ISSN: 2093-5374 / E-ISSN: 2093-5382

Optimization Enabled Ensemble Learning for Leukemia Classification Using Microarray Data

P.C. Chaitra[1*], and Dr.R. Saravana Kumar[2]

[1*]Assistant Professor, Computer Science and Engineering, Christ University, Kengeri; Research scholar, Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Visvesvaraya Technological University, Belagavi, India. chaitrapcjay@gmail.com, https://orcid.org/0009-0009-8103-2293

[2]Research Supervisor, Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Visvesvaraya Technological University, Belagavi, India. dr.saravanakumar@dsatm.edu.in, https://orcid.org/0000-0002-4622-5745

Received: April 06, 2025; Revised: May 23, 2025; Accepted: June 10, 2025; Published: June 30, 2025

Abstract

Leukemia classification involves identification and categorization of various leukemias, a cluster of blood malignancies influencing white blood cells. Proper classification is crucial for selecting the appropriate treatment modalities and predicting outcomes in patients. Historically, leukemia classification was based on clinical and morphological characteristics, but new developments in genomics like microarray and next-generation sequencing tools have facilitated more accurate molecular classifications. Machine learning (ML) and deep learning (DL) methods have transformed leukemia classification by enabling automation of analysis in large and intricate datasets to ensure more accurate and efficient leukemia subtype classification. The primary goal of this research is to suggest a new leukemia classification method using microarray data. Leukemia microarray data first undergoes preprocessing, after which feature selection is performed through Serial Exponential-Secretary Bird Optimization Algorithm (SE-SBOA). SE-SBOA is an optimization method that embeds the exponential weighted moving average concept (EWMA) into Secretary Bird Optimization Algorithm (SBOA). The method helps to find the best feature subset, improving model performance at lower complexity. Lastly, leukemia classification is done using the proposed ensemble method that combines Graph Neural Network (GNN), Multi-Layer Perceptron (MLP) and Random Forest. Utilizing the advantages of GNN, MLP and Random Forest, the model proposed herein attains higher classification accuracy and proves to outperform traditional methods. Experimental results demonstrate that the SE-SBOA-based Ensemble Learning technique outperformed standard methods, attaining an accuracy of 95.9%, a precision of 96.1%, a recall of 96.2%, and an F1-score of 96.2%.

Keywords: Leukemia Classification, Microarray Data, Disease Detection, Optimization Algorithm, Deep Learning.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 16, number: 2 (June), pp. 903-926. DOI: *10.58346/JOWUA.2025.I2.056*

*Corresponding author: Assistant Professor, Computer Science and Engineering, Christ University, Kengeri; Research scholar, Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Visvesvaraya Technological University, Belagavi, India.

1 Introduction

Cancer has emerged as a predominant cause of mortality on a global scale. This phenomenon arises from various pathological issues within the organism or heritable diseases transmitted through genetic lineage. WHO indicates that cancer constitutes a significant factor contributing to mortality among individuals under the age of 70 in 112 countries out of a total of 183 (Gowin et al., 2021). In at least 23 countries, cancer is categorized as the third or fourth most significant cause of death. The available treatment modalities for the most prevalent forms of cancer are relatively limited in their efficacy (Dutta et al., 2025). Typically, medical professionals diagnose cancer through histopathological examination of minute tissue utilizing a microscope. However, this diagnostic approach is associated with substantial costs, prolonged processing times, and the potential for erroneous diagnostic outcomes. Leukemia is a specific type of hematological malignancy and is classified alongside Lymphoma and Myeloma (Hoffmann et al., 2006). These techniques necessitate a high degree of expertise and considerable effort. In addition, medical practitioners conduct hematological assessments and bone marrow analyses to diagnose leukemia; nevertheless, these methodologies are noted for their substantial financial implications and temporal demands. Leukemia is defined by an abnormal proliferation of immature leukocytes, which subsequently invade the bone marrow (Raju et al., 2020). This invasion compromises the marrow's capacity to generate other crucial hematopoietic cells that are vital for maintaining a robust immune response and overall hematological well-being. Pathologists who specialize in identification of leukemia utilize a range of methodologies, including array-based CGH, LDI-PCR, and ACGH (Zakir Ullah et al., 2021). Interventional radiology, as a complementary approach, is also constrained by the impact of hereditary factors on imaging outcomes (Alabdulqader et al., 2024).

Leukemia profoundly impairs normal functions of healthy hematological cells. Bone marrow serves a critical role in the production of diverse blood cell lineages and progenitor cells, which are collectively termed hematopoietic stem cells. In specific circumstances, hematopoietic stem cells may differentiate into aberrant white blood cells, thereby compromising the functionality of healthy blood cells and ultimately precipitating the development of leukemia (Oybek Kizi et al., 2025). The field of leukemia research has garnered significant attention among clinical investigators, who aspire to improve diagnostic and therapeutic modalities for patients due to the disease's high mortality rate (Labaj et al., 2017). Traditional approaches for the classification of distinct leukemia types involve physical examinations, hematological evaluations, bone marrow aspirations, and imaging techniques (Krishnan & Dandekar, 2022). Standard practices include blood smear examination, comprehensive blood counts, differential leukocyte counts, flow cytometry, X-ray imaging, CT scans and PET scans (Agarwal & Singh, 2024; Maaroof et al., 2025). The disease can be categorized into four primary classifications: ALL, CML, AML and CLL. Proper diagnosis plays a fundamental role in managing leukemia since it directs healthcare providers to the best therapeutic approaches that can be administered to the patient. The classification patterns of four fundamental types of leukemia are founded upon particular anatomical sites affected as well as characteristics of disease spread (Oybek Kizi et al., 2025). The acute form of leukemia manifests more aggressive as well as more rapid progression than its chronic equivalents. Hence, the treatment of ALL and AML requires a strong approach, as any delay in their identification can allow for accelerated disease progression (Hartigan, 2023). The treatment processes of ALL and AML differ, thus highlighting the significance of differentiating between these two for planning of treatment (Rupapara et al., 2022). The diagnosis of ALL and AML types of leukemia demands improvement of specialized skill, and the process can prove to be time-consuming and fiscally taxing. One of the major challenges of leukemia management is the morphological resemblance of cells, which could have different reactions to therapeutic compounds and cytotoxic drugs. Such factors highlight the

need for creating sophisticated diagnostic approaches that employ next-generation technologies for precise identification of the particular form of leukemia using gene expression data (Selvaraj et al., 2024).

As a result, research conducted to reveal the genetic causes of Leukemia and to enable it to be diagnosed at an early stage with ease is gaining prominence. The microarray technology, which enables one to analyze and quantify gene expression in thousands of genes at a time, has a central role to play in understanding interactions between genes and cancer and inter-gene relations (Wang et al., 2022). Microarray data contain thousands of genes per sample; yet, the inherent data acquisition challenges lead to a shortage of samples (Mahmoudi & Lailypour, 2015). Microarray cancer data analysis is an important research challenge cutting across disciplines, including statistics, ML, computational biology, pattern recognition, and other associated fields, as it is a key to improving the detection, diagnosis and treatment of cancer. Research in this area is aimed at improving cancer patients' survival rates through the optimization of screening and treatment methods as well as the extension of pertinent knowledge. The primary challenges associated with the classification of microarray datasets stem from various issues, including noisy data, class imbalances, a deficiency of adequate samples and the formidable curse of dimensionality of features, which complicates analysis and results in inaccurate classifications (Selvaraj et al., 2024). A multitude of investigations concerning the categorization of binary-class microarray cancer data have been undertaken. Nonetheless, the categorization of multi-class microarray data persists as a salient research challenge due to the difficulties engendered by class imbalances, whereby classes with a limited quantity of samples are frequently marginalized as a result of the predispositions of the majority of models that tend to favor classes with a greater number of instances (Dagnew & Shekar, 2021). Artificial Intelligence instruments, especially algorithms, play a crucial role in facilitating the understanding of gene expression data and the classification of genes, particularly within the domain of oncology. Meta-heuristic algorithms have been explicitly formulated to tackle such complexities. The principal advantages of utilizing meta-heuristic algorithms for the resolution of NP-hard problems encompass their resilience in discerning optimal solutions for intricate optimization endeavors and their effectiveness in addressing large-scale problems within acceptable computational time constraints (Krishnan & Dandekar, 2022).

The main goal of this work is to classify leukemia using optimization-enabled feature selection and an ensemble learning approach. Initially, Leukemia microarray data is provided for preprocessing, which is executed by handling missing values, data oversampling and Standard Scaler. After preprocessing, feature selection is performed by the proposed SE-SBOA (Ghaderzadeh et al., 2021). Here, SE-SBOA is generated by integrating the EWMA concept in SBOA so as to identify optimal feature subsets, improving model performance and reducing complexity. Lastly, classification of leukemia is performed by proposed Ensemble-based classification, which combines methods like GNN, MLP and Random Forest. Combining GNN, MLP, and Random Forest leverages their strengths to enhance model performance while keeping it simple and avoiding overfitting.

Principal Objectives of this Paper

- To perform leukemia classification, an ensemble learning-based approach utilizes methods, such as GNN, MLP, and Random Forest, to improve model performance while reducing complexity and overfitting.
- For feature selection, a new algorithm, titled Serial Exponential-Secretary Bird Optimization Algorithm (SE-SBOA), is derived by integrating the serial exponential weighted moving average concept and SBOA (Mahmoudi & Lailypour, 2015).

- To evaluate the performance of the proposed method, accuracy, precision, recall, and F1-score are used to compare with those of existing methods.

Paper Organization

Introduction is given in the first section. The second section discusses previous studies and research related to leukemia classification. The third section explains the proposed approach of leukemia classification. Result is given in the fourth section, and conclusion in the fifth section.

2 Literature Survey

This section examines existing research that has classified various types of leukemia using various methodologies. The chosen studies were examined and evaluated for their strengths and weaknesses with regard to leukemia classification.

