In the field of machine learning, a commonly employed approach involves the creation of a prediction model $f : F \to \Re$ using a training set S^{train}. Several K-dimensional feature vectors associated with the resulting observed inputs were found inside the training set. The feature area indicates F, which is expressed as a Cartesian product for separate feature areas $F_1 \times F_2 \times \ldots \times F_K$. The probability weight function p is established. The purpose is to identify the relative contributions of every feature to the calculation of a given test data point $y \in F$. In order to quantify the impact of features, it is anticipated that a subset of the test point data will be selected based on their value. The formula for the value coefficient is as follows:

$$v(S)(y) = \sum_{x \in F} p(x)(f(\sigma(y, x, S)) - f(x)) \qquad (5)$$

Where y represents an explanatory variable, x denotes random data, σ indicated distribution function, and f denotes feature space. According to equation (5), the Shapley value regarding the k^{th} feature of the game (N, u) is as follows:

$$\phi_k(y) = \frac{1}{K!} \sum_{p \in K} \sum_{x \in P} p(x) [f(\sigma(y, x, per^k(p) \cup \{k\})) - f(\sigma(y, x, per^k(p)))] \qquad (6)$$

Thus, $per^k(p)$ denotes the set of every feature preceding k^{th} feature within the permutation $p \in \pi(K)$. Then, $\pi(k)$ denotes every possible scenario for k^{th} distinct features and

$v\left(per^{k}\left(p \cup \{k\}\right)\right)$ *and* $v\left(per^{k}\left(p\right)\right)$ indicates that the functions $f(x)$ are mutually exclusive. In order to ascertain an accurate Shapley value, each potential cooperation must be evaluated, regardless of the k^{th} characteristic. Thus, it was challenging to describe $v(s)$ due to the absence of information regarding distribution $p(x)$. In addition, when working with a collection of N features, there are 2^{N} potential associations, which renders the discovery of a precise solution virtually impossible unless one interacts with a minor amount of structures.

Therefore, one may accurately forecast the Shapley values by combining random sampling with an approximation strategy. The relevance of a feature can be assessed by utilizing the estimated Shapley value, as computed in equation (7),

$$\hat{\phi}_{k}(y) = \frac{1}{M}\sum_{m=1}^{M} \left[f\left(\sigma\left(y, x^{m}, per^{k}\left(p^{m}\right) \cup \{k\}\right)\right) - f\left(\sigma\left(y, x^{m}, per^{k}\left(p^{m}\right)\right)\right)\right] \tag{7}$$

Where, every sample denotes M, modified random data indicates x^{m}. The technique uses $\hat{\phi}_{k}(y)$ to calculate the correlation between the predicted value of a particular data point y and the k^{th} feature. The cooperative game-theoretic feature extraction technique is used to extract the significant feature from a very complex gene appearance dataset. This is achieved by employing the kernel Shapley value.

3.4 Feature selection

To identify the most relevant features in a dataset, a data mining technique known as feature selection is employed. This approach can potentially reduce training time, improve accuracy, and mitigate the influence of irrelevant variables on learning models, enhancing their overall performance. Hybrid Genetic Dung Beetle Optimization (HGDBO) is a novel approach that combines two optimization methodologies. One popular evolutionary algorithm that takes elements from natural selection is the Genetic Algorithm (GA). The GA considers each individual within the population as a potential subset of attributes throughout the feature selection process. DBOA is a bio-inspired system that emulates the exploratory behavior of dung beetles. Dung beetles employ a specific navigational strategy while approaching dung mounds. This characteristic is used by DBOA to determine which feature subset in the dataset is best.

3.4.1 Hybrid genetic dung beetle optimization (HGDBO)

The utilization of a parallel genetic algorithm (PGA) in a global optimization process offers the advantage of exhibiting similarities to the genetic evolution observed in cells. The proposed approach is a heuristic search strategy that utilizes defined replicate operations in a stochastic manner to modify the outcomes of functions for binary classified strings. The chromosome is divided into several pieces called genes [28]. It has been demonstrated that PGA is a trustworthy and efficient search technique that requires little knowledge of the particular problem in order to investigate a wide search space. The work aims to conduct classification by imposing constraints on the number of characteristics. Eliminating factors that could cause an erroneous categorization algorithm streamlines the system. This can be done with a genetic algorithm.

Genetic algorithmic techniques are primarily used to computationally address optimization problems, building upon simple evolution and genetics principles. A chromosome consisting of several genes represents a potential solution direction in a PGA. Within the domain of solution, an individual might be conceptualized as a composite of chromosomes. Initially, a group of N chromosomes with a length of L is established. Next, the fitness function for every chromosome in individuals is assessed. The process of selective chromosomal merging results in the formation of new generations through parenting.

Consequently, the fitness function necessitates the calculation of the possibility of selection, specifically the probability of a certain chromosome being chosen as a parent. Selection probabilities can be determined using the PGA technique. Here y_{i} represents the population's i^{th} chromosome and $f\left(y_{i}\right)$represents fitness, and the parallel gene selection p_{s} is as follows:

$$p_{s} = \frac{f\left(y_{i}\right)}{\sum_{i=1}^{N} f\left(y_{i}\right)} \tag{8}$$

During the crossover technique, the creation of additional offspring occurs through the combination of specific parents. A mutation is an erratic, rare alteration in a chromosomal gene that prevents the PGA from convergent toward locally ideal conditions. The process of applying selection, crossover, and mutation in an iterative manner persists until a predetermined maximum number of rounds is attained and the final requirement is met.

The best chromosome with a population is preserved for the following generation in an elitism version of the fundamental PGA, which is provided to enhance its performance. As a result, there is less chance that the chromosome may be lost due to crossover or mutation, and the most effective method can be implemented more quickly. After the last iteration, the PGA generates a chromosome that indicates the best solution, or the feature subset with the highest classification accuracy.

Dung beetle optimization (DBO): One method for swarm intelligence optimization is the dung beetle optimizer (DBO). This phenomena was impacted by beetle activity, which includes rolling balls, dancing, hunting, stealing, and reproducing [29]. Additionally, it possesses a substantial capacity for optimization and exhibits a rapid rate of convergence. It is possible to interpret the rolling dung beetle's position as

$$\begin{cases} y_i(t+1) = y_i(t) + \beta \times g \times y_i(t-1) + b \times \Delta y \\ \Delta y = \left| y_i(t) - Y^w \right| \end{cases} \qquad (9)$$

Where $y_i(t)$ denotes data regarding i^{th} search agents location at t^{th} iteration, t represents the present iteration number, and $g\ and\ b$ represents the constant value of range between (0,1). Then, β denotes the coefficient assigned a value for either -1 or 1, Δx denotes variations in strong light intensity, and Y^w indicates the global worst location. Selecting the appropriate numbers of two parameters is represented as $g\ and\ b$. More specifically, $\beta = 1$ specifies no deviation while $\beta = -1$ designates a change from the starting path.

Exploration stages: In order to proceed, search agents must decide to relocate themselves when they encounter obstacles that impede their progress. After effectively determining an alternative trajectory, they may proceed with the motion of the object. The current location of the beetle has been adjusted and delineated as follows.

$$y_i(t+1) = y_i(t) + \tan(\theta) \left| y_i(t) - y_i(t-1) \right| \qquad (10)$$

Thus, θ denotes a deflection angle within a range from (0, 1). The tangent function represents θ, obtaining a novel rolling location that can replicate the moving behaviour. The process of boundary selection, which replicates the reproductive area where female beetles establish characteristics, is described by

$$\begin{cases} Lb^* = \max\left(Y^* \times (1-R), Lb\right) \\ Ub^* = \min\left(Y^* \times (1+R), Ub\right) \end{cases} \qquad (11)$$

Hence, Y^* characterizes the present best location, $R = 1 - t/T_{\max}\ and\ T_{\max}$ denotes the maximum amount of iteration, $Lb^*\ and\ Ub^*$ indicates lower and upper limits for searching area, R determine the dynamic variations in the reproducing locations boundary ranges, and $Lb\ and\ Ub$ denotes Lower and upper bounds for optimization issues.

Exploitation stages: The boundary ratio for searching location changes gradually, which was described as R. The possibility of oscillation between directions can be reduced when convergence approaches the optimal location. The female beetles choose which of the nearby brood balls to use for their face characteristics after settling on a spot. It is possible to explain the brood ball's location at each iteration stage as

$$B_i(t+1) = Y^* + b_1 \times \left(B_i(t) - Lb^*\right) + b_2 \times \left(B_i(t) - Ub^*\right) \qquad (12)$$

Thus, $B_i(t)$ denotes the location of i^{th} search agents, $b_1\ and\ b_2$ represents two independent arbitrary vectors. Equation (13), which describes the boundaries of the ideal hunting region for little dung beetles,

$$\begin{cases} Lb^b = \max\left(Y^b \times (1-R), Lb\right) \\ Ub^b = \min\left(Y^b \times (1+R), Ub\right) \end{cases} \qquad (13)$$

Where, $Ub^b\ and\ Lb^b$ denotes upper and lower bounds for optimal hunting area and Y^b represents a global optimal location. The smaller dung beetle's location can be changed as follows:

$$y_i(t+1) = y_i(t) + C_1 \times \left(y_i(t) - Lb^b\right) + C_2 \times \left(y_i(t) - Ub^b\right) \qquad (14)$$

Where $C_1\ and\ C_2$ represents a random vector that ranges from 0 to 1 and $y_i(t)$ represents position of the i^{th} little dung beetle in t^{th} repetition. Because they collect the feces balls of other dung beetles, some dung beetles have earned the nickname "thieves." The following equation describes the thief's location:

$$y_i(t+1) = Y^b + H \times q \times \left(\left| y_i(t) - Y^* \right| + \left| y_i(t) - Y^b \right| \right) \qquad (15)$$

Where H represents constant value, q denotes a random vector that is normally distributed, and $y_i(t)$ represents a location for i^{th} thief in t^{th} iteration. To a certain extent, the addition of q can prevent falling to the optimal local solution by increasing random instability. Algorithm 1 depicts the pseudocode of the HGDBO.

```
Algorithm 1: HGDBO
Start
    Create initial population
      While
    $iteration\_number < \max imum number\ of\ iteration$
        For every chromosome,
            Calculate fitness value using
            $p_s = \frac{f(y_i)}{\sum_{i=1}^{N} f(y_i)}$
        End for
            Crossover
            Mutation
        End while
      Output best solution
End
    // Evaluate Dung Beetles optimization
Require: $n$ denotes the total amount of dung beetle, $T_{\max}$ indicates maximum iterations
        For $i \Leftarrow 1 \rightarrow n$ do
            Update the search agent position utilizing equation (9),
          if $B_i > Ub^*$ then
              $B_i \Leftarrow Ub^*$
        end if
          if $B_i < Lb^*$ then
              $B_i \Leftarrow Lb^*$
        end if
      end for
   Optimal solution
Return
```

3.5 Machine learning methods for classification

Machine learning has emerged as a potent technique for classifying cancer by analyzing gene expression data. The proposed study included machine learning classifiers such as SVM, NB, RF, and KNN [30].

3.5.1 Support vector machine (SVM)

The supervised learning technology known as the Support Vector Machine is used to address classification difficulties. The algorithm applies the kernel trick method to transform the input data and identifies the appropriate boundary between positive and negative samples. The additional samples are assigned to the same space and classified into a class based on the direction of the distance they fall from the established maximal boundary. In addition, it tackles both nonlinear and linear classification through the utilization of the kernel trick. By employing a kernel, it is possible to transform low-dimensional inputs into feature spaces of higher dimensions. Particularly, the traditional soft-margin SVM can be expressed as,

$$Svm = \min_{\alpha_i \delta_i} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j x_i x_j K(y_i, y) - \sum_{i=1}^{N} \alpha_i \tag{16}$$

Thus $\sum_{i=1}^{N} \alpha_i x_i = 0 \ and \ 0 \le \alpha_i \le P, \quad for \ i = 1,2,...,L$,and α_i represents the Lagrange multiplier. P indicates penalty factor, $K(y_i, y)$ denotes the spectral kernel, which is shown using radial basis kernels for Gaussian, polynomial, and linear functions below:

$$K(y_i, y) = \begin{cases} \\ y_i^T y \\ (y_i^T y + b), \ b > 0 \\ \exp\left(-\frac{\|y_i - y\|^2}{2\sigma^2} \right), \ \sigma \ne 0 \end{cases} \tag{17}$$

Eventually, the SVM's output was determined by,

$$f(y) = \operatorname{sgn}\left(\sum_{i \notin 0} \alpha_i x_i K(y_i, y) + b \right) \tag{18}$$

Where σ represents the weight parameter, training samples denotes y , sgn represents a signum function, and $K(y_i, y)$ represents a polynomial kernel value that quantifies the relationship among input pattern y_i . b represents the SVM parameter, which is determined at the final stage of the training phase.

