
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1980–2000, DOI 10.12694/scpe.v25i3.2690

OPTIMAL FEATURE SELECTION FROM HIGH-DIMENSIONAL FUSION OF BLOOD SMEAR IMAGES FOR LEUKEMIA DIAGNOSIS

G. CHINNA PULLAIAH* AND P.M. ASHOK KUMAR†

Abstract. The goal of this study is to improve blood smear image-based blood cancer prediction through medical diagnostic advancements. Blood cancers, particularly leukemia, are challenging to diagnose because of the complexity of biological data and the dimensionality of medical images. There are interpretability and computational problems with each currently in use. We suggest the Random Forest-Recurrent Feature Elimination (RF-RFE) model to increase the precision and dependability of blood cancer diagnosis. This model integrates machine learning and image processing, optimizes feature selection and refinement from high-dimensional data, and applies the XGBoost algorithm to guarantee diagnosis accuracy. Recent model analysis reveals that RF-RFE performs better than them on a wide range of metrics. The RF-RFE offered a sensible, well-rounded strategy. More research on medical diagnostics is made possible by its adaptability in multi-class classification and effectiveness in handling high-dimensional feature values. The optimized feature set and computational efficiency of the model, which may enhance leukemia detection and diagnostics, are highlighted in this study.

Key words: XGBoost, Blood Smear Images, Leukemia Diagnosis, Optimal Feature, Random Forest

1. Introduction. Computational methods improve medical image analysis, including blood disease diagnosis. Example: blood smear image analysis shows blood cell morphology and number. Before, experts manually examined these images, which was tedious, time-consuming, biased, and error-prone [1]. Image processing, machine learning, and digital morphology automate, speed up, and accurately analyze blood smears [2]. Automated analysis can detect subtle blood cell variations that manual methods miss and produce more consistent results [3]. XGBoost and other machine learning methods enable fast, accurate, and automatic blood smear image analysis. XGBoost, scalable and reliable, excels at survival analysis [4] and image classification [5]. It handles large, complex datasets. Using XGBoost to analyze blood smear images requires extracting and selecting blood cell features. Keypoints, color, shape, and texture measures are features. Combining features can increase dimensionality, redundancy, and noise, hurting predictive models. Blood smear image analysis is essential for diagnosing fatal diseases like acute lymphoblastic leukemia (ALL). To simplify and improve feature space, XGBoost-based blood smear image analysis needs efficient feature selection.

This research seeks to create an ALL-diagnostic tool by painstakingly extracting and selecting the most important features from blood smear images. The goals are to analyze blood smear images' many features, create an ideal feature selection procedure to reduce noise and redundant information, and create a diagnostic model using XGBoost's strength with the refined feature set.

Feature selection prepares high-dimensional datasets for accurate predictions. A good choice can reduce computational requirements, improve model precision, and explain data dynamics. ALL diagnoses require top accuracy. Simplified features reduce overfitting by generalizing models.

Advanced feature optimization is possible with RF and RFE. RF's ensemble nature aggregates feature importance while RFE iteratively culls features. XGBoost's gradient boosting framework and unmatched predictive power complete this hybrid approach and handle ALL predictions' complexity.

Blood smear image features are extracted and selected for ALL diagnostics in this research. Although other conditions or domains may benefit from the principles and techniques discussed, ALL prognosis is the main

*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522302 (Corresponding author, pullaiahgcp@gmail.com).

†Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Vaddeswaram, Guntur, Andhra Pradesh, India – 522302 (pmashokk@gmail.com).

focus. Evaluation of all features will show morphological and spatial details' importance. Although predictive modeling is essential to this study, it is mostly used to evaluate the chosen features' reliability and efficiency.