Sharanya Selvaraj et al., (Selvaraj et al., 2024) used a cutting-edge super-learning paradigm combining several ML models to analyze gene expression data and classify leukemia cells. The method used an entropy-based feature importance technique to select gene profiles that are most crucial for the classification task. Efficiency of this super-learning paradigm was emphasized by its final super learner, Random Forest, which effectively classified cross-validated datasets from candidate models. Empirical verification conducted on a gene expression monitoring data set demonstrated that this model outperformed other state-of-the-art models with regard to prediction accuracy. Essam H. Houssein et al., (Mahmoudi & Lailypour, 2015) proposed a BMO algorithm combined with SVM for gene feature selection in cancer classification using microarray data. The BMO algorithm was initially used to determine the most relevant genes, thus optimizing the subset of features in order to increase classification accuracy. Later on, the identified genes were classified using an SVM model, and there was a classification performance improvement. The hybrid approach showed higher accuracy and effectiveness in cancer classification through dimension reduction and focusing on the most significant features in the dataset.

Aiguo Wang et al proposed an ensemble approach to selection of features for the discovery of stable biomarkers and classification of cancer using microarray expression data. First, a range of feature selection methods was used to identify relevant features across models of varying nature. Next, these features identified were integrated using an ensemble approach to ensure both stability and robustness. The approach aimed to enhance accuracy and reliability of cancer classification by focusing on the most informative biomarkers. The ensemble approach improved overall effectiveness by combining several views of feature selection. Vaibhav Rupapara et al., (Krishnan & Dandekar, 2022) proposed a methodological framework for hematological malignancy prediction using leukemia microarray gene expression data along with a hybrid LVT model. First, the gene expression data were preprocessed to deal with missing data and to attain normalization. Next, feature selection was performed to identify the relevant genes required for the classification process. The hybrid LVT model, which combines the concepts of logistic regression and vector trees, was used to classify the leukemia samples. This approach improved predictive performance by skillfully combining the strengths of logistic regression and vector trees in analysis of complex gene expression data relevant to blood cancer classification.

Guesh Dagnew and B.H. Shekar (Dagnew & Shekar, 2021) investigated ensemble learning methods for the classification of microarray cancer data using tree-based features. First, methods for tree-based feature selection were used to identify the most relevant features from the gene expression dataset. Thereafter, a framework for ensemble learning that aggregates several classifiers was used to classify

the cancer samples. This approach greatly enhanced classification accuracy by linking the strengths of heterogeneous models to provide a strong and accurate prediction of cancer based on the chosen features. Ebtisam Abdullah Alabdulqader et al., (Agarwal & Singh, 2024) sought to improve the predictive ability for hematological malignancies using leukemia microarray gene data combined with Chi2 features using a weighted CNN. First, the Chi2 feature selection technique was used to identify the most relevant genes from the microarray data. The identified features were then used to train a weighted CNN model, which enhanced classification effectiveness by assigning differential importance to different features. This approach efficiently enhanced predictive precision for hematological cancers by interpreting intricate patterns in the gene expression profile and optimally improving the model's performance in classifying leukemia.

Surbhi Gupta et al. (Gupta et al., 2022) employed state-of-the-art DL techniques for cancer classification based on microarray gene expression data (Jothiaruna, 2022). Initially, gene expression data sets were processed meticulously to remove the missing values problem and normalize the data set. Next, the CNN architecture was implemented to automatically extract features and identify complex patterns hidden in the data. The model was trained rigorously to classify different cancer types based on the expression levels of relevant genes. This approach greatly enhanced classification accuracy by taking advantage of the strong feature extraction ability built into DL models in handling high-dimensional gene information. Mehmet Bilen et al., (Bilen et al., 2020) proposed a method involving a hybrid ensemble gene selection strategy along with a refined GA towards leukemia classification using microarray gene expression data sets. Initially, the advanced GA was utilized to refine gene selection through the identification of the most pertinent features. The identified genes were subsequently employed within ensemble models to enhance the accuracy of classification. This methodology effectively augmented predictive performance by utilizing GA and ensemble learning for optimal feature selection and resilient classification.

2.1 Research Gaps

Based on challenges delineated in the preceding section, ensuing research deficiencies have been articulated:

- Random Forest methodology delineated in (Selvaraj et al., 2024) effectively classifies cross-validated datasets derived from prospective learners; however, its efficacy is limited by the dimensionality of the training dataset employed for model development.
- BMO-SVM algorithm as presented in (Mahmoudi & Lailypour, 2015) exhibits a significantly elevated informational superiority ratio relative to alternative algorithms; nevertheless, it neglects to assess generalizability of selected features or classification accuracy across varying frameworks.
- The decision trees method in (Wang et al., 2022) obtains better stability scores and achieves classification performance, but it ignores the relative importance of features.
- CNN (Gupta et al., 2022) demonstrates substantial potential within the field of genomics, especially concerning gene microarray datasets, but it does not provide an assessment of accuracy.
- Finally, Enhanced GA study in (Bilen et al., 2020) proves to be the most successful, although it faces the challenge of trouble due to the dimensionality of microarray data.

3 Proposed Ensemble Learning for Leukemia Classification

In this section, an ensemble learning model proposed for leukemia classification is detailed. First, leukemia microarray data undergoes preprocessing. After preprocessing, a novel method, named SE-SBOA, is developed for feature selection. SE-SBOA is developed by incorporating the EWMA concept into SBOA, optimizing feature subsets to improve model performance and reduce complexity. SBOA has various benefits, such as its effectiveness and simplicity in solving optimization problems of complex nature. It emulates the process of choosing an optimal candidate in recruitment, and thus it is extremely effective for both discrete and continuous optimization problems. Lastly, classification of leukemia is performed using a proposed ensemble-based approach, which combines techniques such as GNN, MLP and Random Forest. Through the integration of GNN, MLP and Random Forest, the strengths of each are utilized to bring about improved accuracy. This ensemble method enhances classification performance, reduces overfitting, and improves generalization. Structural representation of the proposed model is described by Figure 1.

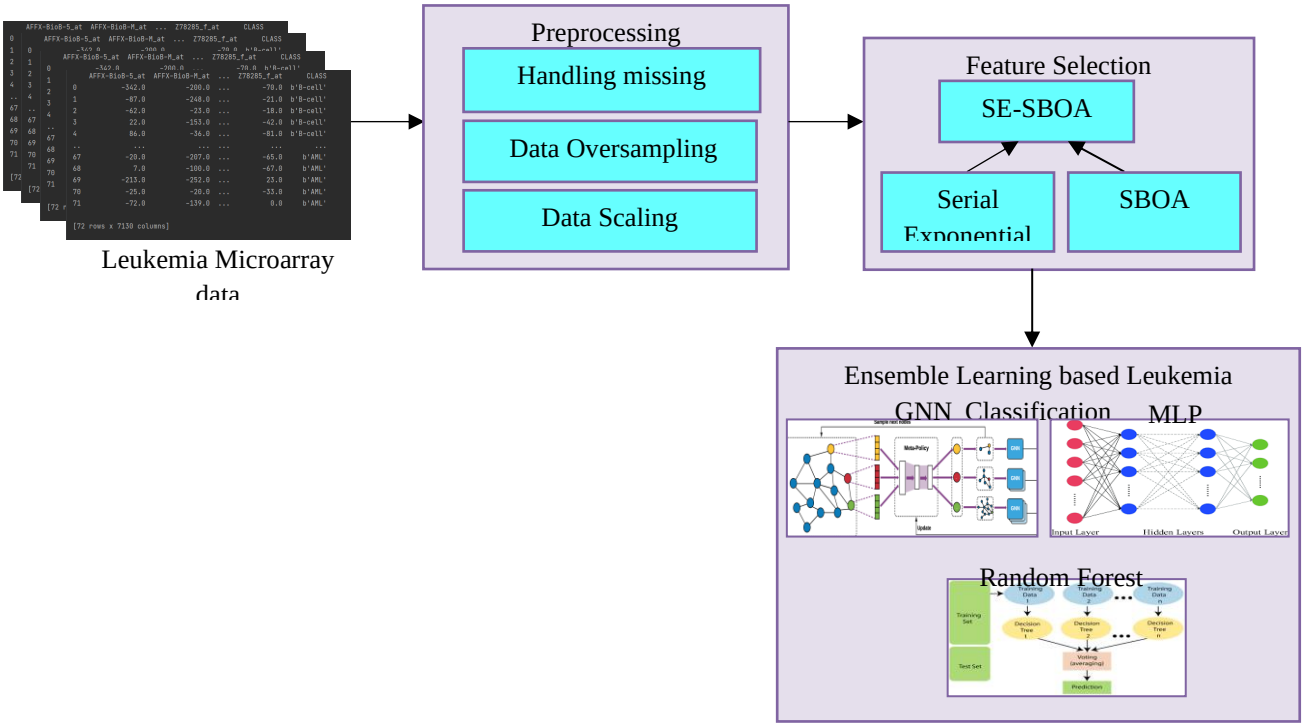


Figure 1: Diagrammatic depiction of SE-SBOA-based Ensemble Learning for Leukemia Classification

3.1 Pre-processing the Microarray Data

Pre-processing involves cleaning and converting raw data to make it suitable for assessment. It is a critical step in data analysis. Handling missing values, noise reduction, and normalizing variables in microarray data ensures accuracy. This process is required to enhance the predictive model and derive meaningful insights. Data scaling, data oversampling and missing value imputation are the three primary pre-processing techniques employed in this study.

3.1.1 Handling the Missing Data

Various factors, such as measurement inaccuracies, insufficiently advanced measurement instruments, equipment malfunctions, data corruption and analogous complications, may lead to a dataset exhibiting a restricted quantity of absent values. The presence of these absent values may introduce bias, which has the potential to diminish the overall effectiveness of predictive models. Plotting missing values using various methods allows for the creation of precise predictions, and this method is usually adaptable and successful in resolving problems involving missing data. In this approach, the arithmetic mean of remaining non-null values is used to substitute each absent value corresponding to a specific attribute. Equation (1) delineates a process for calculating mean value, thereby facilitating replacement of missing values.

$$Mean(A) = \frac{\sum_{w=0}^{H_e} A_w}{H_e} \quad (1)$$

Here, A_w indicates data in a row and H_e states overall data.