3.5.2 Naive Bayes (NB)

Using Naive Bayes Classifiers is the easiest and most effective way to organize. Based on the concept of Bayesian networks, this approach can be conceptualized as a possible graphical model that depicts a set of random elements and their conditional separation. In Bayesian networks, there are several effective ways to receive and process input. Because the features of the applied data are independent, the Naive Bayes approach is used for classification. Every sequence of information initially creates multiple opportunities with this method. When the new data arrives, the total number of possible sequences for each individual is calculated. Consequently, sequences are classified based on the number of probability sequences. With a strong belief in separateness, the Naive Bayes division uses the Bayes theory in a straightforward and practicable manner. Thus, A data can be recorded with a group G_j label in the following ways:

$$NB(G_j \mid A) = \frac{NB(A \mid G_j) * NB(G_j)}{NB(A)} \tag{19}$$

Thus, A represents the total amount of data and G_j denotes a group of labels.

3.5.3 Random forest (RF)

A random forest is made up of predictable trees organized so that each tree in the forest has the same circulation and is dependent on randomly selected, independently tested vector numbers. The highest possible value of the total degradation error is increased as the forest's tree count rises. The overall influence of each tree in the forest and their interconnections determine which tree exhibits the highest frequency of errors. The error rates derived from randomly selecting features to isolate each node exhibit more resilience in the environment. Internal measurement is used to assess the effectiveness, error rate, and ability to fix mistakes within a division. It is applied to show how the division responds as the number of elements rises.

3.5.4 K-nearest neighbours (KNN)

The KNN methodology classifies information by identifying the nearest occurrences in the characteristic space for training purposes. KNN is a machine learning algorithm that is commonly used as a simple and direct method for data isolation. This methodology employs a classification method that gives a category to each concern based on the categorization of the majority of its closest friends. The whole training set is retained throughout the research process. The most straightforward approach for KNN, with K=1, is the Neighborhood rule. The database sample and the surrounding samples should be appropriately segregated.

Therefore, sample segregation can be determined by considering the proximity of neighboring samples when the data is uncertain. In order to determine how far away an unknown object is from each training set sample, one needs only the unknown data and the training set. The training set samples closest to the unidentified group have the same minimum value range. Therefore, a cluster of the closest neighboring data points could be employed to categorize an unidentifiable sample. Ultimately, the accuracy of each classifier is evaluated using different datasets.

4. Result and Discussion

This section explains the suggested model's performance evaluation and outcome analysis. The Python tool is used for the experimental purposes. Eleven datasets are used in the implementation of the proposed parallel selection method. The proposed 90% is utilized for training and 10% for testing.

4.1 Dataset Details

This research examines a few real-world microarray datasets with a remarkably high number of characteristics (genes) in this section. Specifically, the current study considers 11 publicly available datasets. The eleven datasets include Colon tumor, Central nervous system (CNS), Leukemia, Breast cancer, Lung Cancer, Ovarian Cancer, Leukemia_3c, Leukemia_4c, Mixed lineage Leukemia gene (MIL), and small round blue cell tumor (SRBCT). Table 2 provides an overview of 11 datasets [31].

Table 2: Overview of 11 datasets.

Datasets	**Total Genes**	**Initial Data's**	**Total number of data (After augmentation)**	**Total amount of features selected**	**Classes**
Colon tumor	2000	60	1240	1395	2
Central nervous system	7129	60	1250	1125	2
Leukemia	7129	72	1240	1116	2
Breast cancer	24481	97	1250	1389	2
Lung Cancer	12533	181	1421	1279	2
Ovarian Cancer	15154	253	1518	1367	2
Leukemia_3c	7129	72	1512	1361	3
Leukemia_4c	7129	72	1440	1296	2
Lymphoma	4026	62	1452	1307	3
MILL	12582	72	1512	1361	3
SRBCT	2308	83	1494	1345	2

In many prior experiments, the original dataset was randomly divided into two parts: a training set and a test set. The next step is to pick genes from the training set and then evaluate their quality using the unseen test set. However, society today views such an approach as unreliable because of the few cases. In an alternative scenario, the proposed method would use an external 10-fold cross-validation to partition the data. A comparison of different techniques for estimating errors in microarray categorization was proposed. Large-scale gene expression is analyzed via microarray experiments, which produce vast datasets but necessitate careful processing to derive useful information. In general, microarray data experiment findings offer a useful sample of the patterns of gene expression under various circumstances. Table 3 shows the system configuration.

Table 3: System configuration

Processor	Intel® Core(TM) i3-3245 CPU@3.40Ghz 3.40 GHz
Installed memory(RAM)	4.00 GB (3.83 GB usable)
System type	64-bit Operating system
Pen and Touch	No pen and Touch Input is available for this display

4.2 Performance Metrics Evaluation

The performance analysis is assessed to evaluate metrics like accuracy, F1-score, precision, recall, true negative rate (TNR), and true positive rate (TPR). Additionally, an analysis of the outcomes using other existing techniques can be used to estimate the efficiency.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (20)$$

$$\Pr ecision = \frac{T_P}{T_P + F_P} \quad (21)$$

$$\operatorname{Re} call = \frac{T_P}{T_P + F_N} \quad (22)$$

$$F1 - score = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \quad (23)$$

$$Specificity = \frac{T_N}{T_N + F_P} \quad (24)$$

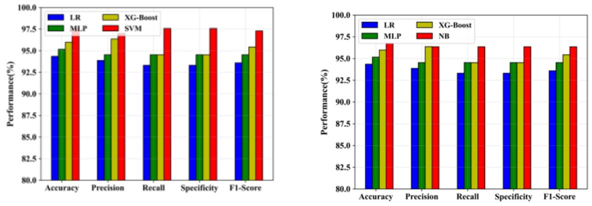
Hence, TP denotes true positive, TN represents true negative, FP indicates false positive, and FN denotes false negative.

4.3 Comparative analysis with other methods

This section offers an analysis of the results and a comparison between the proposed approach and current models using the XG-Boost, Multi-layer perceptron, and logistics regression (LR) techniques.

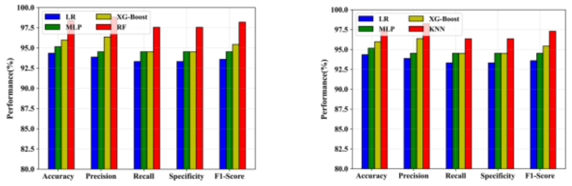
4.3.1 Performance evaluation for Dataset1 (Colon tumor)

The comparison between the proposed and existing approaches for dataset 1 is depicted in Figure 2 (a)-(d). In Figure 2(a), the proposed SVM method succeeds with an accuracy of 0.97, whereas the existing methods are LR (0.94), MLP (0.95), and XG-Boost (0.95). Then, the proposed method achieves a precision (0.97), recall (0.975), specificity (0.97), and F1-score (0.972). The proposed NB approach achieves an accuracy of 0.96 in Figure 2(b). The proposed RF approach achieves an accuracy of 0.98 in Figure 2(c). The proposed KNN approach achieves an accuracy of 0.97 in Figure 2(d). Therefore, the proposed method yields superior results. The values of the proposed and current approaches for dataset 1 are displayed in Table 4.



(a) SVM

(b) NB



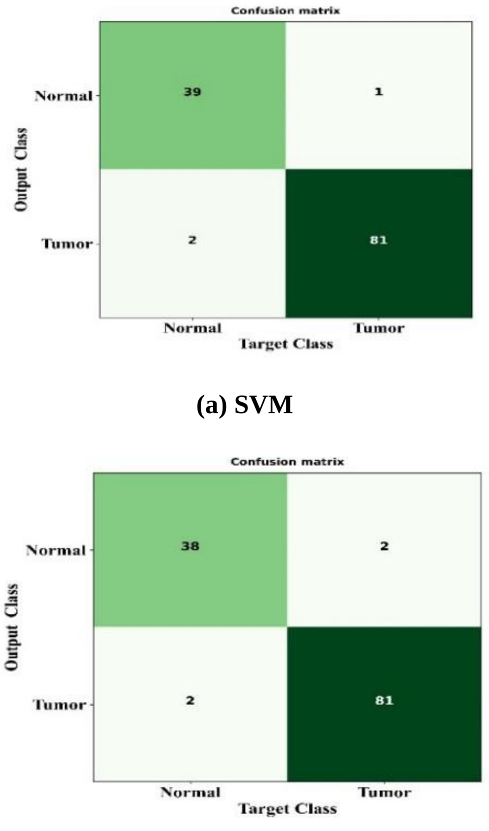
(c) RF

(d) KNN

Fig 2: Performance metrics for Dataset 1

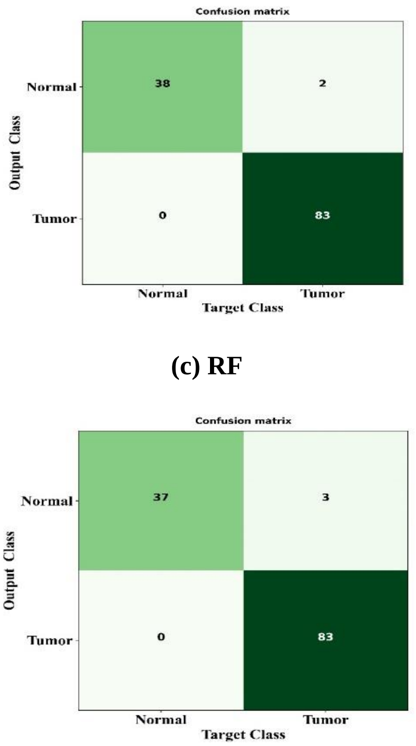
Table 4: Values of proposed and existing methods for dataset1.

Metrics	SVM	NB	RF	KNN	XG-Boost	MLP	LR
Accuracy	0.9758	0.9677	0.9839	0.9758	0.9597	0.9516	0.9435
precision	0.9700	0.9635	0.9882	0.9825	0.9635	0.9453	0.9386
Sensitivity	0.9757	0.9634	0.9756	0.9634	0.9451	0.9450	0.9331
Specificity	0.9756	0.9612	0.9756	0.9634	0.9450	0.9438	0.9320
F1-score	0.9729	0.9610	0.9818	0.9728	0.9543	0.9430	0.9359



(a) SVM

(b) NB



(c) RF

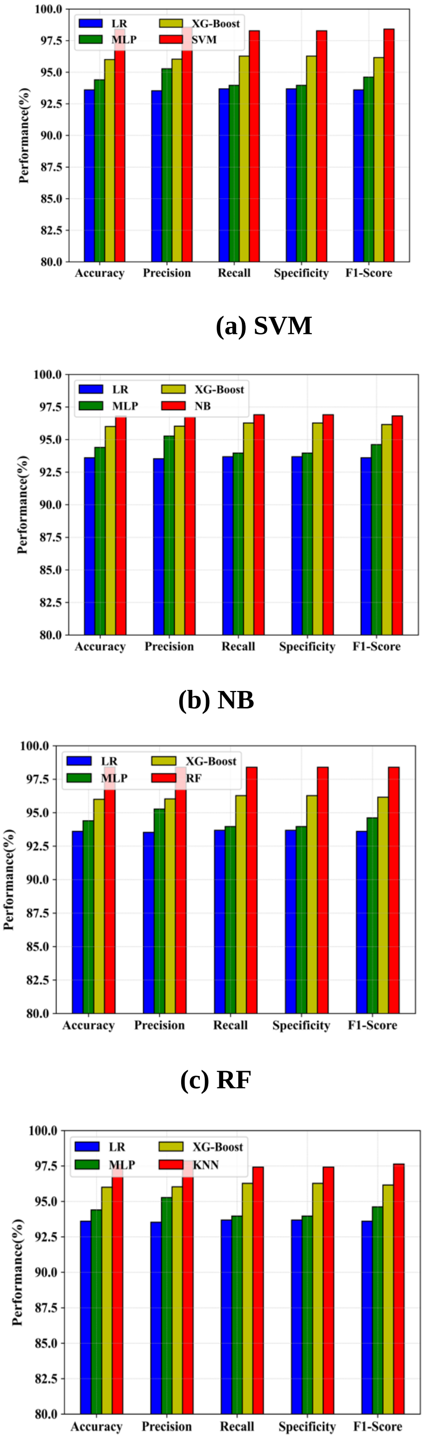
(d) KNN

Fig 3: Confusion matrix for dataset1

Figure 3(a)-(d) shows the confusion matrix attained by the proposed technique for colon tumor classification. The proposed study categorized two classes of colon tumor classification: normal and tumor. In figure 3 (a), when detecting the normal class, the proposed method recognized 39 as normal classes and 2 classified as tumor classes. For classifying the tumor classes, the proposed method recognized 81 as tumor, and one is classified as a normal class. In figure 3 (b), when detecting the normal samples, the proposed method recognized 38 as normal classes and 2 classified as tumor classes. For classifying the tumor classes, the proposed method recognized 81 as tumor, and 2 were classified to be normal class. In figure 3 (c), when detecting the normal samples, the proposed method recognized 38 as normal classes and none classified as tumor classes. For classifying the tumor classes, the proposed method recognized 83 as tumor, and 2 were classified to be normal class. In figure 3 (d), when detecting the normal samples, the proposed method recognized 37 as normal classes and none classified as tumor classes. For classifying the tumor classes, the proposed method recognized 83 as tumor, and 3 were classified to be normal class. Therefore, the proposed approach allows for more accurate gene classification.