A comprehensive medical image analysis literature review, including XGBoost diagnostics, follows. The paper will then discuss RF-RFE feature selection and XGBoost predictive modeling. Results and discussions will compare findings to prior research, discuss implications, and suggest improvements or more research. The conclusions will summarize key findings and suggest future research

2. Related Research. The diagnosis and early detection of blood cancer using computational techniques remains a critical research avenue, as evidenced by the array of pioneering work in the domain. The following section delves into the recent advancements in this field, encapsulating studies that employ various methods, from matrix-based feature extraction to intricate deep learning frameworks.

Arif Muntasa et al., [6] has presented a method to classify Acute Lymphoblastic Leukemia (ALL) using the Gray Level Co-occurrence Matrix (GLCM) and sixteen distance models, resulting in 192 features for each object. This method achieved an impressive accuracy rate of 96.97% with minimal false positives and negatives, outperforming other existing approaches. Aldinata Rizky Revanda et al. [7] introduced an efficient approach for classifying ALL on white blood cell microscopy images. They propose using Mask R-CNN for instance segmentation and contrast enhancement to improve classification accuracy, achieving an accuracy of 83.72%, precision of 85.17%, and sensitivity of 81.61%.

Zeinab Moshavash et al.'s study [8] focused on accurately diagnosing acute leukemia using blood microscopic images. They introduce a reliable and automatic leukocyte segmentation and feature extraction technique that sets a new benchmark for ALL recognition with 98.10% cell and 89.81% image accuracy.

In order to maximize ALL detection, Nada M. Sallam et al. [9] employed Grey Wolf Optimization (GWO) for feature selection. They increase the efficiency and accuracy of ALL diagnosis to 99.69%, 99.5% sensitivity, and 99% specificity.

Ghada Emam Atteia et al. [10] addressed early ALL prognosis with a hybrid deep learning system. By merging autoencoder networks and pretrained convolutional neural networks, this system achieves feature extraction and ALL diagnosis accuracy better than state-of-the-art techniques.

Using Multiple Instance Learning for Leukocyte Identification (MILLIE), Petru Manescu et al. [11] automate the analysis of blood films and bone marrow aspirates for the diagnosis of acute promyelocytic leukemia through the use of deep learning. MILLIE's high accuracy in identifying APL in bone marrow aspirates and blood films made clinical evaluations easier in environments with limited resources.

Segu Praveena et al. [12] introduced the segmentation and classification of acute lymphoblastic leukemia (ALL) using Deep CNN, Grey Wolf-based Jaya Optimization Algorithm (GreyJOA), and Sparse Fzzy C-Means (Sparse FCM). With its promising sensitivity, specificity, and accuracy, this ALL diagnosis method has the potential to lower patient mortality.

A social spider optimization-based computer-aided diagnosis system for acute lymphoblastic leukemia was proposed by Ahmed T. Sahlol et al. [13]. This innovative model surpasses previous approaches and may help with early ALL diagnosis thanks to its unique integration of multiple features and 95.67% classification accuracy.

Bayesian-optimized CNNs were applied to microscopic blood smear images by Ghada Atteia et al. [14] in order to diagnose ALL. Their model outperformed other cutting-edge techniques with a 96.81% accuracy rate on the test set, demonstrating its enormous potential for ALL detection.

A computer system based on image analysis was created by Ahmed M. Abdeldaim et al. [15] to diagnose ALL. With the K-NN classifier in particular, the system segments and classifies cells as normal or affected with good accuracy. Although statistical results are not provided in the article, the suggested system might aid in the diagnosis of ALL.

The inefficiencies of manually identifying acute lymphoblastic leukemia (ALL) were investigated by Adel Sulaiman et al. [16]. ResRandSVM increases the accuracy of automated diagnosis by utilizing Random Forest for feature selection, ResNet50 for feature extraction, and Support Vector Machine for classifier. Three methods are used to refine the deep features that multiple models extract. The improved features for blood smear leukemia detection are tested by four classifiers. ResRandSVM performs well when using InceptionV3 for feature extraction, Random Forest for feature refinement, and SVM for classification. ResRandSVM performs

better in experiments than in other comparisons, indicating that it has the potential to expedite ALL diagnosis.