3.1.2 Data Oversampling

Data oversampling involves increasing the size of the training dataset to advance performance, particularly when dealing with imbalanced classes. A common method used for oversampling is bootstrapping-based oversampling (Shorten et al., 2021), where innovative samples are created by recurrently sampling from the original dataset with replacement. This technique ensures that the resulting dataset contains duplicate instances or samples not originally present, providing more data for the model. Bootstrapping helps reduce overfitting and enhances the simplification instinct of the model by diversifying training data without requiring the collection of additional data. It is an effective strategy for improving model accuracy. Before data augmentation, the dataset size for the Leukemia_3c dataset was 72 x 7130. After augmentation, the dataset size increased to 3000 x 7130. For the GSE28497 dataset, the original data size was 281x22285, and following augmentation, the dataset size grew to 7000 x 22285.

3.1.3 Data Scaling

The data is standardized by using StandardScaler (Li et al., 2015). This methodology presupposes that the input variables adhere to a normal distribution and subsequently standardizes these variables to attain a mean of 0 and a standard deviation of 1. It calculates the mean and standard deviation for each distinct feature and then modifies the feature by these computations:

$$S_d = \frac{Z - \vartheta}{\rho} \quad (2)$$

where υ signifies mean, Z implies actual value of a feature in the dataset before scaling and ρ states standard deviation.

3.2 Serial Exponential-Secretary Bird Optimization Algorithm-based Feature Selection

The methodology of determining suitable characteristics to reduce input variables of a model through application of relevant data and data assembly techniques is referred to as feature selection. In this, EWMA (Saccucci et al., 1992) and SBOA (Ananthi et al., 2025) are incorporated to form the SE-SBOA framework for the purpose of executing feature selection. EWMA offers adaptability, flexibility and

computational effectiveness, whereas the SBOA augments optimization by efficiently balancing exploration and exploitation, ensuring global convergence and adaptability across diverse, high-dimensional problems. This combination may improve convergence speed, exploration-exploitation balance, robustness, adaptability, search effectiveness and stability in complex optimization problems. The steps involved are as follows:

3.2.1 Solution Encoding

Solution encoding specifies preferred matrix order. For optimization, solution dimension is the quantity of variables or decision-making components that are concurrently modified to attain improved results. The presence of a feature is indicated in this context by its representation as a vector with values 0 to 1. The position of an active feature is indicated by a value of 1. Figure 2 shows the encoding of solution.

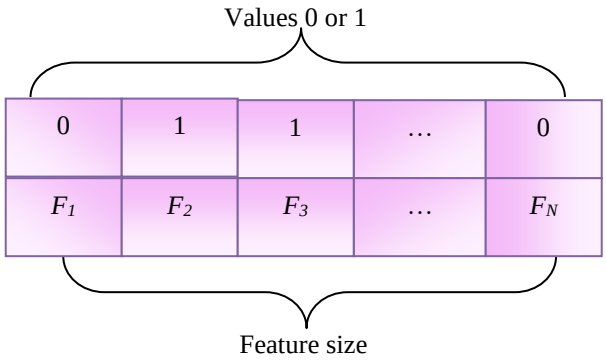


Figure 2: Solution Encoding.

3.2.2 Fitness Function

Accuracy of selected data is considered a noteworthy attribute of neural networks. The accuracy can be improved as demonstrated below.

$$A_c = \frac{h+e}{h+e+y+x} \quad (3)$$

Where, h implies true positive, e signifies true negative, A_c states accuracy, y and x denotes false positive and false negative.

3.2.3 Detailed description of Serial Exponential-Secretary Bird Optimization Algorithm

The SE-SBOA integrates EWMA (Saccucci et al., 1992) with SBOA (Ananthi et al., 2025) to improve optimization performance by combining statistical analysis with the bird's hunting strategy. SBOA enhances the flexibility and efficiency. EWMA helps prioritize recent data, enhancing trend analysis and improving predictive accuracy. By combining these two techniques, SE-SBOA enhances accuracy, exploration, and trend analysis efficiency.

Phase 1: Initialization

The SBOA involves setting up the problem parameters and defining the optimization goals. It begins with initializing a population of candidate solutions, which are then evaluated based on a predefined

fitness function. The algorithm prepares for the selection process by determining key factors, such as the number of iterations, termination criteria, and the range of solution variables. This phase lays the groundwork for the algorithm to efficiently explore and exploit solution space. The initialization procedure of SBOA is illustrated in Equation (4).

$$F = Lowerb + E \times (Upperb - Lowerb) \quad (4)$$

Where $Lowerb$ shows lower bound, $Upperb$ shows upper bound, and E is a random count ranging from 0 to 1.

Phase 2: Hunting strategy of secretary bird

Predatory conduct exhibited by secretary birds during ingestion of serpents is characteristically classified into three different states: exploration for prey, attack on prey, and eating of prey.

a) Searching for Prey

Hunting strategy of secretary bird comprises a unique approach where the bird actively hunts on foot, using its long legs to strike prey. It primarily targets insects, small vertebrates, and reptiles. The bird spots prey from a distance, then swiftly approaches, delivering powerful kicks to subdue its catch. This strategy highlights agility, precision, and the use of visual and physical prowess to capture prey.

$$while\ l < \frac{1}{3}L, \qquad F^{l+1} = F^{l} + (F_{E1} - F_{E2}) \times R_d \quad (5)$$

where L implies total number of iterations. To advance the performance of SBOA, EWMA (Saccucci et al., 1992) is employed, which is stated as follows,

$$F_J^l = \upsilon \times F^l + (1-\upsilon) \times F_J^{l-1} \quad (6)$$

Rewriting weights up to $(l-2)^{th}$ period, which is why serial EWMA is obtained.

$$F_J^l = \upsilon \times F^l + (1-\upsilon) \times (\upsilon * F^{l-1} + (1-\upsilon) \times (\upsilon \times F^{l-2} + (1-\upsilon) \times F_J^{l-3})) \quad (7)$$

$$F^l = \frac{1}{\upsilon} \times [F_J^l + (1-\upsilon) \times (\upsilon \times F^{l-1} + (1-\upsilon) \times (\upsilon \times F^{l-2} + (1-\upsilon) \times F_J^{l-3}))] \quad (8)$$

Assuming that solutions of EWMA (Saccucci et al., 1992) and SBOA (Ananthi et al., 2025) are corresponding at the current iteration, equation (8) is substituted into equation (5).

$$F^l = \left\{\frac{1}{\upsilon} \times [F_J^l + (1-\upsilon) \times (\upsilon \times F^{l-1} + (1-\upsilon) \times (\upsilon \times F^{l-2} + (1-\upsilon) \times F_J^{l-3}))]\right\} + (F_{E1} + F_{E2}) \times R_d \quad (9)$$

Wherein, F_{E1} and F_{E2} are the random candidate solution, R_d is the randomly created array in interval [0, 1], F^l denotes current position of candidate at timestep l, F^{l+1} represents the updated position at timestep l, and υ signifies smoothing factor.

b) Consuming Prey

Consuming prey refers to the act of selecting candidates (prey) during the selection process. The goal is to maximize the chance of choosing the best candidate by first "scanning" a portion of available candidates without choosing, which is similar to observing potential prey. After this observation phase, the next best candidate encountered is selected. Algorithm aims to balance between exploration (observing candidates) and exploitation (seizing the optimal candidate), and consuming prey optimally ensures the highest chance of success.

c) Attacking Prey

Attacking prey refers to the decision-making step where, after a certain observation phase, the algorithm chooses a candidate (prey) based on prior information. During the observation phase, a fixed number or percentage of candidates are reviewed without selection. Once this phase concludes, the next candidate that exceeds the qualities of any previously observed candidates is selected, symbolizing the "attack" on the optimal choice. This strategy aims to maximize the probability of acquiring the best candidate by balancing exploration (searching for the best option) and exploitation (seizing the best available option).

Phase 3: Secretary bird's escape tactic (exploitation stage)

The exploitation stage replicates the decision-making process of a secretary bird. After exploring a portion of candidates without selecting (the observation phase), the algorithm moves into the exploitation stage, where it must choose the best available option. The escape strategy involves choosing the first candidate that surpasses all previous ones encountered, ensuring the algorithm "escapes" suboptimal choices and maximizes the likelihood of selecting the best candidate. This stage emphasizes swift action and optimal decision-making.

Phase 4: Termination

This iterative process repeats until the optimal solution is reached.

3.3 Ensemble Learning based Leukemia Classification

Soft voting represents an ensemble learning methodology whereby multiple classifiers generate probability distributions for each class, culminating in a final prediction derived from averaging of these probabilities, ultimately selecting the class with highest mean probability, which frequently enhances accuracy and generalization relative to hard voting. Ensemble techniques are ML approaches that initiate with a collection of base mechanisms and amalgamate them to fabricate a more resilient and effective approach. When specific conditions, such as model diversity, are fulfilled, an ensemble typically surpasses the performance of individual base models. It is crucial to acknowledge that formulation of these base models is not constrained by limitations, implying that they need not be homogeneous. Ensemble methods can be classified into two types, predicated on their training techniques: parallel and sequential (Manconi et al., 2022).

Sequential Ensemble Approaches:

At each step, a novel base classifier is introduced, focusing on correcting the residual misclassification errors to improve the existing solution. The boosting algorithms, which are the most widely used sequential ensemble techniques, give extra attention to misclassified examples, increasing their chances

of being incorporated into the training set for any innovative base approach that is established (Houssein et al., 2021).

Parallel Ensemble Mechanisms

In order to enhance autonomy concerning misclassification inaccuracies, foundational models are concurrently trained. Given that amalgamation of models exhibiting diverse error patterns can significantly augment overall efficacy, this autonomy is of paramount importance. A widely recognized parallel ensemble methodology, known as the bagging algorithm, was adopted to yield an ensemble characterized by reduced error variability in comparison to individual base classifiers. To achieve this, a set of distinct models is generated through bootstrap sampling applied to various training datasets.