4.3.2 Performance evaluation for Dataset 2 (Central nervous system)

The comparison of proposed and current techniques for dataset 2 is displayed in Figure 4(a)–(d). In Figure 4(a), the proposed SVM method obtains an accuracy of 0.98, whereas the existing methods are LR (0.93), MLP (0.94), and XG-Boost (0.96). Then, the proposed method achieves a precision (0.985), recall (0.982), specificity (0.982), and F1-score (0.984). In Figure 4(b), the proposed NB method attains an accuracy of 0.968. In Figure 4(c), the proposed RF method succeeds with an accuracy of 0.98. In Figure 4(d), the proposed KNN method accomplishes an accuracy of 0.976. Thus, the proposed method has enhanced performance and reduced processing time. Table 5 depicts values for dataset 2 for proposed and existing approaches.



(a) SVM

(b) NB

(c) RF

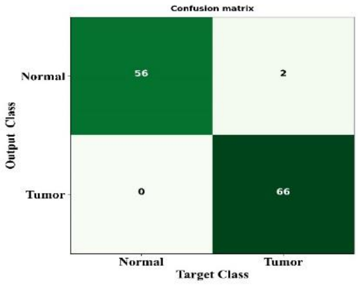
(d) KNN

Fig 4: Performance metrics for Dataset 2

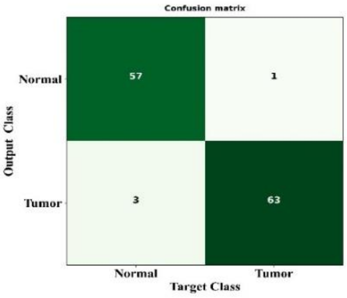
Table 5: Values for dataset 2 for proposed and existing approaches.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.984	0.968	0.984	0.976	0.944	0.96	0.936
precision	0.9855	0.9673	0.9839	0.9785	0.9527	0.9603	0.9352
Sensitivity	0.9827	0.9689	0.9739	0.9741	0.9396	0.9626	0.9368
Specificity	0.9826	0.9680	0.9732	0.9740	0.9390	0.9620	0.9360
F1-score	0.9841	0.9681	0.9631	0.9763	0.9461	0.9615	0.93604

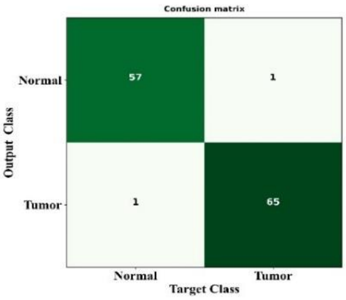
Figure 5(a)-(d) shows the confusion matrix attained by the proposed technique for Central nervous system classification. In figure 5 (a), when detecting the normal class, the proposed method recognized 56 as normal classes and none classified as tumor classes. For classifying the tumor classes, the proposed method recognized 66 as tumor, and 2 were classified to be normal class. In figure 5 (b), when detecting the normal samples, the proposed method recognized 57 as normal classes and 3 classified as tumor classes. For classifying the tumor classes, the proposed method recognized 63 as tumor, and one is classified as a normal class. In figure 5 (c), when detecting the normal samples, the proposed method recognized 57 as normal classes and one classified as tumor classes. For classifying the tumor classes, the proposed method recognized 65 as tumor, and one is classified as a normal class. In figure 5 (d), when detecting the normal samples, the proposed method recognized 55 as normal classes and none classified as tumor classes. For classifying the tumor classes, the proposed method recognized 66 as tumor, and 3 were classified to be normal class.



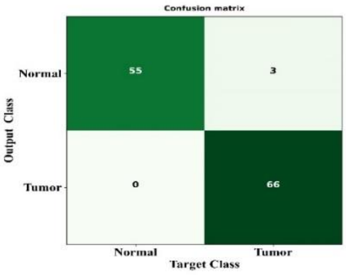
(a) SVM



(b) NB



(c) RF

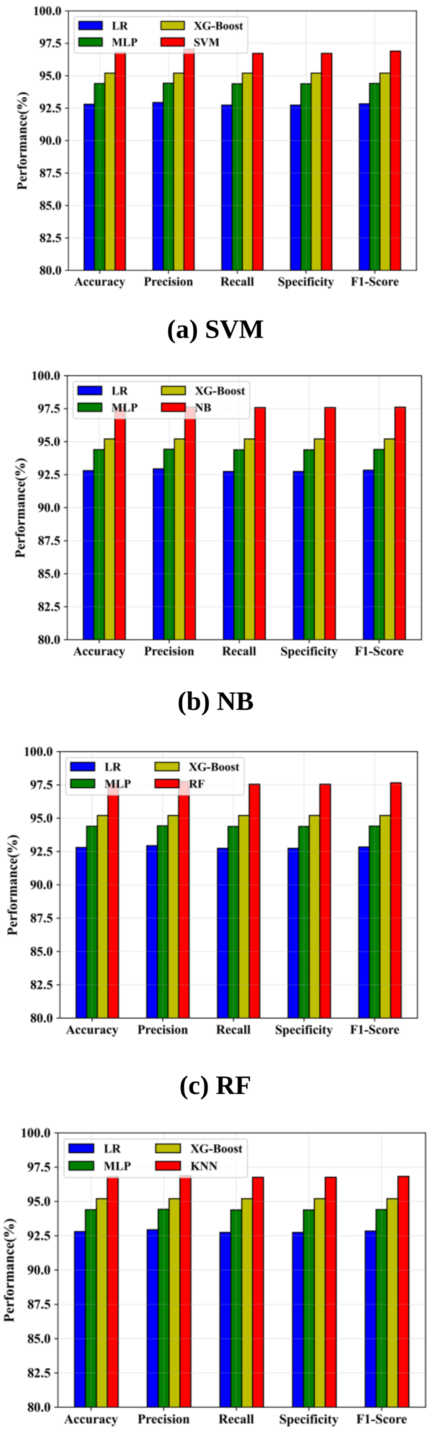


(d) KNN

Fig 5: Confusion matrix for dataset1

4.3.3 Performance evaluation for Dataset3 (Breast cancer)

Figure 6 (a)-(d) compares proposed and existing methods for dataset 3. In Figure 6(a), the proposed SVM method attains an accuracy of 0.96, whereas the existing methods are LR (0.92), MLP (0.94), and XG-Boost (0.95). Then, the proposed method achieves a precision (0.97), recall (0.967), specificity (0.96), and F1-score (0.968). The proposed NB approach obtains an accuracy of 0.976 in Figure 6(b). The proposed RF approach obtains an accuracy of 0.97 in Figure 6(c). The proposed KNN approach obtains an accuracy of 0.968 in Figure 6(d). Therefore, the proposed method was identifying essential gene and faster to select the optimal features. Table 6 shows the values of proposed and existing methods for dataset 3.



(a) SVM

(b) NB

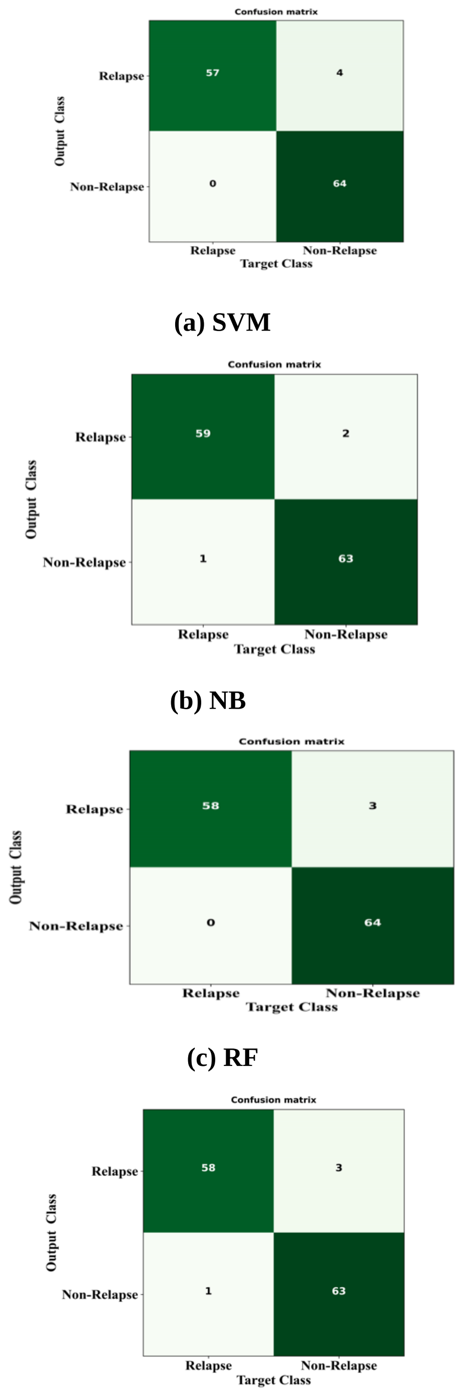
(c) RF

(d) KNN

Fig 6: Performance metrics for Dataset 3

Table 6: Values of proposed and existing methods for dataset 3.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.968	0.976	0.976	0.968	0.944	0.952	0.928
precision	0.9705	0.9762	0.9776	0.9687	0.9442	0.9519	0.9293
Sensitivity	0.9672	0.9757	0.9754	0.9675	0.9437	0.9515	0.9273
Specificity	0.9670	0.9754	0.9752	0.9673	0.9435	0.9510	0.9270
F1-score	0.9688	0.9760	0.9765	0.9681	0.9440	0.9519	0.9283



(a) SVM

(b) NB

(c) RF

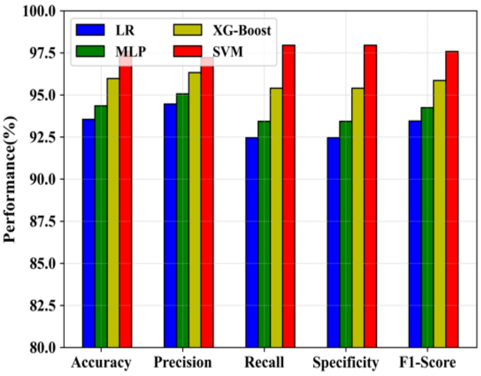
(d) KNN

Fig 7: Confusion matrix for dataset3

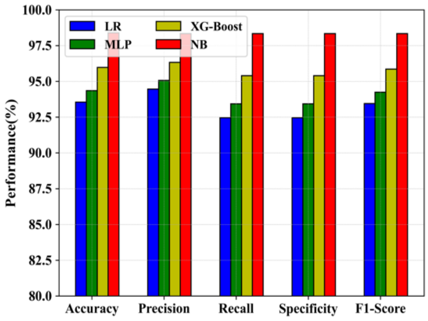
Figure 7(a)-(d) shows the confusion matrix attained by the proposed technique for Breast cancer classification. In Figure 7(a), when selecting the relapse class, the proposed method recognized 57 as relapse classes and none classified as non-relapse classes. For classifying the non-relapse classes, the proposed method recognized 4 as relapse, and 64 were classified as non-relapse classes. In figure 7 (b), when selecting the relapse samples, the proposed method recognized 59 as relapse classes and one classified as non-relapse classes. For classifying the non-relapse classes, the proposed method recognized 63 as non-relapse, and 2 were classified as relapse class. In figure 7 (c), when selecting the relapse samples, the proposed method recognized 58 as relapse classes and none classified as non-relapse classes. For classifying the non-relapse classes, the proposed method recognized 64 as non-relapse, and 3 were classified as relapse classes. In figure 7 (d), when selecting the relapse samples, the proposed method recognized 58 as relapse classes and none classified as non-relapse classes. For classifying the non-relapse classes, the proposed method recognized 63 as non-relapse, and 3 were classified as relapse class.