The recent advancements in Leukemia diagnosis present a range of methodologies, each contributing valuable insights to the domain. Arif Muntasa et al.'s work [6] utilizes the Gray Level Co-occurrence Matrix (GLCM) for feature extraction, whereas other studies, like that of Ghada Emam Atteia et al. [10], delve into deep learning frameworks.

One notable trend from the related research is the focus on robust feature extraction and optimization. Arif Muntasa et al.'s approach [6], while achieving high accuracy rates, extracts 192 features for each object. Such a comprehensive feature space, while detailed, might introduce redundancy, potentially leading to computational inefficiencies and overfitting. This is where our RF-RFE model's strength becomes evident. By streamlining feature sets and removing non-essential attributes, RF-RFE offers an optimized and relevant feature set for diagnosis.

Deep learning models, such as Ghada Atteia et al.'s Bayesian-based CNNs [14] and Petru Manescu et al.'s MILLIE [11], are powerful but often demand significant computational resources. Moreover, their complexity can sometimes challenge interpretability, which is essential in medical applications.

Optimization techniques also find representation in this array of research. Nada M. Sallam et al.'s work [9] introduces the Grey Wolf Optimization algorithm for feature selection. Their method, with its impressive accuracy metrics, illustrates the potential of nature-inspired algorithms. In contrast, our RF-RFE, rooted in Random Forests, offers a method that seeks to understand the inherent structure of the data.

Furthermore, Adel Sulaiman et al.'s ResRandSVM [16] shares similarities with our approach by integrating feature extraction, refinement, and classification. However, our RF-RFE stands apart in its explicit focus on addressing redundancy in high-dimensional data, ensuring the most optimal feature set powers the subsequent XGBoost mechanism.

Considering the diverse methodologies in blood cancer diagnosis, our RF-RFE model emerges as a balanced approach that emphasizes precision, computational efficiency, and clarity, marking its significance in the field.

3. Methods and Materials. The sequential phase architecture of the RF-RFE framework is depicted in Figure 3.1. Microscopic blood smear features are used by the novel machine learning-based blood cancer prediction model RF-RFE. Preprocessing techniques are prominently featured in this first stage to enhance image clarity and quality. Next, combine the texture, color, and morphology to create a comprehensive feature vector. RF-RFE, designed for high-dimensional data, controls refinement. Through a series of iterations, this process carefully eliminates less significant attributes to produce an ideal feature set for analysis. The gradient-boosting algorithm XGBoost performs well with complex biological data. XGBoost [17] starts off with this precisely calibrated feature set. This machine learning model trains, adapts, and improves its predictive capabilities using gradient-boosted tree algorithms [16]. Throughout the modeling process, performance is monitored to guarantee the best possible outcomes. This system detects leukemia-variant blood cancers early by using machine learning and image analysis.

3.1. Preprocessing.

Image Resizing. Image resizing standardizes the dimensions of all images to a consistent size. This ensures that features extracted from each image are comparable and consistent. Images acquired from different sources or devices can have varying dimensions. Resizing them to a consistent dimension helps in managing the computational cost and ensuring uniformity in feature extraction.

Let I be the input image with dimensions (h, w). Resizing it to dimensions (h', w') is achieved by a spatial transformation function T, such that $I' = T(I)$ where I' is the resized image.

Noise Reduction. Noise reduction involves filtering the image to remove unwanted artifacts and noise, which could distort the image's actual content. Medical images might contain noise due to various reasons like electronic interference, transmission errors, or imperfect sensors. Removing this noise is essential for clear visualization and accurate feature extraction. A common method is Gaussian blurring, represented as: $I' = G * I$ where I' is the denoised image, $*$ denotes convolution, and G is a Gaussian kernel.