Blending Policies for Ensemble Methods

Despite the theoretical independence of blending strategies and training methods in practice, the training approach that is selected frequently results in a preference for a particular output blending policy. Following a review of some of the most popular voting techniques, insightful information about how to apply them to sequential and parallel ensemble techniques is given.

Voting

Soft voting and hard voting are two primary methods of voting. The process of soft voting is depicted in Figure 3. Three base models, denoted by the numbers 0 and 1 in this instance, forecast two potential classes. Class with most votes from base models is chosen as concluding forecast in a hard voting process. Two of the three models in this case suggest that class 1 will be the ultimate prediction. To ascertain which class has the highest average probability, soft voting, on the other hand, computes the average class pseudo-probabilities from the base models (Manconi et al., 2022).

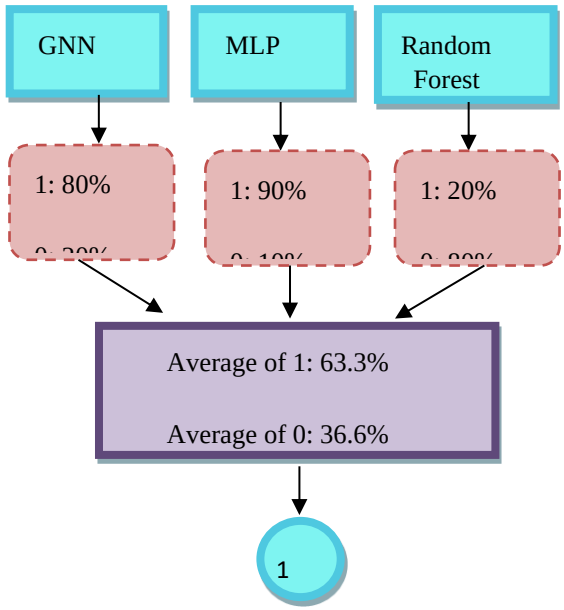


Figure 3: Process of soft voting

An ensemble of three classifiers is utilized to ascertain a class label from a set of k potential labels $\{C_1, C_2, ..., C_\kappa\}$. For a specified input $\wp$, each classifier $\Re_i$ representing i^{th} a model produces a k

dimensional vector $\left[\Re_i^1(\wp), \Re_i^2(\wp), \ldots\ldots, \Re_i^\kappa(\wp)\right]$. In this vector, $\Re_i^j(\wp)$ it signifies output of the i^{th} model for j^{th} category label. Following the process of averaging individual outputs and modifying class labels, overall categorization is established. The label κ^* that has garnered the highest number of votes is subsequently chosen. Each classifier is afforded an equivalent degree of consideration:

$$\kappa^* = Arg\ Max\ \Re(\wp) = ArgMax\ \frac{1}{\hbar} \cdot \sum_{i=1}^{\hbar} \Re_i^j(\wp) = ArgMax \sum_{i=1}^{\hbar} \Re_i^j(\wp) \qquad (10)$$

Soft voting in ensemble learning combines the probabilistic outputs of models like GNN, MLP, and Random Forests to improve Leukemia classification. By averaging predictions, it reduces overfitting and leverages each model's strengths: GNN for graph data, MLP for complex patterns, and Random Forests for robustness, enhancing overall forecast reliability.

3.3.1 GNN

GNN (Jiang & Luo, 2022) constitute a specialized category of neural networks adept at processing data organized in graph structures, wherein nodes symbolize entities and edges denote interrelations among these entities. GNN exhibit considerable efficacy in various tasks, including but not limited to node classification, graph classification and link prediction. They operate by passing messages between nodes and aggregating information from their neighbors to learn node or graph-level representations. The core operation in GNNs is graph convolution, where node features are updated by aggregating information from neighboring nodes. Mathematical formulation of Graph is represented as:

$$U = (B, J, T) \qquad (11)$$

wherein B denotes a set of node J s, implies a set of edges connecting nodes and T represents an adjacency matrix representing connections between nodes. Each node b_r has a set of neighboring nodes, denoted as $E(b_r)$. The degree matrix P_{rr} is defined as:

$$P_{rr} = = |E(b_r)| \qquad (12)$$

The Laplacian matrix of a graph is demonstrated as:

$$Q = P - T \qquad (13)$$

Normalized version of the Laplacian matrix is given by,

$$\tilde{Q} = M = P^{-1/2} \times TP^{-1/2} \qquad (14)$$

Herein, M the identity matrix, P is the degree matrix**,** $E(b_r)$which denotes the set of neighboring nodes b_r. Q defines Laplacian matrix, T denotes adjacency matrix**,** $\tilde{Q}$ represents normalized Laplacian matrix**,** $P^{-1/2}$ implies inverse square root of degree matrix**,** which is used to normalize the influence of nodes with different degrees. GCNs apply convolutional operations to graph structures using spectral graph theory. The graph convolution operation is defined as:

$$A_{*U} = Y\left(P^{-1/2} \times TP^{-1/2}\right) \times A \qquad (15)$$

Here, Y signifies learnable weight matrix, A represents the node feature matrix, and A_{*U} signifies graph convolution operation applied to node features A. To stabilize learning and prevent gradient explosion, an improved form is:

$$A_{*U} = Y\left(\tilde{P}^{-1/2} \times \tilde{T}\tilde{P}^{-1/2}\right) \times A \qquad (16)$$

Wherein, $\tilde{T} = T + M$ the self-loop is added to include the node's features $\tilde{P}$ and the corresponding degree matrix $\tilde{P}_{rr} = \sum_u \tilde{T}_{ur}$. Figure 4 represents GNN architecture.

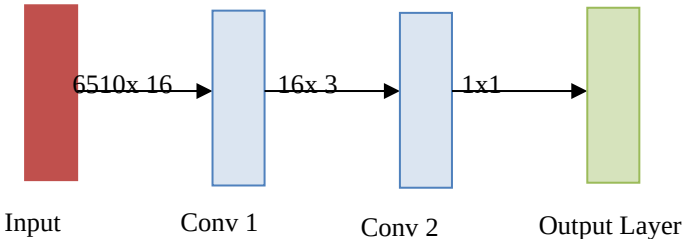


Figure 4: GNN architecture

3.3.2 MLP

MLP (Massaoudi et al., 2021) represents a specific category of ANN characterized by several layers of interconnected nodes, commonly referred to as neurons. The architectural framework of an MLP is composed of an input layer, one or more intermediary hidden layers and a final output layer. Each neuron in a layer has connections to neurons in the layers that surround it, and these connections have weights attached to them. The MLP learns by modifying these weights in response to the input data and the associated output during the training phase. The network can model complex relationships in data by introducing non-linearity through the use of activation functions like sigmoid, ReLU or tanh. A technique known as backpropagation, which recursively upgrades weights to diminish error between actual and predicted output, is normally used to train MLPs using supervised learning algorithms. The MLP can discover patterns and forecast outcomes using the given data thanks to this process. The arithmetical illustration of MLP is:

$$Z = \varphi_T\left\{\sum_{g=1}^{f} \aleph_\Gamma\left[\varphi_\Omega\left(\sum_{g=1}^{N} \aleph_\Omega E_g\right)\right]\right\} \qquad (17)$$

Here, $\aleph_\Gamma$ $\aleph_\Omega$ serve as representations of weights corresponding to hidden and output layers, φ_Γ and φ_Ω denote activation functions relevant to hidden and output layers, whereas E_g signifies input parameters, z implies output of MLP, N represents total number of input neurons and f represents total number of neurons in output layer. Figure 5 represents MLP architecture.

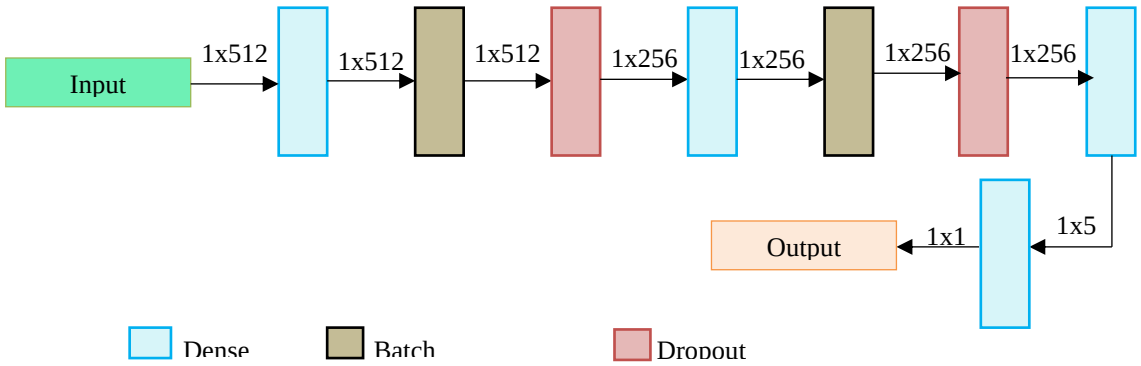


Figure 5: Architecture of MLP

3.3.3 Random Forest

Random Forest classification is a supervised learning technique employed to classify data points into given classes based on decision trees. A Random Forest classifier (Jaiswal & Samikannu, 2017) works by building many decision trees from training data subsets and combining their predictions to enhance accuracy and minimize overfitting. Every tree individually predicts a class label, and final classification is determined using a majority voting scheme. For instance, in a medical diagnosis system, Random Forest can predict whether a tumor is malignant or benign based on characteristics such as size, shape, and texture. Efficacy of Random Forest is attributed to its capacity to adeptly manage data characterized by high dimensionality while maintaining robustness against noise and missing values. The construction of a classification tree in a Random Forest involves splitting data into branches based on class variables. The goal is to create pure nodes, meaning each node contains data from a single class. The process starts with impure nodes and gradually refines them using relative class frequencies.

$$\varpi(\wp) = \varphi(\hbar_1, \hbar_2, \ldots\ldots, \hbar_z \qquad (18)$$

Here, $\hbar_z$ indicates relative frequencies of z diverse classes at the considered node, and φ is a symmetric function of $(\hbar_1, \hbar_2, \ldots, \hbar_z)$. Various algorithms for tree implementation adhere to distinct metrics of inaccuracy. These metrics are strategically chosen to mitigate the degree of inaccuracy that split induces at a specified node. The effectiveness of a split at a node f is determined by:

$$\Lambda\varpi(e, f) = \varpi(f) - \ell(\kappa) \times \varpi(\kappa) - \ell(o) \times \varpi(o) \qquad (19)$$

Here, $\Lambda\varpi(e, f)$ represents the change in impurity (or error) due to splitting at node f, $\varpi(f)$ signifies impurity (or inaccuracy measure) of the current node f before splitting, $\ell(\kappa)$ denotes proportion of data samples that go to left child node after split, $\varpi(\kappa)$implies impurity measure of left child node after split, $\ell(o)$implies proportion of data samples that go to the right node after the split and $\varpi(o)$signifies impurity measure of right child node after the split.