4.3.4 Performance evaluation for Dataset 4 (Leukemia)

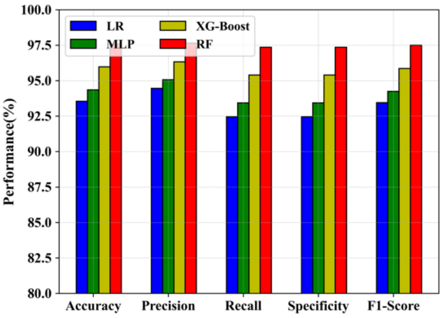
Figure 8 (a)-(d) compares proposed and existing methods for dataset 4. In Figure 8(a), the proposed SVM method accomplishes an accuracy of 0.97, whereas the existing methods are LR (0.93), MLP (0.943), and XG-Boost (0.959). Then, the proposed method achieves a precision (0.972), recall (0.979), specificity (0.97), and F1-score (0.975). Figure 8(b) shows the proposed NB approach with an accuracy of 0.983. Figure 8(c) shows the proposed RF technique with an accuracy of 0.975. Figure 8(d) shows the proposed KNN algorithm with an accuracy of 0.967. So, the proposed method is a more precise classification of genes related to illness and reduced dimensionality. Table 7 depicts values for dataset 4 for proposed and existing approaches.



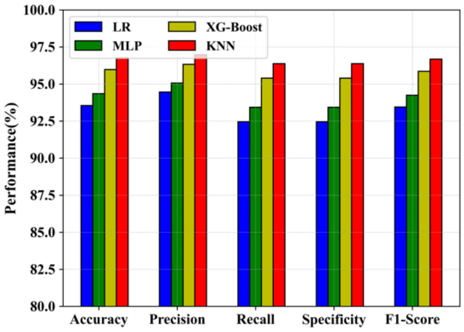
(a) SVM



(b) NB



(c) RF

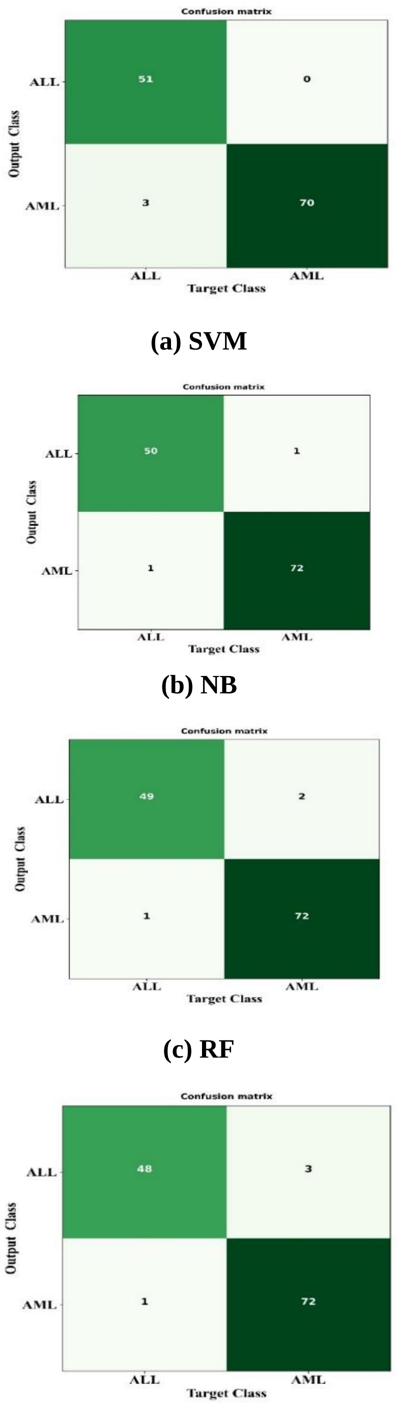


(d) KNN

Figure 8: Performance metrics for Dataset 4

Table 7: Values for dataset 4 for proposed and existing approaches.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9758	0.9839	0.9758	0.9677	0.9435	0.9597	0.9355
precision	0.9722	0.9833	0.9764	0.9697	0.9506	0.9632	0.9445
Sensitivity	0.9794	0.9830	0.9735	0.9637	0.9343	0.9539	0.9245
Specificity	0.9792	0.9828	0.9733	0.9635	0.9340	0.9535	0.9243
F1-score	0.9758	0.9824	0.9750	0.9667	0.9424	0.9585	0.9344



(a) SVM

(b) NB

(c) RF

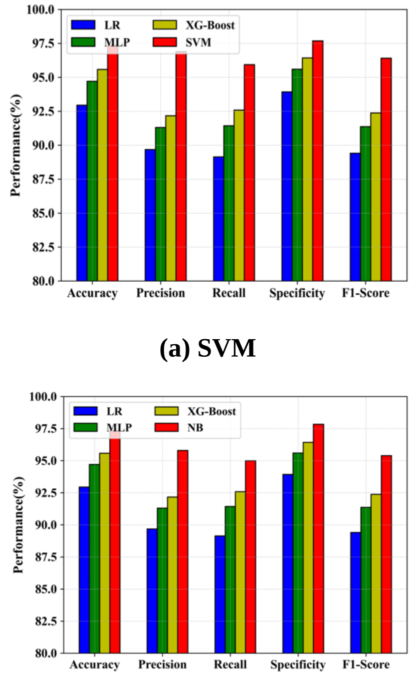
(d) KNN

Figure 9: Confusion matrix for dataset 4

Figure 9(a)-(d) shows the confusion matrix attained by the proposed technique for Leukemia classification. The proposed study categorized two classes of gene prediction: Acute Lymphocytic Leukemia (ALL) and Acute Myeloid Leukemia (AML). In Figure 9(a), when selecting the ALL class, the proposed method recognized 51 as ALL classes and 3 classified as AML classes. For classifying the AML classes, the proposed method recognized 70 as AML, and none were classified as ALL classes. In figure 9 (b), when selecting the ALL samples, the proposed method recognized 50 as ALL classes and one classified as AML classes. For classifying the AML classes, the proposed method recognized 72 as AML, and one is classified as an ALL class. In figure 9 (c), when selecting the ALL samples, the proposed method recognized 49 as ALL classes and one classified as AML classes. For classifying the AML classes, the proposed method recognized 72 as AML, and 2 were classified as an ALL class. In figure 9 (d), when selecting the ALL samples, the proposed method recognized 48 as ALL classes and one classified as AML classes. For classifying the AML classes, the proposed method recognized 72 as AML, and 3 were classified as an ALL class.

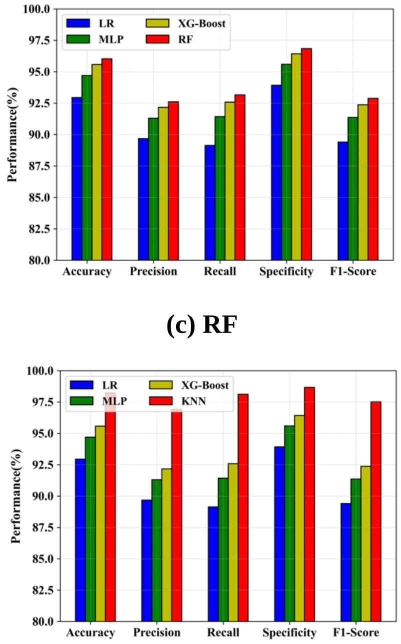
4.3.5 Performance evaluation for Dataset 5 (Leukemia_3c)

Figure 10 (a)–(d) shows how the proposed and current methods for dataset 5 compare to each other. In Figure 10(a), the proposed SVM method accomplishes an accuracy of 0.973, whereas the existing methods are LR (0.92), MLP (0.947), and XG-Boost (0.955). Then, the proposed method achieves a precision (0.96), recall (0.95), specificity (0.976), and F1-score (0.964). The accuracy achieved by the proposed NB approach in Figure 10(b) is 0.97. The accuracy achieved by the proposed RF technique in Figure 10(c) is 0.96. The accuracy achieved by the proposed KNN approach in Figure 10(d) is 0.98. Thus, the proposed method is an effective device that can speed up gene discovery and increase precision and effectiveness. Table 8 shows the values of proposed and existing methods for dataset 5.



(a) SVM

(b) NB



(c) RF

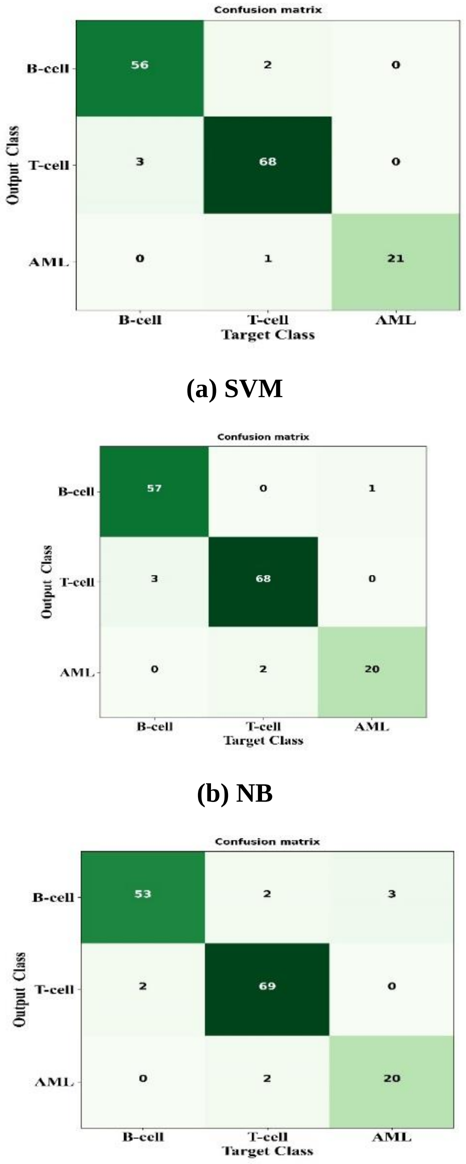
(d) KNN

Fig 10: Performance metrics for Dataset 5

Table 8: Values of proposed and existing methods for dataset 5.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9735	0.9735	0.9603	0.9823	0.947	0.9558	0.9294
precision	0.9689	0.9579	0.9261	0.9691	0.9129	0.9216	0.8968
Sensitivity	0.9592	0.9498	0.9315	0.9812	0.9143	0.9258	0.8913
Specificity	0.9767	0.9783	0.9684	0.9866	0.9559	0.9642	0.9392
F1-score	0.9640	0.9538	0.9288	0.9751	0.9136	0.9237	0.8940

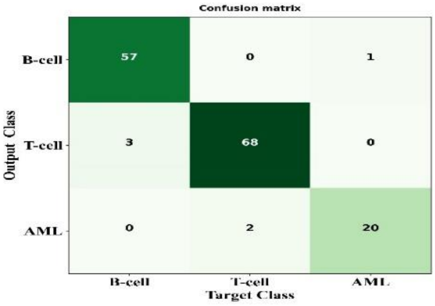
Figure 11(a)-(d) shows the confusion matrix attained by the proposed technique for Leukemia_3c classification. The proposed study categorized three classes of gene prediction: B-cell, T-cell, and Acute Myeloid Leukemia (AML). In Figure 11(a), when selecting the B-cell class, the proposed method recognized 56 as B-cell classes, 3 classified as T-cell classes, and none classified as AML classes. For classifying the T-cell classes, the proposed method recognized 2 as B-cell classes, 68 classified as T-cell classes, and one classified as AML classes. In Figure 11(b), when selecting the B-cell class, the proposed method recognized 57 B-cell classes, 3 classified as T-cell classes, and none classified as AML classes. For classifying the T-cell classes, the proposed method recognized none as B-cell classes, 68 classified as T-cell classes, and two classified as AML classes. In Figure 11(c), when selecting the B-cell class, the proposed method recognized 53 as B-cell classes, 2 classified as T-cell classes, and none classified as AML classes. For classifying the T-cell classes, the proposed method recognized two B-cell classes, 69 classified as T-cell classes, and two classified as AML classes. In Figure 11(d), when selecting the B-cell class, the proposed method recognized 57 B-cell classes, 3 classified as T-cell classes, and none classified as AML classes. For classifying the T-cell classes, the proposed method recognized none as B-cell classes, 68 classified as T-cell classes, and two classified as AML classes.