Contrast Enhancement. Contrast enhancement amplifies the differences between pixel values in an image, making features more distinguishable. Some medical images might have low contrast due to the nature of the tissue or the acquisition process. Enhancing contrast aids in better visualization and differentiation of regions of interest.

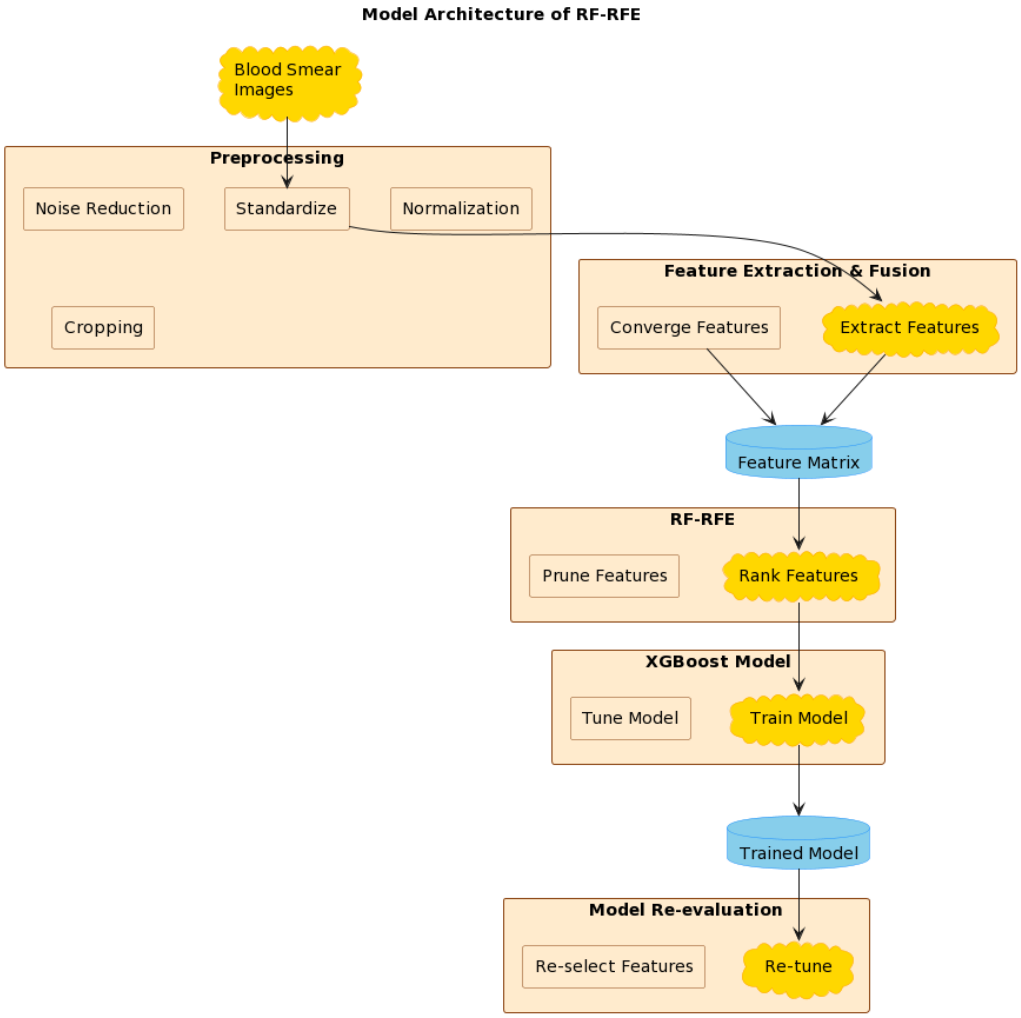

Fig. 3.1: Architecture of RF-RFE

Histogram equalization is one technique:

$$p_r(_k) = n_k/MN \tag{3.1}$$

where p_γ is the normalized histogram, γ_k are pixel intensities, n_k is the number of pixels with intensity γ_k, and $M \times N$ is the image size.