3.3.4 Ensemble learning by soft voting

Soft voting in ensemble learning, incorporating models such as GNN, MLP and Random Forests, enhances leukemia classification. Each model outputs probabilistic predictions, and the final class label is determined by averaging these probabilities, selecting the class with the highest average. This method

leverages strengths of each model: GNNs excel at processing graph-based data, MLPs are effective at modeling complex non-linear relationships, and Random Forests are resilient to overfitting. By combining these models, soft voting reduces classification errors and improves overall performance in leukemia diagnosis. The approach strengthens the ensemble's robustness, providing valuable insights into the confidence of each model, allowing for better understanding of their strengths and weaknesses. Ultimately, soft voting improves the accuracy and reliability of leukemia classification, leading to more precise medical decisions and optimized healthcare outcomes.

4 Results

Outcome based on suggested scheme is outlined in this section. To implement, a Python tool is used. For measuring how good the presented methodology is, four measures are employed: accuracy, recall, precision and F1-Score.

4.1 Dataset Description

GSE28497 dataset comprises gene expression data for leukemia aimed at discovering molecular signatures that are linked to different leukemia subtypes. It is obtained through high-throughput microarray experiments and applies to gene expression analysis in the field of bioinformatics. Scientists use this dataset for subtype classification, biomarker discovery, and elucidating molecular mechanisms of leukemia. It contributes to the advancement of diagnostic instruments and therapeutic measures for leukemia patients.

Leukemia_3c dataset has gene appearance information for the classification of leukemia. It finds application in biomedical research for classifying various kinds of leukemia depending on gene expression profiles. Dataset is generally applied to ML functions such as classification, to efficiently classify leukemia subtypes, resulting in better diagnostics and treatments within medical environments. It is applied extensively in cancer research and bioinformatics applications.

Figure 6 illustrates visualizations of the GSE28497 and Leukemia_3c datasets. Figures 6 (a) and 6 (b) show the original data from each dataset. The GSE28497 dataset originally consists of 281 × 22,285 data, whereas the Leukemia_3c dataset is 72 × 7,130. Figure 6 (c) and 6 (d) depict augmented data to enhance the data for processing. After the augmentation, the GSE28497 dataset expands in size to 7,800 × 22,285, and the Leukemia_3c dataset increases to 3,000 × 7,130. Figures 6 (e) and 6 (f) present the selected features for each dataset, with GSE28497 containing 7,800 × 18,344 and Leukemia_3c comprising 3,000 × 6,511. These visualizations highlight the most relevant variables for analysis and classification tasks, demonstrating the effectiveness of augmentation and feature selection.

GSE28497 dataset

	samples	type	...	AFFX-TrpnX-5_at	AFFX-TrpnX-M_at
0	GSM705467.CEL.gz	B-CELL_ALL	...	2.860505	2.608381
1	GSM705468.CEL.gz	B-CELL_ALL	...	2.908509	2.634063
2	GSM705469.CEL.gz	B-CELL_ALL	...	2.626871	2.673293
3	GSM705470.CEL.gz	B-CELL_ALL	...	2.762835	2.624163
4	GSM705471.CEL.gz	B-CELL_ALL	...	2.748511	2.738165
..	...	...	...	...	...
276	GSM706002.CEL.gz	B-CELL_ALL_ETV6-RUNX1	...	2.705429	2.762513
277	GSM706003.CEL.gz	B-CELL_ALL_ETV6-RUNX1	...	2.942468	2.666361
278	GSM706004.CEL.gz	B-CELL_ALL_ETV6-RUNX1	...	2.811914	2.697337
279	GSM706005.CEL.gz	B-CELL_ALL_ETV6-RUNX1	...	2.710848	2.769975
280	GSM706007.CEL.gz	B-CELL_ALL_ETV6-RUNX1	...	2.698757	2.753662

[281 rows x 22285 columns]

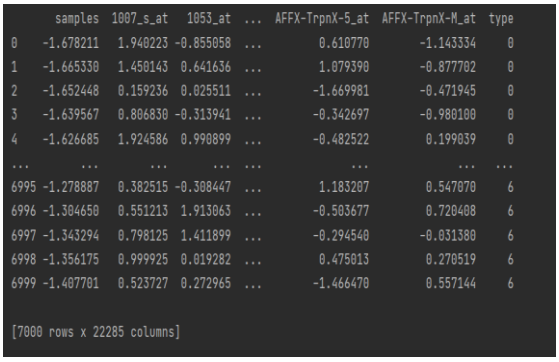
(a)

Leukemia_3c dataset

	AFFX-BioB-5_at	AFFX-BioB-M_at	...	Z78285_f_at	CLASS
0	-342.0	-200.0	...	-70.0	b'B-cell'
1	-87.0	-248.0	...	-21.0	b'B-cell'
2	-62.0	-23.0	...	-18.0	b'B-cell'
3	22.0	-153.0	...	-42.0	b'B-cell'
4	86.0	-36.0	...	-81.0	b'B-cell'
..	...	...	...	...	...
67	-20.0	-207.0	...	-65.0	b'AML'
68	7.0	-100.0	...	-67.0	b'AML'
69	-213.0	-252.0	...	23.0	b'AML'
70	-25.0	-20.0	...	-33.0	b'AML'
71	-72.0	-139.0	...	0.0	b'AML'

[72 rows x 7130 columns]

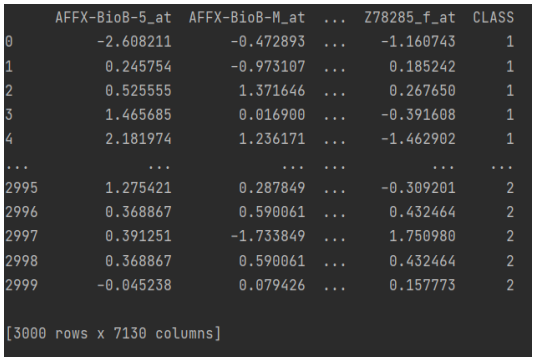
(b)



	samples	1007_s_at	1053_at	...	AFFX-TrpnX-5_at	AFFX-TrpnX-M_at	type
0	-1.678211	1.940223	-0.855058	...	0.610770	-1.143334	0
1	-1.665330	1.450143	0.641636	...	1.079390	-0.877702	0
2	-1.652448	0.159236	0.025511	...	-1.669981	-0.471945	0
3	-1.639567	0.806830	-0.313941	...	-0.342697	-0.980100	0
4	-1.626685	1.924586	0.990899	...	-0.482522	0.199039	0
...	...	...	...	...	...	...	...
6995	-1.278887	0.382515	-0.308447	...	1.183207	0.547070	6
6996	-1.304650	0.551213	1.913063	...	-0.503677	0.720408	6
6997	-1.343294	0.798125	1.411899	...	-0.294540	-0.031380	6
6998	-1.356175	0.999925	0.019282	...	0.475013	0.270519	6
6999	-1.407781	0.523727	0.272965	...	-1.466470	0.557144	6

[7000 rows x 22285 columns]

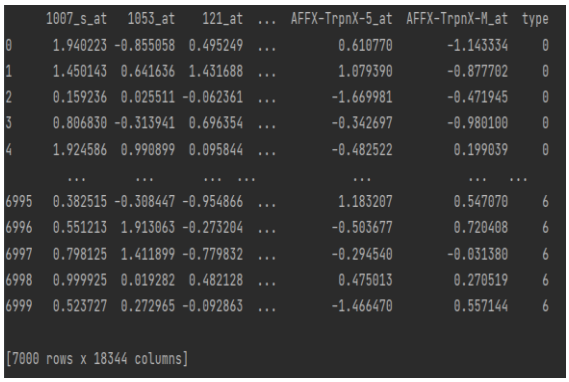
(c)



	AFFX-BioB-5_at	AFFX-BioB-M_at	...	Z78285_f_at	CLASS
0	-2.608211	-0.472893	...	-1.160743	1
1	0.245754	-0.973107	...	0.185242	1
2	0.525555	1.371646	...	0.267650	1
3	1.465685	0.016900	...	-0.391608	1
4	2.181974	1.236171	...	-1.462902	1
...	...	...	...	...	...
2995	1.275421	0.287849	...	-0.309201	2
2996	0.368867	0.590061	...	0.432464	2
2997	0.391251	-1.733849	...	1.750980	2
2998	0.368867	0.590061	...	0.432464	2
2999	-0.045238	0.079426	...	0.157773	2

[3000 rows x 7130 columns]

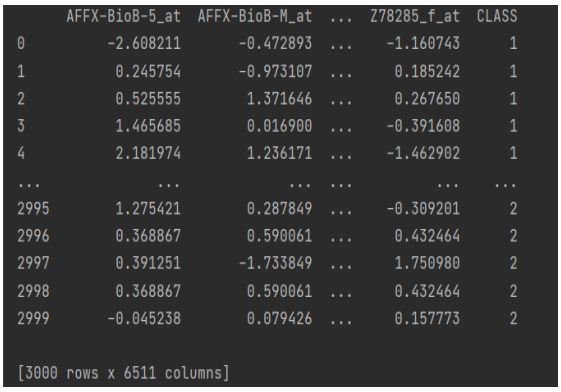
(d)