(a) SVM

(b) NB

(c) RF

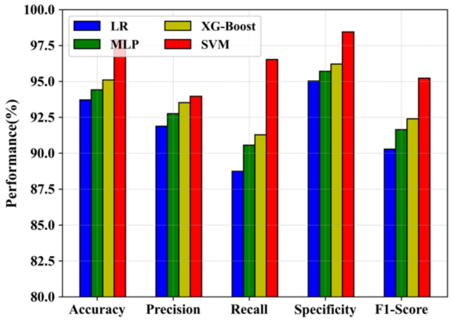


(d) KNN

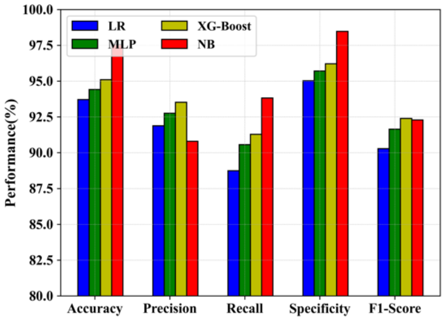
Fig 11: Confusion matrix for dataset 5

4.3.6 Performance evaluation for Dataset 6 (Leukemia_4c)

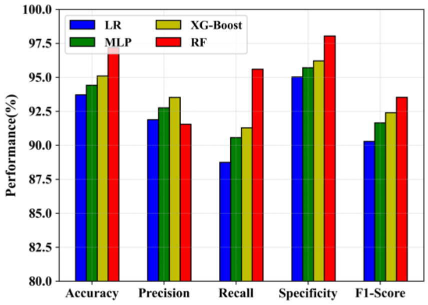
Figure 12 (a)-(d) depicts a comparison of the proposed and current approaches for dataset 6. In Figure 12(a), the proposed SVM method reaches an accuracy of 0.979, whereas the existing methods are LR (0.93), MLP (0.944), and XG-Boost (0.951). Then, the proposed method achieves a precision (0.93), recall (0.96), specificity (0.98), and F1-score (0.95). The accuracy of the proposed NB approach in Figure 12 (b) is 0.975. The accuracy achieved by the proposed RF technique in Figure 12 (c) is 0.97. The accuracy achieved by the suggested KNN approach in Figure 12 (d) is 0.96. In order to find significant genes, the proposed machine learning techniques are more objective and concentrate on the data. Table 9 depicts values for dataset 6 for proposed and existing approaches.



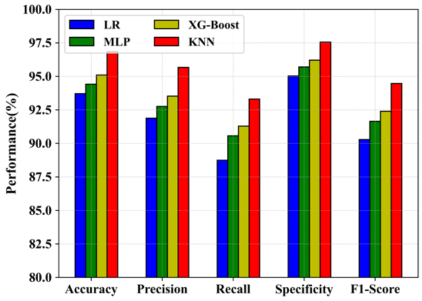
(a) SVM



(b) NB



(c) RF



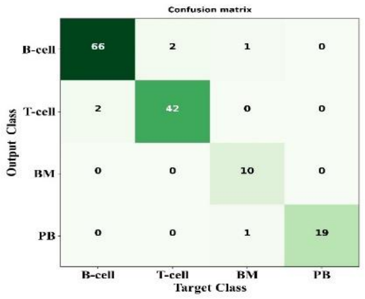
(d) KNN

Fig 12: Performance metrics for Dataset 6

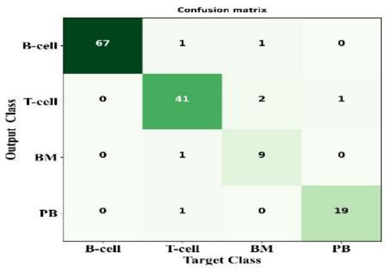
Table 9: Values for dataset 6 for proposed and existing approaches.

mod/per	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.979	0.9755	0.972	0.9685	0.9441	0.951	0.9371
precision	0.9396	0.9079	0.9154	0.9567	0.9275	0.9352	0.9188
Sensitivity	0.9652	0.9382	0.9559	0.9330	0.9055	0.9128	0.8874
Specificity	0.9844	0.9847	0.9803	0.9755	0.9570	0.9620	0.95028
F1-score	0.9522	0.9228	0.9352	0.9447	0.9164	0.9238	0.9028

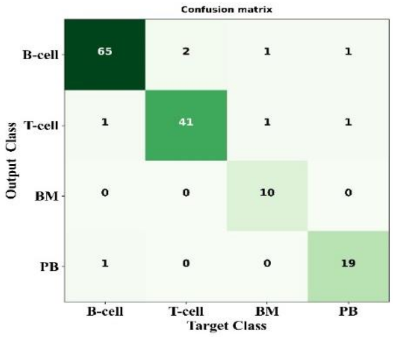
Figure 13(a)-(d) shows the confusion matrix attained by the proposed technique for Leukemia_4c classification. The proposed study categorized four classes of gene selection: B-cell, T-cell, bone marrow (BM), and peripheral blood (PB). In Figure 13(a), when selecting the B-cell class, the proposed method recognized 66 as B-cell classes, 2 classified as T-cell classes, and none classified as BM and PB classes. For classifying the T-cell classes, the proposed method recognized 2 as B-cell classes, 42 classified as T-cell classes, and none classified as BM and PB classes. In Figure 13(b), when selecting the B-cell class, the proposed method recognized 67 as B-cell classes, with none classified as T-cell, BM, and PB classes. For classifying the T-cell classes, the proposed method recognized one as B-cell classes, 41 classified as T-cell classes, and one classified as BM and PB classes. In Figure 13(c), when selecting the B-cell class, the proposed method recognized 65 as B-cell classes, one classified as T-cell classes, none classified as BM class, and one classified as PB class. For classifying the T-cell classes, the proposed method recognized two as B-cell classes, 41 classified as T-cell classes, and none classified as BM and PB classes. In Figure 13(d), when selecting the B-cell class, the proposed method recognized 64 as B-cell classes, 2 classified as T-cell classes, and none classified as BM and PB classes. For classifying the T-cell classes, the proposed method recognized 5 as B-cell classes, 42 as classified as T-cell classes, and one classified as BM and PB classes.



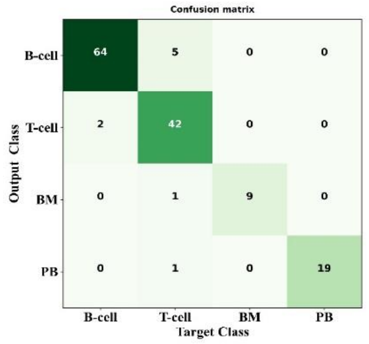
(a) SVM



(b) NB



(c) RF

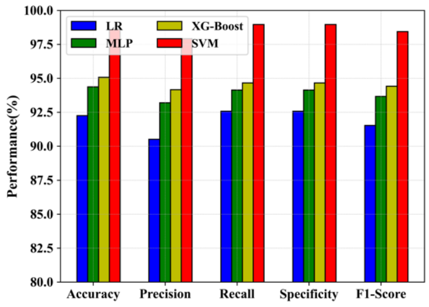


(d) KNN

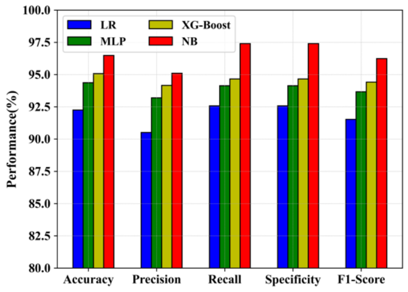
Fig 13: Confusion matrix for dataset 6

4.3.7 Performance evaluation for Dataset 7 (Lung cancer)

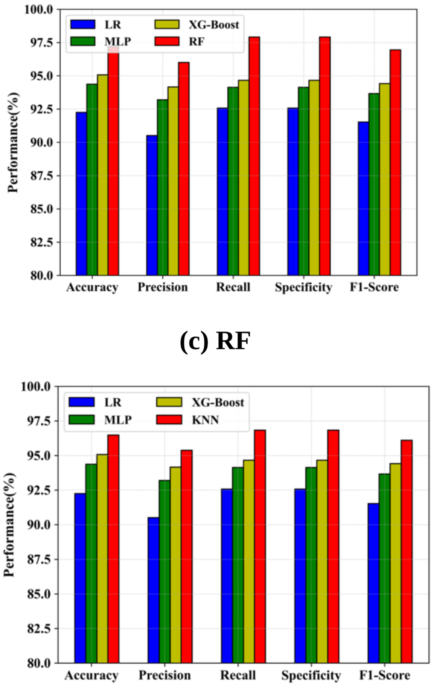
Figure 14 (a)-(d) compares the proposed and existing methods for dataset 7. In Figure 14(a), the proposed SVM method achieves an accuracy of 0.98, whereas the existing methods are LR (0.92), MLP (0.943), and XG-Boost (0.950). Then, the proposed method achieves a precision (0.97), recall (0.98), specificity (0.989), and F1-score (0.98). The proposed NB method in Figure 14(b) has an accuracy of 0.96. The proposed RF approach in Figure 14(c) yielded an accuracy of 0.971. In Figure 14(d), the accuracy attained by the proposed KNN technique is 0.964. Hence, the proposed method can assist in the creation of more precise and understandable models for tasks like cancer classification by choosing a smaller collection of relevant genes. Table 10 shows the values of proposed and existing methods for dataset 7.



(a) SVM



(b) NB



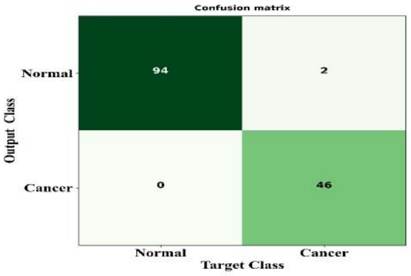
(c) RF

(d) KNN

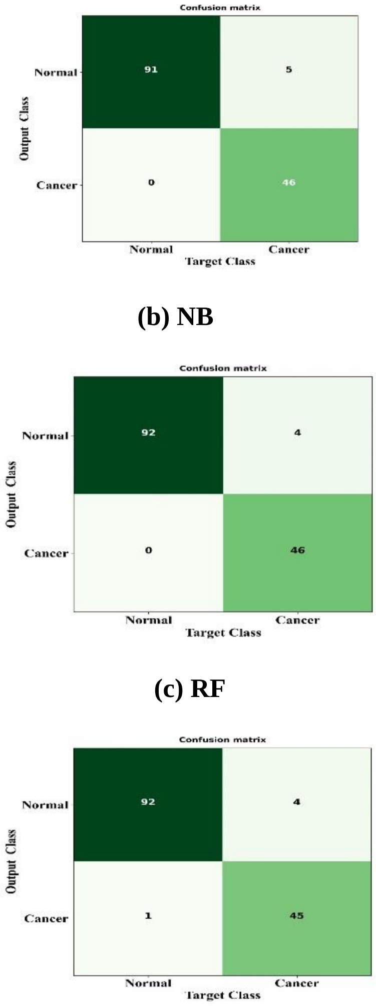
Fig 14: Performance metrics for Dataset 7

Table 10: Values of proposed and existing methods for dataset 7.

mod/per	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9859	0.9648	0.9718	0.9648	0.9437	0.9507	0.9225
precision	0.9791	0.950	0.960	0.9538	0.9319	0.9416	0.9050
Sensitivity	0.9895	0.9739	0.9791	0.9682	0.9413	0.9465	0.9257
Specificity	0.9893	0.9735	0.9790	0.9681	0.9410	0.9462	0.9254
F1-score	0.9843	0.9623	0.9694	0.9609	0.9366	0.9441	0.9152