Image Segmentation. Image segmentation partitions an image into multiple segments or regions, often separating objects of interest from the background. In medical imaging, segmenting out regions of interest, like tumors or specific organs, allows for targeted analysis and reduces computational costs.

One method is the Otsu's thresholding:

$$\sigma_\omega^2(t) = \omega_0(t)\omega_1(t)[\mu_0(t) - \mu_1(t)]^2 \tag{3.2}$$

3.2. Feature Engineering. Feature engineering is a cornerstone in machine learning applications for predicting ALL from blood smear images. Fusion of diverse features – including keypoints and key descriptors, morphological attributes, color distributions, texture patterns, spatial relationships, boundary contours, and the

Nucleus to Cytoplasm Ratio – provides a comprehensive representation of the intricate details present in blood smear images. Each of these feature types captures a unique aspect of cellular structures and their potential abnormalities, ensuring the model receives a holistic understanding of the image content. The significance of this fusion lies in its ability to enhance the model's robustness and predictive capability; while some features, like color, might capture staining intensity variations indicative of ALL, others, such as morphological features, can hint at cell structure anomalies. By integrating these diverse features, we harness the collective strength of each feature type, thereby justifying their fusion for an optimal and accurate ALL prediction model.

3.2.1. The Features. Keypoints and Descriptors: These features highlight the salient features of key-points found in images. Scale [18], orientation [19], location [20], contrast [21], edge response, Harris response, main orientation, descriptor vector [22], Laplacian sign [23], and magnitude [24] are some of them. For locating specific areas of interest in images and understanding their characteristics, these features are essential. While the descriptor vector encodes local appearance, scale, orientation, and location are particularly crucial for spatial information.

Physical characteristics [25]. Morphological features identify characteristics of the object's size and shape. Area, perimeter, compactness, major and minor axis lengths, eccentricity, convexity, area, solidity, extent, orientation, and equivalent diameter are important characteristics. The morphology and structure of objects, such as cells in medical images, can be described using these features. Fundamental size and shape descriptors include area and perimeter, while eccentricity and convexity shed light on any irregularities in an object's shape.

Features of color. Color features concentrate on the color data present in objects. For the red, green, and blue channels, they include mean intensities and standard deviations as well as chroma, hue, value (brightness), color variance, and color entropy. For distinguishing objects based on their color attributes, these features are crucial. The distribution and variation of color can be better understood using mean intensities and standard deviations.

Details of the texture. The spatial arrangement of pixel intensities within objects is described by texture features. They consist of Haralick textures [26], Gabor filters [27], energy (uniformity), entropy [28], homogeneity, correlation, dissimilarity, second moment, and fractal dimension [29]. Understanding the minute details and patterns within objects requires these features, which are essential. For instance, contrast and entropy quantify the complexity and randomness of a texture.

Features of spatial relationships. The arrangement and relationships between the objects in an image are described by spatial relationship features. They include the following metrics: the nearest neighbor distance, pairwise distance statistics (mean and standard deviation), the clustering coefficient, the convex hull area ratio, the object separation index, the object density, the object orientation, the object eccentricity, the object area ratio, and the object perimeter ratio. Analysis of object spatial distributions and clustering patterns benefits greatly from these features.

Features of the boundary and contour. The shape and boundary characteristics of objects are the focus of boundary and contour features. They consist of the following: perimeter, compactness, aspect ratio, circularity, solidity, convexity, bending energy, curvature, and skeletonization. Insights into the object's general shape, roundness, and curvature are provided by these features, which are crucial for identifying object classes.

Ratio of Cytoplasm to Nucleus. The interaction between the nucleus and cytoplasm in cells is quantified by these features. They include the nucleus to cytoplasm area ratio, the nucleus to cytoplasm perimeter, the nucleus to cytoplasm roundness ratio, the nucleus to cytoplasm eccentricity ratio, the nucleus to cytoplasm eccentricity ratio, the nucleus to cytoplasm eccentricity ratio, and the nucleus to cytoplasm eccentricity. The balance between the properties of the nucleus and the cytoplasm, which is a feature of these features, can be a sign of the health of a cell in the context of medical image analysis.