	1007_s_at	1053_at	121_at	...	AFFX-TrpnX-5_at	AFFX-TrpnX-M_at	type
0	1.940223	-0.855058	0.495249	...	0.610770	-1.143334	0
1	1.450143	0.641636	1.431688	...	1.079390	-0.877702	0
2	0.159236	0.025511	-0.062361	...	-1.669981	-0.471945	0
3	0.806830	-0.313941	0.696354	...	-0.342697	-0.980100	0
4	1.924586	0.990899	0.095844	...	-0.482522	0.199039	0
...	...	...	...	...	...	...	...
6995	0.382515	-0.308447	-0.954866	...	1.183207	0.547070	6
6996	0.551213	1.913063	-0.273204	...	-0.503677	0.720408	6
6997	0.798125	1.411899	-0.779832	...	-0.294540	-0.031380	6
6998	0.999925	0.019282	0.482128	...	0.475013	0.270519	6
6999	0.523727	0.272965	-0.092863	...	-1.466470	0.557144	6

[7000 rows x 18344 columns]

(e)



	AFFX-BioB-5_at	AFFX-BioB-M_at	...	Z78285_f_at	CLASS
0	-2.608211	-0.472893	...	-1.160743	1
1	0.245754	-0.973107	...	0.185242	1
2	0.525555	1.371646	...	0.267650	1
3	1.465685	0.016900	...	-0.391608	1
4	2.181974	1.236171	...	-1.462902	1
...	...	...	...	...	...
2995	1.275421	0.287849	...	-0.309201	2
2996	0.368867	0.590061	...	0.432464	2
2997	0.391251	-1.733849	...	1.750980	2
2998	0.368867	0.590061	...	0.432464	2
2999	-0.045238	0.079426	...	0.157773	2

[3000 rows x 6511 columns]

(f)

Figure 6: Dataset Visualization of GSE28497 and Leukemia_3c datasets, (a) original GSE28497 dataset, (b) original Leukemia_3c dataset, (c) Augmented data from GSE28497 dataset, (d) Augmented data from Leukemia_3c dataset, (e) selected features of GSE28497 dataset, and (f) selected features of Leukemia_3c dataset.

4.1 Performance Metrics

This section elucidates various metrics employed to evaluate the efficacy of the proposed methodology, contrasting it with traditional techniques:

a) Accuracy

Accuracy signifies the degree to which measurements or predictions correspond with actual value, as articulated in equation (3).

b) Precision

Precision measures the proportion of true positive predictions among all positive predictions made by a model.

$$P_R = \frac{h+e}{h+y} \qquad (20)$$

Here, P_R signifies precision.

c) Recall

Recall evaluates the proportion of true positive predictions among all actual positive instances in the dataset.

$$R_c = \frac{h+e}{h+x} \qquad (21)$$

Here, R_c implies recall

d) F1 Score

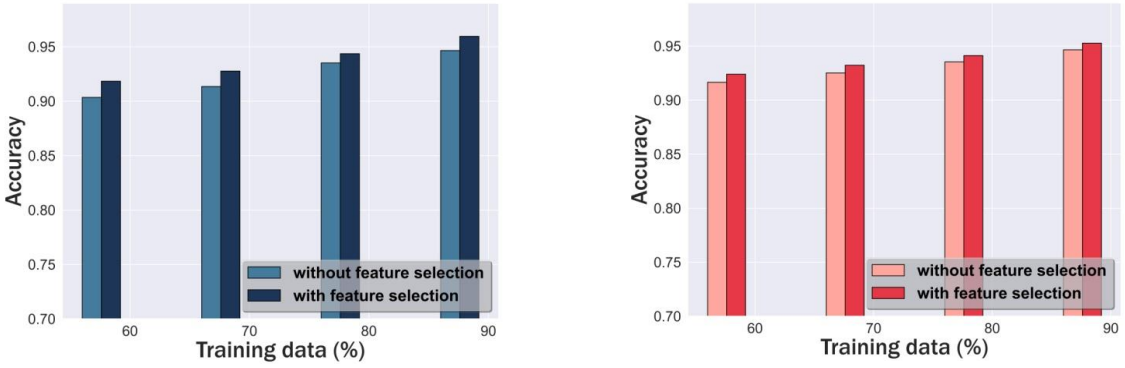
F-measure is the harmonic mean of precision and recall, providing a single metric that balances both concerns.

$$F_S = 2 * \frac{(P_R \times R_C)}{(P_R \times R_C)} \qquad (22)$$

here, F_S implies F1 Score

4.3 Performance Analysis

Figure 7 depicts performance examination of the proposed approach using GSE28497 dataset and Leukemia_3c dataset with and without feature selection. Figure 7(a) shows performance analysis using GSE28497 dataset. For 60% of training data, presented approach achieved an accuracy of 0.903 without feature selection, which improved to 0.918 with feature selection. As training data increased to 70%, accuracy increased further, reaching 0.913 without feature selection and 0.927 with feature selection. With 90% of training data, accuracy reached 0.946 without feature selection, improving to 0.959 with feature selection. Figure 7(b) shows the performance analysis for Leukemia_3c dataset with and without feature selection. For 60% of training data, SE-SBOA-based Ensemble Learning achieved an accuracy of 0.916 without feature selection, which improved to 0.924 with feature selection. As training data increased to 90%, accuracy reached 0.946 without feature selection and further augmented to 0.953 with feature selection.



(a) (b)

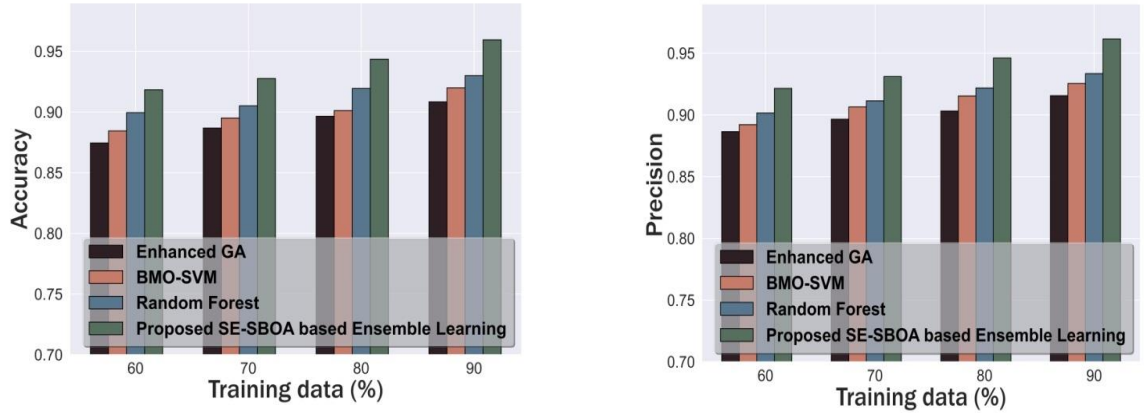
Figure 7: Performance assessment based on (a) GSE28497 dataset and (b) Leukemia_3c dataset

4.4 Comparative Assessment

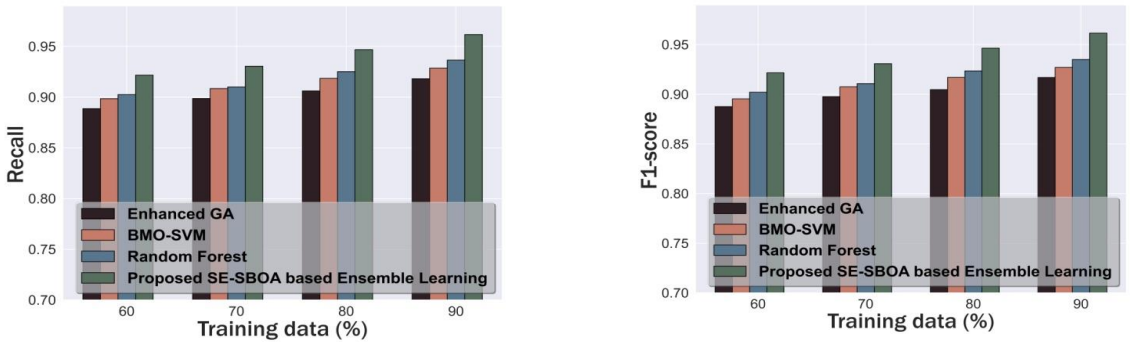
This segment elaborates the assessment of proposed methodology, emphasizing its comparative analysis with conventional techniques, such as Enhanced GA (Hartigan, 2023), BMO-SVM (Mahmoudi et al., 2015) and Random Forest (Dagnew & Shekar, 2021), employing datasets, GSE28497 dataset and Leukemia_3c dataset.

4.4.1 Using GSE28497 dataset

Figure 8 provides a comparative analysis of proposed approach using GSE28497 dataset. In Figure 8(a), accuracy of the SE-SBOA-based Ensemble Learning model is compared to conventional methods. With 90% of the training data, SE-SBOA-based Ensemble Learning model attained an impressive accuracy of 0.959, significantly outperforming conventional methods. In comparison, the Enhanced GA, BMO-SVM, and Random Forest models achieved accuracies of 0.908, 0.919, and 0.930, highlighting the superior performance of the SE-SBOA-based Ensemble Learning model. Figure 8(b) presents precision analysis. Using 60% of training data, SE-SBOA model reached a precision of 0.921, notably outperforming the Enhanced GA, BMO-SVM, and Random Forest models, which had precisions of 0.886, 0.892, and 0.901. When 80% of the training data was used, the SE-SBOA model maintained a high precision of 0.946, demonstrating its stable and exceptional performance across different data sizes. Figure 8(c) illustrates recall metric. With 70% of training data, the SE-SBOA-based Ensemble Learning model executed a recall of 0.931, surpassing recall values of Enhanced GA, BMO-SVM, and Random Forest models, which were 0.898, 0.908, and 0.910. This further illustrates superior performance of presented approach based on of recall. Lastly, comparative assessment of F1-score is shown in Figure 8(d). With 90% of training data, SE-SBOA-based Ensemble Learning model reached a remarkable F1-score of 0.962, outperforming traditional methods such as Enhanced GA, BMO-SVM, and Random Forest, which recorded F1-scores of 0.916, 0.927, and 0.935.