(a) SVM



(b) NB

(c) RF

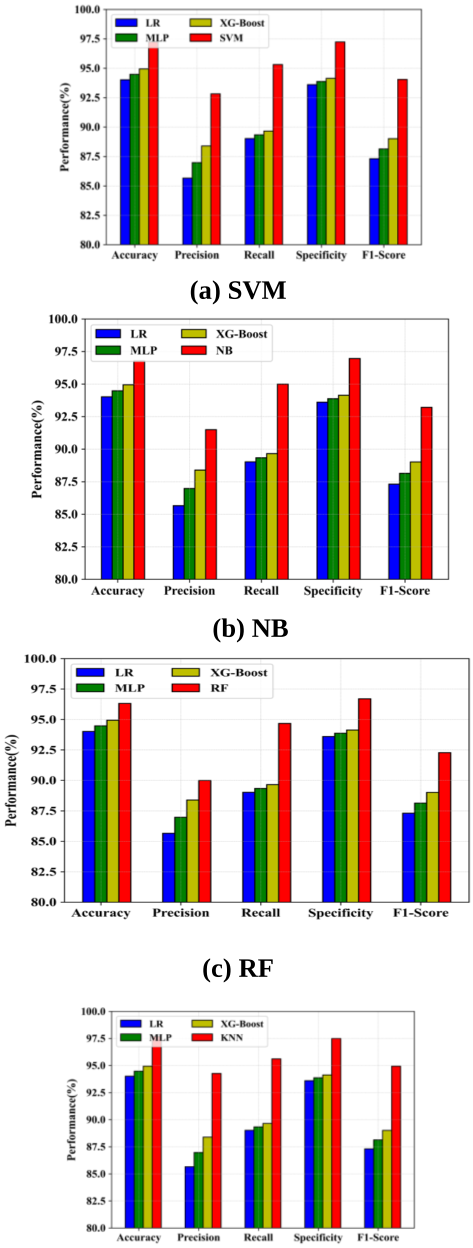
(d) KNN

Fig 15: Confusion matrix for dataset 7

Figure 15(a)-(d) shows the confusion matrix attained by the proposed technique for lung cancer classification. In Figure 15(a), when detecting the normal class, the proposed method recognized 94 as normal classes and none classified as cancer classes. For classifying the cancer classes, the proposed method recognized 46 as cancer, and 2 were classified to be normal class. In Figure 15(b), when detecting the normal samples, the proposed method recognized 91 as normal classes and none classified as cancer classes. For classifying the cancer classes, the proposed method recognized 46 as cancer, and 5 were classified as normal. In Figure 15(c), when detecting the normal samples, the proposed method recognized 92 as normal classes and none classified as cancer classes. For classifying the cancer classes, the proposed method recognized 46 as cancer, and 4 were classified as normal. In Figure 15(d), when detecting the normal samples, the proposed method recognized 92 as normal classes and one classified as cancer classes. For classifying the cancer classes, the proposed method recognized 45 as cancer, and 4 were classified to be normal class.

4.3.8 Performance evaluation for Dataset 8(Lymphoma)

The comparison between proposed and existing approaches for dataset 8 is depicted in Figure 16 (a)-(d), utilizing metrics such as accuracy, precision, recall, F1-score, and specificity. The accuracy achieved by the suggested SVM approach in Figure 16 (a) is 0.97, whereas the existing methods are LR (0.94), MLP (0.944), and XG-Boost (0.949). Then, the proposed method achieves a precision (0.92), recall (0.95), specificity (0.97), and F1-score (0.94). The accuracy achieved by the proposed NB approach in Figure 16(b) is 0.96. The accuracy achieved by the proposed RF technique in Figure 16 (c) is 0.963. The accuracy achieved by the suggested KNN algorithm in Figure 16(d) is 0.97. The proposed algorithms are capable of efficiently identifying the most relevant genes and are built to process high-dimensional data. Table 11 depicts values for dataset 8 for proposed and existing approaches.



(a) SVM

(b) NB

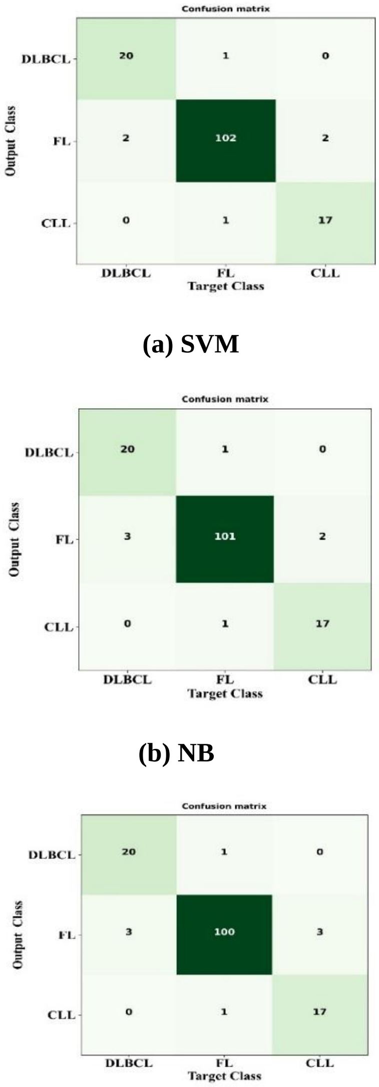
(c) RF

(d) KNN

Figure 16: Performance metrics for Dataset 8

Table 11: Values for dataset 8 for proposed and existing approaches.

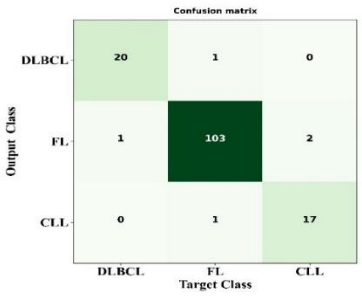
Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9724	0.9678	0.9632	0.977	0.9448	0.9494	0.9402
precision	0.9281	0.9149	0.8999	0.9426	0.8696	0.8838	0.8565
Sensitivity	0.9530	0.9498	0.9467	0.95617	0.8933	0.8964	0.8901
Specificity	0.9722	0.9695	0.9669	0.97496	0.9387	0.9413	0.9360
F1-score	0.9404	0.9320	0.9227	0.94938	0.8813	0.8901	0.8730



(a) SVM

(b) NB

(c) RF



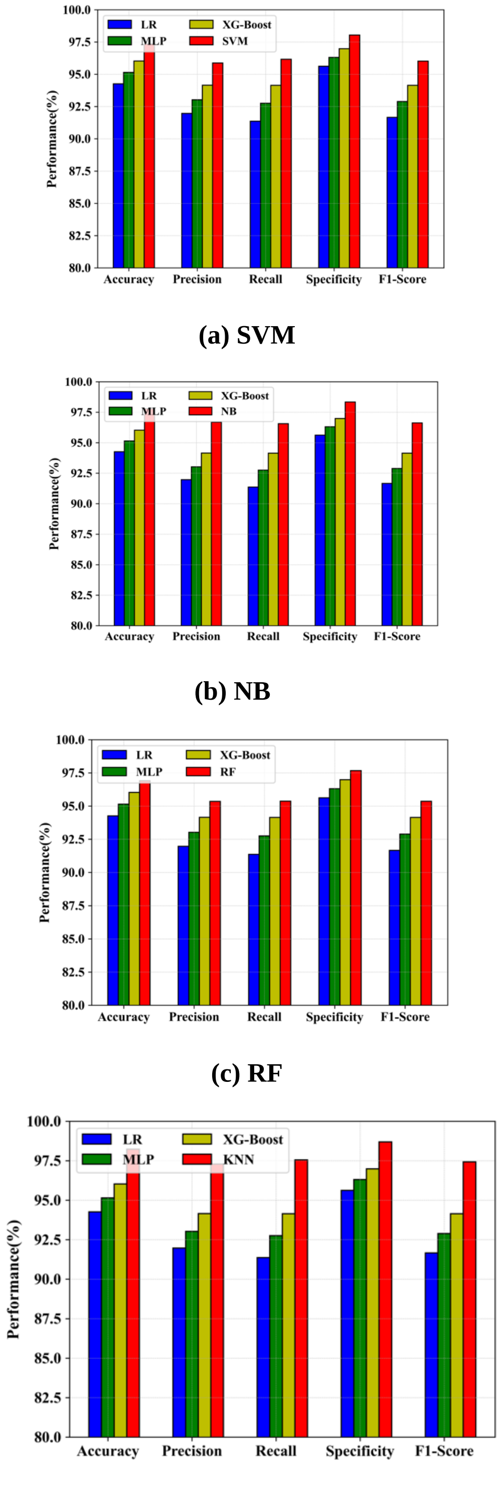
(d) KNN

Fig 17: Confusion matrix for dataset 8

Figure 17(a)-(d) shows the confusion matrix attained by the proposed technique for Lymphoma classification. The proposed study categorized three classes of gene selection: Diffuse large B cell lymphoma (DLBCL), Follicular lymphoma (FL), and chronic lymphocytic Leukemia (CLL). In Figure 17(a), when selecting the DLBCL class, the proposed method recognized 20 as DLBCL classes, 2 classified as FL classes, and none classified as CLL classes. For classifying the FL classes, the proposed method recognized one as DLBCL classes, 102 classified as FL classes, and one classified as CLL classes. In Figure 17(b), when selecting the DLBCL class, the proposed method recognized 20 DLBCL classes, 3 classified as FL classes and none classified as CLL classes. For classifying the FL classes, the proposed method recognized one as DLBCL classes, 101 classified as FL classes, and one classified as CLL classes. In Figure 17(c), when selecting the DLBCL class, the proposed method recognized 20 as DLBCL classes, 3 classified as FL classes, and none classified as CLL classes. For classifying the FL classes, the proposed method recognized one as DLBCL classes, 100 classified as FL classes, and one classified as CLL classes. In Figure 17(d), when selecting the DLBCL class, the proposed method recognized 20 DLBCL classes, one classified as FL classes and none classified as CLL classes. For classifying the FL classes, the proposed method recognized one as DLBCL classes, 103 classified as FL classes, and one classified as CLL classes.

4.3.9 Performance evaluation for Dataset 9(Mixed lineage Leukemia gene (MLL))

Figure 18 (a)-(d) compares the proposed and existing methods for dataset 9. In Figure 18(a), the proposed SVM method achieves an accuracy of 0.973, whereas the existing methods are LR (0.94), MLP (0.951), and XG-Boost (0.960). Then, the proposed method achieves a precision (0.95), recall (0.961), specificity (0.98), and F1-score (0.96). In Figure 10(b), the proposed NB method achieves an accuracy of 0.97. In Figure 18(c), the proposed RF method attains an accuracy of 0.969. In Figure 18(d), the proposed KNN method obtains an accuracy of 0.98. So, the proposed model may manage the high dimensionality by concentrating on the genes that have the greatest impact on class separation, lowering the model's complexity and increasing accuracy. Table 12 shows the values of proposed and existing methods for dataset 9.



(a) SVM

(b) NB

(c) RF

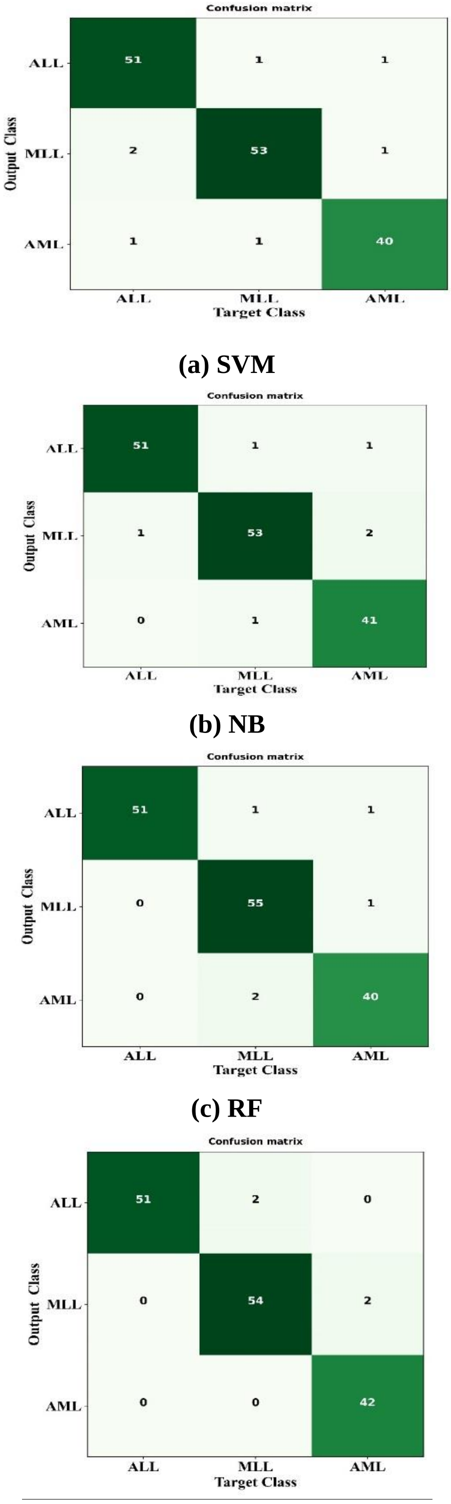
(d) KNN

Fig 18: Performance metrics for Dataset 9

Table 12: Values of proposed and existing methods for dataset 9.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9735	0.9779	0.9691	0.9823	0.9514	0.9603	0.9426

precision	0.9587	0.9668	0.9534	0.9729	0.9302	0.9414	0.9196
Sensitivity	0.9616	0.9655	0.9536	0.9755	0.9275	0.94145	0.913672
Specificity	0.9804	0.9833	0.9766	0.9868	0.9630	0.9698	0.956254
F1-score	0.9601	0.96624	0.9535	0.9742	0.9288	0.9414	0.916659



(a) SVM

(b) NB

(c) RF

(d) KNN

Fig 19: Confusion matrix for dataset 9

Figure 19(a)-(d) shows the confusion matrix attained by the proposed technique for mixed lineage Leukemia gene classification. The proposed study categorized three classes of gene prediction: Acute Lymphocytic Leukemia (ALL), mixed lineage Leukemia gene (MLL), and Acute Myeloid Leukemia (AML). In Figure 19(a), when selecting the ALL class, the proposed method recognized 51 as ALL classes, 2 classified as MLL classes, and one classified as AML classes. For classifying the MLL classes, the proposed method recognized one as ALL, 53 classified as MLL classes, and one classified as AML classes. In Figure 19(b), when selecting the ALL class, the proposed method recognized 51 as ALL classes, 1 classified as MLL classes, and none classified as AML classes. For classifying the MLL classes, the proposed method recognized one as ALL, 53 classified as MLL classes, and one classified as AML classes. In Figure 19(c), when selecting the ALL class, the proposed method recognized 51 as ALL classes and none classified as MLL and AML classes. For classifying the MLL classes, the proposed method recognized one as ALL, 55 classified as MLL classes, and two classified as AML classes. In Figure 19(d), when selecting the ALL class, the proposed method recognized 51 as ALL classes and none classified as MLL and AML classes. For classifying the MLL classes, the proposed method recognized two as ALL, 54 classified as MLL classes, and none classified as AML classes.