3.2.2. Feature Extraction. Feature extraction from blood smear images is a process tailored to capture intricate cellular details pivotal for diagnostics. It begins with key-points and key-descriptors, identifying unique patterns within the cells that are invariant to image transformations. Morphological features elucidate the shape and structure of cells, highlighting any irregularities. While texture features depict subtle patterns and variations within the cell structures, the color spectrum captures the variation in staining intensities, which can be indicative of pathological changes. Context is provided by spatial relationship features, which show how cells are positioned and distributed in relation to one another. The Nucleus to Cytoplasm Ratio

provides information about cellular composition, which is frequently disturbed in conditions like ALL, while boundary and contour features highlight the edges and outline of cells. With their combined ability to provide a multidimensional view of blood cells, these features are crucial for sophisticated diagnostic machine learning models.

Key Ideas and Key Descriptives A well-known method for extracting key details and their descriptors from images that guarantees invariance to scale, rotation, and lighting changes is called SIFT (Scale-Invariant Feature Transform) [30]. These key points can indicate particular unique patterns or anomalies in cells in the context of blood smear images. SIFT is a viable option for thorough blood cell analysis because of its capacity to recognize and describe these unique features, despite possible variations in image capture conditions.

Scale-space Extrema Detection. The first step in SIFT is generating a scale space. This is achieved by convolving the original image $I(x, y)$ with Gaussian functions $G(x, y, \sigma)$ over a range of scales. This can be represented as: $L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$. Here, '*' denotes the convolution operation.

Key point Localization. After creating the scale space, we search for potential keypoints. These are identified at the maxima and minima of the difference-of-Gaussians (DoG) [31] function. The DoG is formed as: $D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$. Here '$k$' is a multiplicative constant.

Orientation Assignment. For rotation invariance, each key point is given an orientation based on the local gradient directions of the image. The gradient magnitude $m(x, y)$ and direction $\theta(x, y)$ at each pixel are given by:

$$m(x, y) = \sqrt{((L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2)} \tag{3.3}$$

$$\Theta(x, y) = arctan((L(x, y + 1) - L(x, y - 1))/(L(x + 1, y) - L(x - 1, y)))) \tag{3.4}$$

Keypoint Descriptor. Finally, a descriptor for each keypoint is formed by accumulating gradient magnitudes and orientations in a localized region around the keypoint. This step ensures the descriptor's robustness to changes in appearance, such as lighting or affine transformations.

Morphological Features [32]. Morphological features are crucial in differentiating various cell structures within blood smear images. The method of Watershed segmentation, aided by gradient information, is paramount in delineating these attributes. It effectively discriminates between cells that are adjacent or slightly overlapping, making it apt for precisely defining boundaries. Given the essential nature of cell morphology in diagnostics, this technique's accuracy and versatility make it indispensable.

Gradient Computation. Given an image $I(x, y)$, the gradient magnitude is:

$$G(x, y) = ((I_x(x, y))^2 + (I_y(x, y))^2) \tag{3.5}$$

Distance Transform. For a binary image $B(x, y)$, the distance $D(x, y)$ to the nearest zero pixels is:

$$D(x, y) = min_{(i,j)}((x - i)^2 + (y - j)^2) \tag{3.6}$$

Watershed Segmentation. Using $G(x, y)$, basins are formed meeting at watershed lines, indicating cell boundaries.