(a) (b)

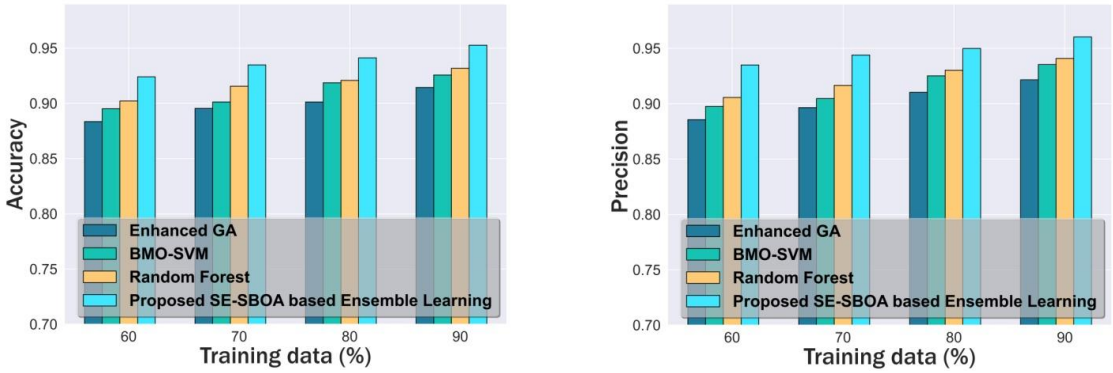


(c) (d)

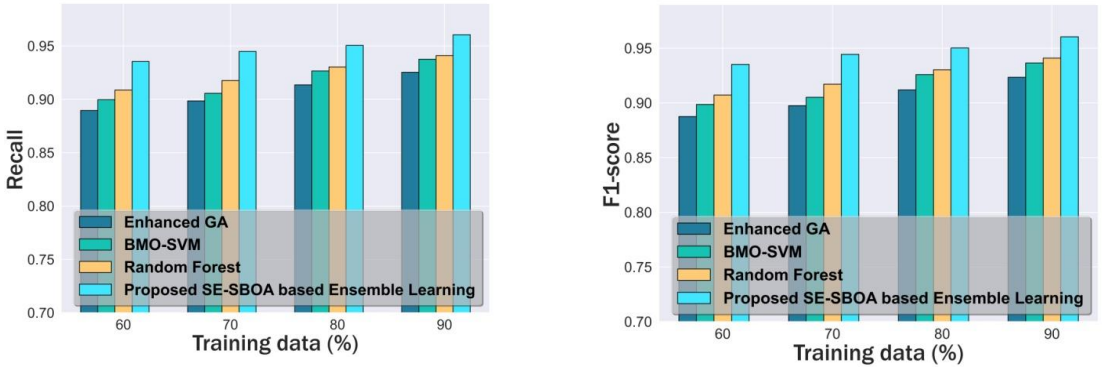
Figure 8: Comparative examination for (a) accuracy, (b) precision, (c) Recall and (d) F1-score

4.4.2 Using Leukemia_3c dataset

Figure 9 provides a comparative analysis of presented approach using the Leukemia_3c dataset, where the SE-SBOA-based Ensemble Learning model is compared with conventional approaches. In Figure 9(a), accuracy of SE-SBOA-based Ensemble Learning model is evaluated against conventional methods. With 80% of the training data, the SE-SBOA-based Ensemble Learning model attained an improved accuracy of 0.941, surpassing conventional mechanisms like, Enhanced GA, BMO-SVM, and Random Forest, which recorded accuracies of 0.901, 0.918, and 0.921. Figure 9(b) illustrates precision analysis. With 70% of the training data, the SE-SBOA-based Ensemble Learning model achieved a precision of 0.944, outperforming precision of BMO-SVM with 0.904. As the training data increased to 90%, the model demonstrated excellent precision, reaching 0.960, surpassing Enhanced GA and Random Forest, have precision scores of 0.921 and 0.941. Figure 9(c) presents recall metric. With 60% of the training data, the SE-SBOA-based Ensemble Learning model achieved a recall score of 0.935, outperforming Enhanced GA (0.889), BMO-SVM (0.899), and Random Forest (0.908). As the training data increased to 90%, model maintained a high recall score of 0.961. Finally, Figure 9(d) displays comparative F1-score analysis. With 90% of training data, SE-SBOA-based Ensemble Learning model achieved an impressive F1-score of 0.960, outperforming conventional approaches, stated F1-scores of 0.923 for Enhanced GA, 0.936 for BMO-SVM, and 0.941 for Random Forest.



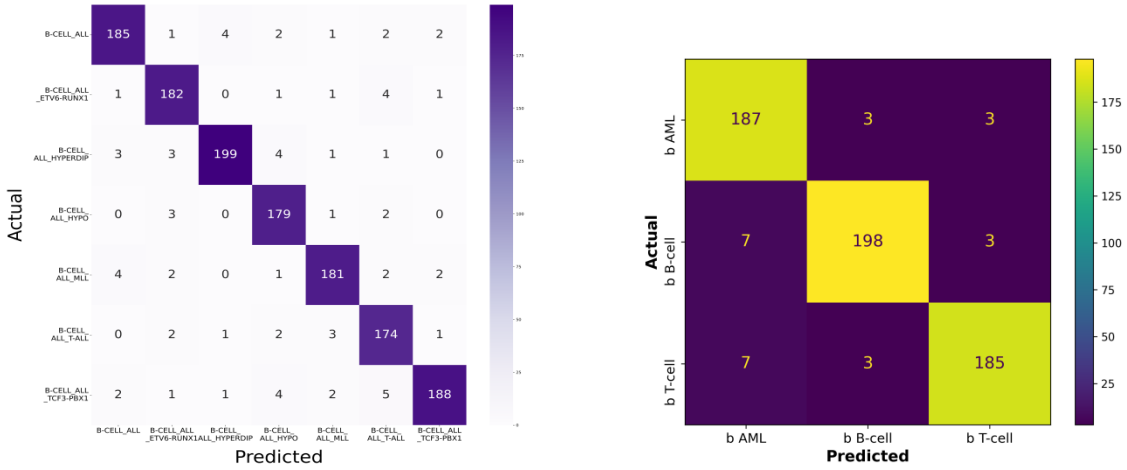
(a) (b)



(c) (d)

Figure 9: Comparative evaluation for (a) accuracy, (b) precision, (c) Recall and (d) F1-score

Figure 10 delineates confusion matrix, which elucidates anticipated class labels positioned in columns and actual class labels arranged in rows. In Figure 10(a), confusion matrix corresponding to the SE-SBOA-based Ensemble Learning model, utilizing GSE28497 dataset is illustrated. The matrix is structured to accommodate two distinct labels, namely actual and predicted. In this instance, testing dataset comprises 1361 samples. Diagonal elements signify instances that were accurately predicted. Among 1361 samples, 1288 were classified correctly, culminating in an accuracy rate of 95.9% for the SE-SBOA-based Ensemble Learning model. Figure 10(b) presents confusion matrix for Leukemia_3c dataset. Testing dataset includes a total of 596 samples. Diagonal elements within this matrix denote correctly predicted instances, as illustrated in graphical analysis. From 596 samples, 570 were classified accurately, resulting in an excellent accuracy rate of 94.1% for the SE-SBOA-based Ensemble Learning model.



(a) (b)

Figure 10: Confusion matrix for (a) GSE28497 dataset and (b) Leukemia_3c dataset

Table 1 provides a comparative evaluation of presented methodologies alongside conventional techniques like Enhanced GA, BMO-SVM and Random Forest. The proposed SE-SBOA-based Ensemble Learning approach outperformed existing methods, delivering superior results in accuracy, precision, recall, and F1-score. In terms of accuracy, it achieved a 3.02% improvement over Random Forest. Compared to BMO-SVM, the proposed method demonstrated a 3.74% increase in precision. Additionally, its recall performance surpassed that of Enhanced GA by 4.54%. Moreover, the F1-score of SE-SBOA-based Ensemble Learning exhibited a 2.8% enhancement over Random Forest, highlighting its efficiency in classification tasks.

Table 1: Comparative discussion

Methods	Enhanced GA	BMO-SVM	Random Forest	Proposed SE-SBOA-based Ensemble Learning
Accuracy	0.908	0.919	0.930	**0.959**
Precision	0.916	0.925	0.934	**0.961**
Recall	0.918	0.928	0.936	**0.962**
F1-score	0.917	0.927	0.935	**0.962**

5 Conclusion

This study aimed to develop a leukemia classification model using optimized DL techniques. Initially, leukemia microarray data underwent a preprocessing phase, which involved handling missing values, applying data oversampling and performing standard scaling. Subsequently, feature selection was performed using the suggested SE-SBOA, an optimization method that implemented EWMA in SBOA. This method allowed the selection of best feature subsets, thus resulting in improved model performance and lower computational complexity. Last but not least, leukemia classification was conducted using the suggested ensemble-based classification, which combined GNN, MLP, and Random Forest. With an integration of GNN, MLP, and Random Forest strengths, the model proved to have a higher classification accuracy, showing efficacy in leukemia classification. Empirical testing verifies the efficacy of the proposed methodology as shown by accuracy levels of 95.9%, precision level of 96.1%, recall level of 96.2%, and an F1-score of 96.2%. These results indicate that SE-SBOA-based ensemble model outperforms conventional classification methods, thereby ensuring greater accuracy and efficacy in leukemia diagnosis. The proposed model would be improved in the future by combining sophisticated DL architectures to achieve improved classification accuracy. The methodology would also be expanded to large-scale and diversified leukemia datasets for achieving wider applicability and stability.