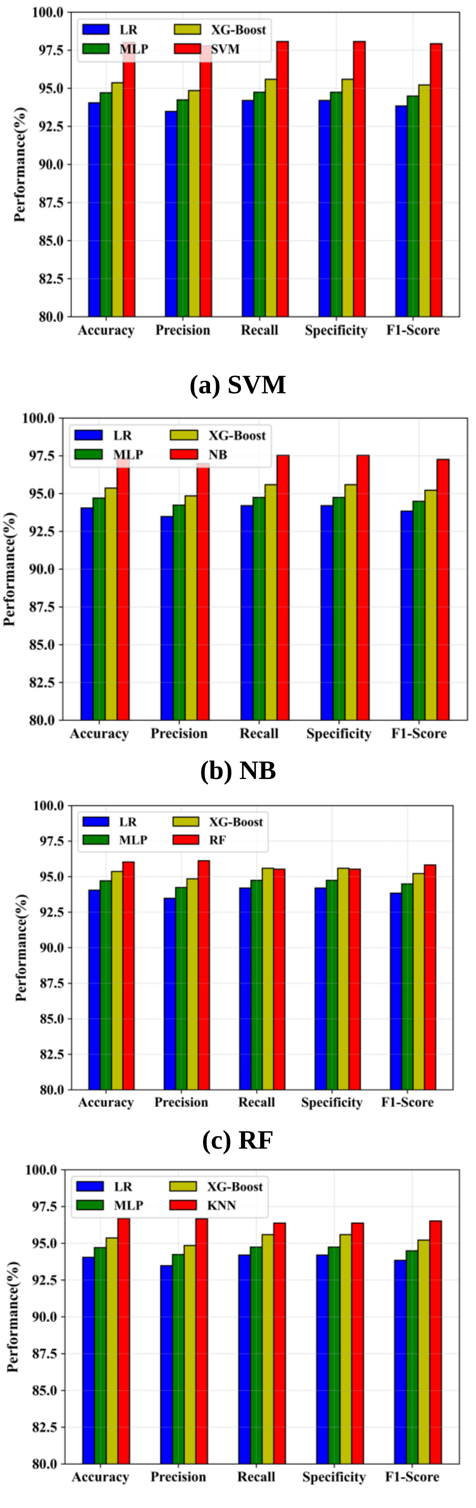
4.3.10 Performance evaluation for Dataset 10(Ovarian cancer)

Figure 20 (a)-(d) compares the proposed and existing methods for dataset 10. In Figure 20(a), the proposed SVM method accomplishes an accuracy of 0.980, whereas the existing methods are LR (0.94), MLP (0.94), and XG-Boost (0.95). Then, the proposed method achieves a precision (0.97), recall (0.98), specificity (0.98), and F1-score (0.97). In Figure 20(b), the proposed NB method obtains an accuracy of 0.973. In Figure 20(c), the proposed RF method attains an accuracy of 0.960. In Figure 20(d), the proposed KNN method reaches an accuracy of 0.96. Large quantities of gene expression data can be effectively analyzed using the proposed method because it is a computationally efficient and easy-to-build system. Table 13 depicts values for dataset 10 for proposed and existing approaches.

Table 13: Values for dataset 10 for proposed and existing approaches.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9801	0.9735	0.9603	0.9669	0.947	0.9536	0.9404
precision	0.9778	0.9698	0.9611	0.9666	0.94235	0.9484	0.9347

Sensitivity	0.9806	0.9752	0.9552	0.9637	0.9474	0.9558	0.9419
Specificity	0.9804	0.9750	0.9550	0.9635	0.9472	0.9554	0.9415
F1-score	0.9792	0.97253	0.9581	0.9651	0.9448	0.9521	0.9383



(a) SVM

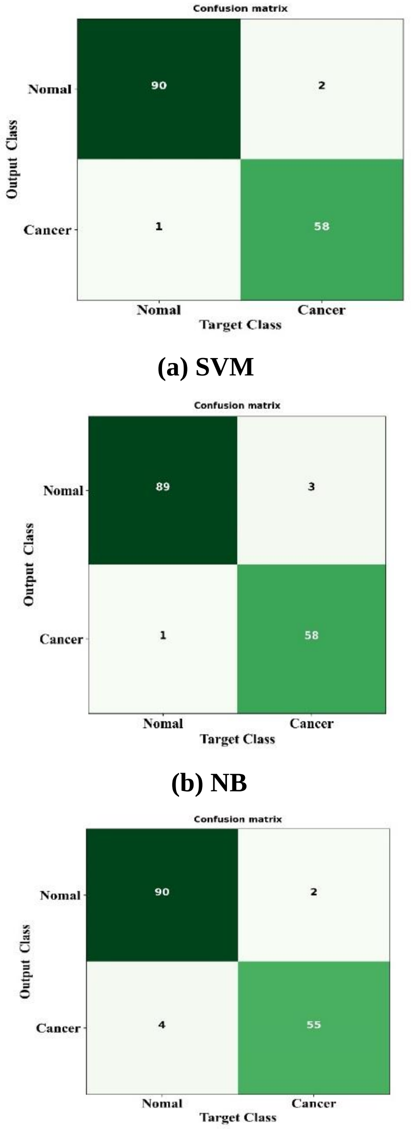
(b) NB

(c) RF

(d) KNN

Fig 20: Performance metrics for Dataset 10

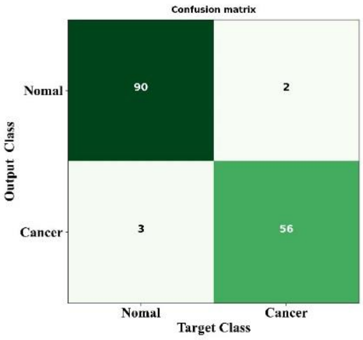
Figure 21(a)-(d) shows the confusion matrix attained by the proposed technique for ovarian cancer classification. In Figure 21(a), when detecting the normal class, the proposed method recognized 90 as normal classes and one classified as cancer classes. For classifying the cancer classes, the proposed method recognized 58 as cancer, and 2 were classified to be normal class. In Figure 21(b), when detecting the normal samples, the proposed method recognized 89 as normal classes and one classified as cancer classes. For classifying the cancer classes, the proposed method recognized 58 as cancer, and 3 were classified to be normal class. In Figure 21(c), when detecting the normal samples, the proposed method recognized 90 as normal classes and 4 classified as cancer classes. For classifying the cancer classes, the proposed method recognized 55 as cancer, and 2 were classified to be normal class. In Figure 21(d), when detecting the normal samples, the proposed method recognized 90 as normal classes and 3 classified as cancer classes. For classifying the cancer classes, the proposed method recognized 56 as cancer, and 2 were classified to be normal class.



(a) SVM

(b) NB

(c) RF

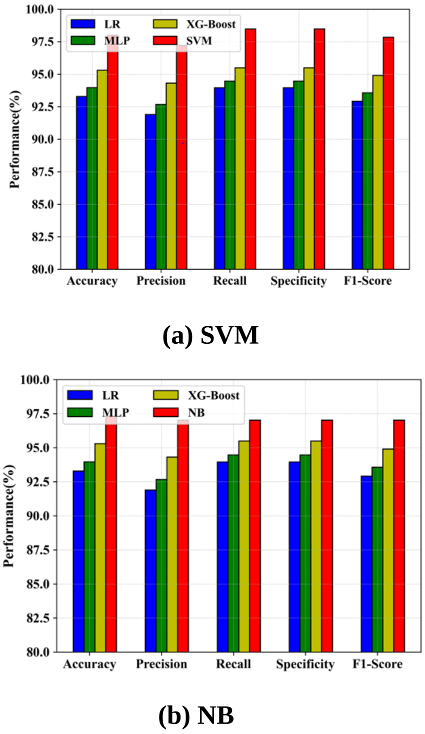


(d) KNN

Figure 21: Confusion matrix for dataset 10

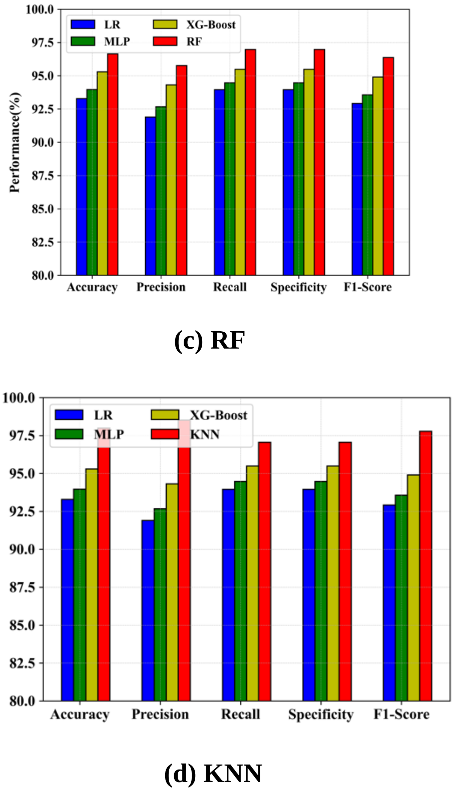
4.3.11 Performance evaluation for Dataset 11(SRBCT)

Dataset 11 is used to compare the proposed and current procedures in Figure 22(a)–(d). The proposed SVM approach achieves an accuracy of 0.97 in Figure 22(a), while the current methods include LR (0.93), MLP (0.939), and XG-Boost (0.953). The proposed method then yields an F1-score of 0.978, recall of 0.984, specificity of 0.98, and accuracy of 0.972. The proposed NB approach achieves an accuracy of 0.972 in Figure 22(b). The proposed RF approach achieves an accuracy of 0.966 in Figure 22(c). The proposed KNN approach obtains an accuracy of 0.97 in Figure 22(d). The values of the proposed and current procedures for dataset 11 are displayed in Table 14.



(a) SVM

(b) NB



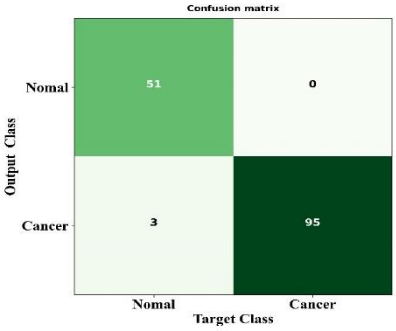
(c) RF

(d) KNN

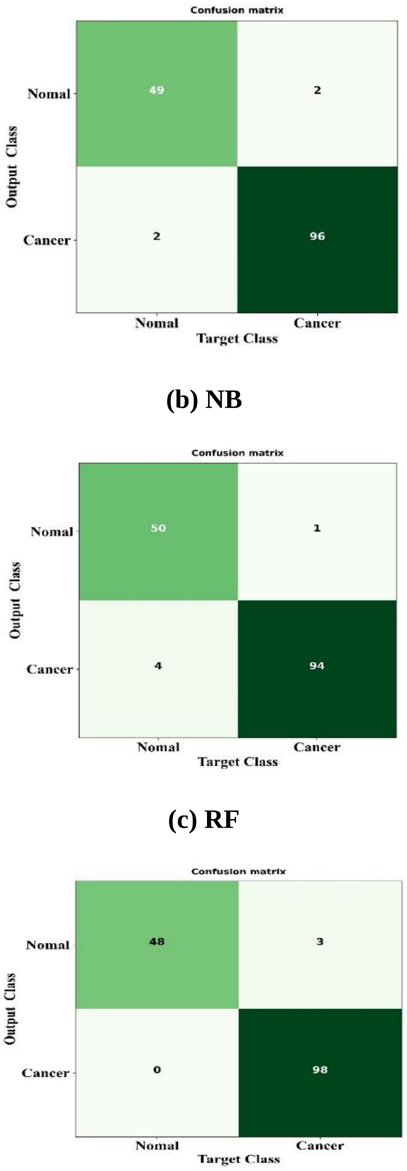
Figure 22: Performance metrics for Dataset 11

Table 14: Values of proposed and existing methods for dataset11.

Metrics	SVM	NB	RF	KNN	MLP	XG-Boost	LR
Accuracy	0.9799	0.9732	0.9664	0.9799	0.9396	0.953	0.9329
precision	0.9722	0.9701	0.9577	0.9851	0.9267	0.9431	0.9189
Sensitivity	0.9846	0.9701	0.9697	0.9705	0.9446	0.9548	0.9395
Specificity	0.9845	0.9701	0.9697	0.9703	0.9443	0.9545	0.9393
F1-score	0.9784	0.9700	0.9637	0.9778	0.9356	0.9489	0.9291



(a) SVM



(b) NB

(c) RF

(d) KNN

Fig 23: Confusion matrix for dataset 11

Figure 23(a)-(d) shows the confusion matrix attained by the proposed technique for SRBCT classification. In Figure 23(a), when detecting the normal class, the proposed method recognized 51 as normal classes and 3 classified as cancer classes. For classifying the cancer classes, the proposed method recognized 95 as cancer, and none were classified to be normal class. In Figure 23(b), when detecting the normal samples, the proposed method recognized 49 as normal classes and 2 classified as cancer classes. For classifying the cancer classes, the proposed method recognized 96 as cancer, and 2 were classified to be normal class. In Figure 23(c), when detecting the normal samples, the proposed method recognized 50 as normal classes and 4 classified as cancer classes. For classifying the cancer classes, the proposed method recognized 94 as cancer, and one is classified to the normal class. In Figure 23(d), when detecting the normal samples, the proposed method recognized 40 as normal classes and none classified as cancer classes. For classifying the cancer classes, the proposed method recognized 98 as cancer, and 3 were classified to be normal class. So, the proposed method identifies gene selection more accurately.

4.4 Discussion

Data on gene expression has been effectively used for a number of applications, especially the classification of cancer. The drawbacks of generalizability issues, inaccurate feature selection, and over-fitting provide difficulties in the creation of efficient classifiers for expression data. Overcoming the aforementioned challenges and improving a classifier's predicted accuracy can be accomplished with efficiency and effectiveness using gene selection. Initially, the data are augmented using the SMOTE model to enhance microarray data. Next, features are extracted, and Shapley values are calculated using the CkSV technique. The HGDBO method was utilized to select the most essential features. Moreover, the procedure is run on the Apache Hadoop Distributed File System for economical storage of big datasets. Furthermore, a variety of machine learning techniques, including Support Vector Machine, Naive Bayes, Random Forest, and K-nearest neighbor are used to classify the characteristics. Then, the real-world microarray dataset for the SVM classifier, dataset 1 has an accuracy (0.97), dataset 2 (0.98), dataset3 (0.968), dataset 4 (0.975), dataset 5 (0.973), dataset 6 (0.979), dataset 7 (0.985), dataset 8(0.972), dataset 9 (0.973), dataset 10(0.980) and dataset 11(0.979). Thus, the proposed method produces superior outcomes than compared to existing methods. Table 15 compares the existing work with the proposed method.

Table 15: Comparison of existing work with the proposed method.

Author name and Reference	Technique used	Performance
Ali et al. [21]	Hybrid filter-genetic feature selection strategy	Attain 93.81% accuracy, 93.8% recall, precision, and F-measure by RF
Akhavan et al. [22]	Two-phase microarray data gene selection technique	Obtain at least 99% accuracy
Alomari et al. [23]	MGWO	Attain an accuracy of 0.9586
Deng et al. [24]	XGBoost-MOGA	Produces accuracy of 83.33% in CNS dataset
Azadifar et al. [25]	SMCEC	Attain accuracy of about 92.09% in the leukemia dataset

Proposed	SMOTE , CkSV, HGDBO, GA and DBO	The real-world microarray dataset for SVM classifier, dataset 1 has an accuracy (0.97), dataset 2 (0.98), dataset3 (0.968), dataset 4 (0.975), dataset 5 (0.973), dataset 6 (0.979), dataset 7 (0.985), dataset 8(0.972), dataset 9 (0.973), dataset 10(0.980) and dataset 11(0.979).

5. Conclusion and Future Scope

The identification of important cancer-related genes has drawn the attention of biologists and is important for both cancer diagnosis and treatment. The proposed SMOTE equalizes the distribution of classes by effectively increasing the quantity of data points within the minority class. To improve interpretability, the major feature from the higher level gene data is extracted using CkSV. Thus, the proposed method uses a novel hybrid bio-inspired model for gene selection that speeds up the learning procedure, referred to as HGDBO, which combines features of GA and DBO algorithms. In addition, this study also presented a novel machine learning approach that could be used to identify significant genes and improve classification accuracy using microarray datasets. Experiments for SVM classifier, dataset 1 has an accuracy (0.97), dataset 2 (0.98), dataset3 (0.968), dataset 4 (0.975), dataset 5 (0.973), dataset 6 (0.979), dataset 7 (0.985), dataset 8(0.972), dataset 9 (0.973), dataset 10(0.980) and dataset 11(0.979). Additionally, gradient boosting and other boosted algorithm families can be tried in future study to increase the model's predicted accuracy. Some distinct boosting algorithms have mathematical formulas, including AdaBoost and Gentle Boost. Gradient Boosting's development and extension contribute to the criterion-fitting process.

Conflicts of Interest

Authors declare that they have no conflict of interest.

Funding Statement

No funding is provided for the preparation of manuscript.

Acknowledgement

None

References

[1] E.A. Alhenawi, R. Al-Sayyed, A. Hudaib and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review." *Computers in Biology and Medicine* vol. 140, pp. 105051, 2022.

[2] Wang, H. Liu, J. Yang and G. Chen, "Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data." Computers in biology and medicine vol. 142, pp. 105208, 2022.

[3] S. Osama, H. Shaban, and A.A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review." *Expert Systems with Applications* vol. 213, pp. 118946, 2023.

[4] H.Z. Almarzouki, "Deep-learning-based cancer profiles classification using gene expression data profile." *Journal of Healthcare Engineering* vol. 2022, 2022.

[5] Jahwar and N. Ahmed, "Swarm intelligence algorithms in gene selection profile based on classification of microarray data: a review." Journal of Applied Science and Technology Trends vol. 2, no. 01, pp. 01-09, 2021.

[6] X. Zheng and C. Zhang, "Gene selection for microarray data classification via dual latent representation learning." *Neurocomputing* vol. 461, pp. 266-280, 2021.

[7] H. Almazrua and H. Alshamlan, "A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data." *IEEE Access* vol. 10, pp. 71427-71449, 2022.

[8] S. Gupta, M.K. Gupta, M. Shabaz and A. Sharma, "Deep learning techniques for cancer classification using microarray gene expression data." *Frontiers in Physiology* vol. 13, pp. 952709, 2022.

[9] H.S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification." *CAAI Transactions on Intelligence Technology* vol. 5, no. 1, pp. 22-33, 2020.

[10] Dabba, A. Tari, S. Meftali and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm." Expert Systems with Applications vol. 166, pp. 114012, 2021.

[11] Haznedar, M.T. Arslan and A. Kalinli, "Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data." Medical & Biological Engineering & Computing vol. 59, pp. 497-509, 2021.

[12] E.H. Houssein, D.S. Abdelminaam, H.N. Hassan, M.M. Al-Sayed and E. Nabil, "A hybrid barnacles mating optimizer algorithm with support vector

machines for gene selection of microarray cancer classification." *IEEE Access* vol. 9, pp. 64895-64905, 2021.

[13] M.L.R. AbdElNabi, M.W. Jasim, H.M. El-Bakry, M.H.N. Taha, and N.E.M. Khalifa, "Breast and colon cancer classification from gene expression profiles using data mining techniques." *Symmetry* vol. 12, no. 3, pp. 408, 2020.

[14] Y.K. Saheed, "Effective dimensionality reduction model with machine learning classification for microarray gene expression data." *In Data science for genomics, Academic Press*, pp. 153-164, 2023.

[15] M. Rostami, K. Berahmand and S. Forouzandeh, "A novel community detection based genetic algorithm for feature selection." *Journal of Big Data* vol. 8, no. 1, pp. 2, 2021.

[16] U. Rahardja, A. Sari, A.H. Alsalamy, S. Askar, A.H.R. Alawadi and B. Abdullaeva, "Tribological properties assessment of metallic glasses through a genetic algorithm-optimized machine learning model." *Metals and Materials International* vol. 30, no. 3, pp. 745-755, 2024.

[17] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari and M. Oussalah, "Gene selection for microarray data classification via multi-objective graph theoretic-based method." *Artificial Intelligence in Medicine* vol. 123, pp. 102228, 2022.

[18] Ž. Avsec, V. Agarwal, D. Visentin, J.R. Ledsam, A. Grabska-Barwinska, K.R. Taylor, Y. Assael, J. Jumper, P. Kohli and D.R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions." *Nature methods* vol. 18, no. 10, pp. 1196-1203, 2021.

[19] S.H. Shah, M.J. Iqbal, I. Ahmad, S. Khan and J.J.P.C. Rodrigues, "Optimized gene selection and classification of cancer from microarray gene expression data using deep learning." *Neural Computing and Applications* pp. 1-12, 2020.

[20] S.K. Baliarsingh, S. Vipsita and B. Dash, "A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm." *Neural Computing and Applications* vol. 32, pp. 8599-8616, 2020.

[21] W. Ali and F. Saeed, "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data." *Processes* vol. 11, no. 2, pp. 562, 2023.

[22] M. Akhavan and S.M.H. Hasheminejad, "A two-phase gene selection method using anomaly detection and genetic algorithm for microarray data." *Knowledge-Based Systems* vol. 262, pp. 110249, 2023.

[23] O.A. Alomari, S.N. Makhadmeh, M.A. Al-Betar, Z.A.A. Alyasseri, I.A. Doush, A.K. Abasi, M.A. Awadallah and R.A. Zitar, "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators." *Knowledge-Based Systems* vol. 223, pp. 107034, 2021.

[24] X. Deng, M. Li, S. Deng, and L. Wang, "Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification." *Medical & Biological Engineering & Computing* vol. 60, no. 3, pp. 663-681, 2022.

[25] S. Azadifar, M. Rostami, K. Berahmand, P. Moradi, and M. Oussalah, "Graph-based relevancy-redundancy gene selection method for cancer diagnosis." *Computers in Biology and Medicine* vol. 147, pp. 105766, 2022.

[26] H. Mansourifar and W. Shi, "Deep synthetic minority over-sampling technique." *arXiv preprint arXiv:2003.09788, 2020*.

[27] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using shapley values." *In Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4, Springer International Publishing*, pp. 17-38, 2020.

[28] M.A. Albadr, S. Tiun, M. Ayob and F. Al-Dhief, "Genetic algorithm based on natural selection theory for optimization problems." *Symmetry* vol. 12, no. 11 (2020): 1758.

[29] Xue, Jiankai, and Bo Shen. "Dung beetle optimizer: A new meta-heuristic algorithm for global optimization." *The Journal of Supercomputing* 79, no. 7, pp. 7305-7336, 2023.

[30] G. Gunawan, H. Hanes and C. Catherine, "C4. 5, K-Nearest Neighbor, Naïve Bayes, and Random Forest Algorithms Comparison to Predict Students' on TIME Graduation." *Indonesian Journal of Artificial Intelligence and Data Mining* vol. 4, no. 2, pp. 62-71, 2021.

https://csse.szu.edu.cn/staff/zhuzx/datasets.html