References

[1] Agarwal, S., & Singh, V. (2024). Visual Terminology Aids in Diagnostic Imaging: Improving Radiology Report Accessibility. *Global Journal of Medical Terminology Research and Informatics*, *2*(3), 16-19.

[2] Alabdulqader, E. A., Alarfaj, A. A., Umer, M., Eshmawi, A. A., Alsubai, S., Kim, T. H., & Ashraf, I. (2024). Improving prediction of blood cancer using leukemia microarray gene data and Chi2 features with weighted convolutional neural network. *Scientific Reports*, *14*(1), 15625. https://doi.org/10.1038/s41598-024-65315-7

[3] Ananthi, A., Subathra, M. S. P., George, S. T., Peter, G., Stonier, A. A., & Sairamya, N. J. (2025). A fusion wavelet-based binary pattern approach for enhanced electroencephalogram signal classification. *Computers and Electrical Engineering, 123,* 110019. https://doi.org/10.1016/j.compeleceng.2024.110019

[4] Bilen, M., Işik, A. H., & Yiğit, T. (2020). A new hybrid and ensemble gene selection approach with an enhanced genetic algorithm for classification of microarray gene expression values on leukemia cancer. *International Journal of Computational Intelligence Systems*, *13*(1), 1554-1566. https://doi.org/10.2991/ijcis.d.200928.001

[5] Dagnew, G., & Shekar, B. H. (2021). Ensemble learning-based classification of microarray cancer data on tree-based features. *Cognitive Computation and Systems*, *3*(1), 48-60. https://doi.org/10.1049/ccs2.12003

[6] Dutta, M., Mojumdar, M. U., Kabir, M. A., Chakraborty, N. R., Siddiquee, S. M. T., & Abdullah, S. (2025). LEU3: An Attention Augmented-Based Model for Acute Lymphoblastic Leukemia Classification. *IEEE Access*. doi: 10.1109/ACCESS.2025.3542609.

[7] Ghaderzadeh, M., Asadi, F., Hosseini, A., Bashash, D., Abolghasemi, H., & Roshanpour, A. (2021). Machine learning in detection and classification of leukemia using smear blood images: a systematic review. *Scientific Programming*, *2021*(1), 9933481. https://doi.org/10.1155/2021/9933481

[8] Gowin, K., Skerget, S., Keats, J. J., Mikhael, J., & Cowan, A. J. (2021). Plasma cell leukemia: a review of the molecular classification, diagnosis, and evidenced-based treatment. *Leukemia Research*, *111*, 106687. https://doi.org/10.1016/j.leukres.2021.106687

[9] Gupta, S., Gupta, M. K., Shabaz, M., & Sharma, A. (2022). Deep learning techniques for cancer classification using microarray gene expression data. *Frontiers in physiology*, *13*, 952709.

[10] Hartigan, P. (2023). Diabetic diet essentials for preventing and managing chronic diseases. *Clinical Journal for Medicine, Health and Pharmacy*, *1*(1), 16-31.

[11] Hoffmann, K., Firth, M. J., Beesley, A. H., de Klerk, N. H., & Kees, U. R. (2006). Translating microarray data for diagnostic testing in childhood leukaemia. *BMC cancer*, *6*(1), 229. https://doi.org/10.1186/1471-2407-6-229

[12] Houssein, E. H., Abdelminaam, D. S., Hassan, H. N., Al-Sayed, M. M., & Nabil, E. (2021). A hybrid barnacles mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification. *IEEE Access*, *9*, 64895-64905. doi: 10.1109/ACCESS.2021.3075942

[13] Jaiswal, J. K., & Samikannu, R. (2017, February). Application of random forest algorithm on feature subset selection and classification and regression. In *2017 world congress on computing and communication technologies (WCCCT)* (pp. 65-68). Ieee. doi: 10.1109/WCCCT.2016.25.

[14] Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert systems with applications*, *207*, 117921. https://doi.org/10.1016/j.eswa.2022.117921

[15] Jothiaruna, N. (2022). SSDMNV2-FPN: A cardiac disorder classification from 12 lead ECG images using deep neural network. *Microprocessors and Microsystems*, *93*, 104627. https://doi.org/10.1016/j.micpro.2022.104627

[16] Krishnan, R., & Dandekar, G. (2022). Medical Image Analysis for Disease Detection and Categorisation Using Deep Learning Algorithm. *International Academic Journal of Science and Engineering*, *9*(3), 16–20.

[17] Labaj, W., Papiez, A., Polanski, A., & Polanska, J. (2017). Comprehensive analysis of MILE gene expression data set advances discovery of leukaemia type and subtype biomarkers. *Interdisciplinary Sciences: Computational Life Sciences*, *9*(1), 24-35. https://doi.org/10.1007/s12539-017-0216-9

[18] Li, X., Yang, T., Li, S., Jin, L., Wang, D., Guan, D., & Ding, J. (2015). Noninvasive liver diseases detection based on serum surface enhanced Raman spectroscopy and statistical analysis. *Optics express*, *23*(14), 18361-18372. https://doi.org/10.1364/OE.23.018361

[19] Maaroof, M. K. A., Al-Msarhed, N. K. H. R., & Jaber, A. S. (2025). Fractal-Hyper Net: Elevating X-ray diagnostic visualization through deep hyper-dimensional feature learning and fractal-scale structural integrity. *Journal of Internet Services and Information Security*, *15*(2), 871–891. https://doi.org/10.58346/JISIS.2025.I2.058

[20] Mahmoudi, S., & Lailypour, C. (2015). A discrete binary version of the Forest Optimization Algorithm. In *International Conference on Information Technology, Computer & Communication*.

[21] Manconi, A., Armano, G., Gnocchi, M., & Milanesi, L. (2022). A soft-voting ensemble classifier for detecting patients affected by COVID-19. *Applied Sciences*, *12*(15), 7554. https://doi.org/10.3390/app12157554

[22] Massaoudi, M., Refaat, S. S., Chihi, I., Trabelsi, M., Oueslati, F. S., & Abu-Rub, H. (2021). A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting. *Energy*, *214*, 118874. https://doi.org/10.1016/j.energy.2020.118874

[23] Oybek Kizi, R. F., Theodore Armand, T. P., & Kim, H. C. (2025). A review of deep learning techniques for leukemia cancer classification based on blood smear images. *Applied Biosciences*, *4*(1), 9. https://doi.org/10.3390/applbiosci4010009

[24] Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August). Study the influence of normalization/transformation process on the accuracy of supervised classification. In *2020 third international conference on smart systems and inventive technology (icssit)* (pp. 729-735). IEEE. doi: 10.1109/ICSSIT48917.2020.9214160.

[25] Rupapara, V., Rustam, F., Aljedaani, W., Shahzad, H. F., Lee, E., & Ashraf, I. (2022). Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. *Scientific reports*, *12*(1), 1000. https://doi.org/10.1038/s41598-022-04835-6

[26] Saccucci, M. S., Amin, R. W., & Lucas, J. M. (1992). Exponentially weighted moving average control schemes with variable sampling intervals. *Communications in Statistics-simulation and Computation*, *21*(3), 627-657. https://doi.org/10.1080/03610919208813040

[27] Selvaraj, S., Alsayed, A. O., Ismail, N. A., Kavin, B. P., Onyema, E. M., Seng, G. H., & Uchechi, A. Q. (2024). Super learner model for classifying leukemia through gene expression monitoring. *Discover Oncology*, *15*(1), 499. https://doi.org/10.1007/s12672-024-01337-x

[28] Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, *8*(1), 101. https://doi.org/10.1186/s40537-021-00492-0

[29] Wang, A., Liu, H., Yang, J., & Chen, G. (2022). Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. *Computers in biology and medicine*, *142*, 105208. https://doi.org/10.1016/j.compbiomed.2021.105208

[30] Zakir Ullah, M., Zheng, Y., Song, J., Aslam, S., Xu, C., Kiazolu, G. D., & Wang, L. (2021). An attention-based convolutional neural network for acute lymphoblastic leukemia classification. *Applied Sciences*, *11*(22), 10662. https://doi.org/10.3390/app112210662

Authors Biography

P.C. Chaitra is currently pursuing a Ph.D. as a Research Scholar in the Department of Computer Science at Dayananda Sagar Academy of Technology & Management, Bengaluru, Karnataka, India, under Visvesvaraya Technological University (VTU), Belagavi. She holds an M.Tech in Computer Science and Engineering (2012) and a B.E. in Computer Science and Engineering (2010), both from VTU, Belagavi. She is presently serving as an Assistant Professor in the Department of Computer Science at Christ (Deemed to be University), Bangalore, India. Prior to this, she worked as an Assistant Professor in the Department of Computer Science at Dayananda Sagar Academy of Technology & Management from 2013 to 2022.Her research interests include Artificial Intelligence and Machine Learning, Application Development using Python, Image Processing, and Compiler Design. She has contributed to several international journals and conferences and actively participates in Faculty Development Programs. She has a keen interest in AI and Data Science education and research.

Dr.R. Saravanakumar R holds degrees in Computer Science and Engineering, including a Ph.D. from Anna University, Chennai in 2015 along with a Master's Degree in Computer Science & Engineering from Anna University, Chennai, and a Bachelor's Degree in CSE from Bharathiyar University, Coimbatore. He has 18+ years of experience in teaching and research, and is currently working as a Professor at Jain (Deemed-to-be University), Bangalore. He has published over 55 refereed publications in renowned Scopus/WoS and SCIE indexed journals and conferences and authored 4 textbooks. He holds 5 Indian patents and has earned an Innovation Grand patent from the UK Patent Office. Additionally, he serves as a Corresponding Editor for IGI Global Publications, CRC Press, and Bentham Science Publications, and actively reviews for international journals. He is also a member of ACM, IAENG, and CSI, focusing his research on Data Analytics, Big Data Analytics, Data Science & Cloud Computing. His expertise is acknowledged by the academic community, actively contributing as a reviewer, technical committee member, Assessment and Accreditation Process (NBA, NIRF, and NAAC)