Color Features. Color histograms and moments are foundational for extracting color features from blood smear images, capturing variations in staining intensities. These intensities can hint at abnormalities, making the method invaluable for diagnostics. The approach is justified as it's computationally efficient and offers a broad representation of cellular color distribution. Histogram Computation: Given image I, histogram H for a color channel c is:

$$H_c(k) = |(x, y)|I_c(x, y) = k| \tag{3.7}$$

Moments. The n^{th} moment of a histogram H_c is:

$$M_n = \overset{255}{(k=0)} k^n H_c(k) \tag{3.8}$$

Texture Features. Gray Level Co-occurrence Matrix (GLCM) [33] stands out for texture feature extraction. By analyzing pixel pair frequencies at specific positions, GLCM encapsulates patterns and textures in cells, pivotal for discerning abnormalities. Its efficacy in capturing local variations makes it a justified choice.

GLCM Computation. For an offset $(\Delta x, \Delta y)$, $GLCM_P(i,j)$ is the frequency of pixel pairs with intensities i and j.

Spatial Relationship Features. To understand the spatial positioning and orientation of cells, Delaunay triangulation is optimal. This method creates triangles connecting nearby cells, offering insights into cell distribution and proximity. Given the importance of cell relationships in diagnostic contexts, this approach is vital.

Delaunay Triangulation. Given a set of points P, a triangle (p,q,r) belongs to the Delaunay triangulation if no other point in P lies within the circumcircle of the triangle.

Circumcircle condition. For each triangle ΔABC in T, let O be the center of the circumcircle passing through A, B, and C. The triangulation T is Delaunay if and only if no point P from the set lies inside the circle with O as the center.

Empty Circle Property. For every triangle ΔABC in the Delaunay Triangulation, the circumcircle of ΔABC does not contain any other point of P in its interior.

Boundary and Contour Features. Active Contour Model or Snakes is a potent method for boundary and contour feature extraction. By iteratively evolving curves based on internal and external forces, it clings to cell boundaries, ensuring precise contour delineation. This method's adaptability to subtle boundary nuances justifies its adoption. Snake Evolution: The snake $\upsilon(s) = [x(s), y(s)]$ evolves according to:

$$F_{total}() = {}^{1}_{0}F_{int}() + F_{image}() + F_{con}()ds \tag{3.9}$$

Spatial Relationship Features. To understand the spatial positioning and orientation of cells, Delaunay triangulation is optimal. This method creates triangles connecting nearby cells, offering insights into cell distribution and proximity. Given the importance of cell relationships in diagnostic contexts, this approach is vital.

Boundary and Contour Features. Active Contour Model or Snakes is a potent method for boundary and contour feature extraction. By iteratively evolving curves based on internal and external forces, it clings to cell boundaries, ensuring precise contour delineation. This method's adaptability to subtle boundary nuances justifies its adoption.

Nucleus to Cytoplasm Ratio [34]. Thresholding and region-based segmentation are crucial for delineating the nucleus and cytoplasm in cells. By computing their areas separately and determining their ratio, insights into cellular health are gleaned. This feature is critical given its prominence in many pathological conditions, including ALL.

Thresholding: For image I, binary image $B(x,y)$ is:

$$(1if I(x,y) > T; 0otherwise) \tag{3.10}$$

Ratio Computation: For segmented nucleus area A_N and cytoplasm area A_C:

$$Ratio = AN/AC \tag{3.11}$$

3.3. Optimal Feature Selection. Modern medical image analysis relies heavily on extracting comprehensive features from images to improve the accuracy of disease predictions. When dealing with blood smear images, especially in the context of ALL prediction, a fusion of various features — including keypoints and key descriptors, morphological features, color features, texture features, spatial relationship features, boundary and contour features, and the Nucleus to Cytoplasm Ratio — provides a rich representation of data. Yet, such fusion can also introduce redundancy and noise. Therefore, there's a pressing need for an optimal selection process to retain only the most significant features, ensuring efficient and accurate diagnosis models.

The combination of Random Forest (RF) [35] with Recursive Feature Elimination (RFE) [36] offers a systematic approach to tackle this challenge. RF inherently ranks features based on their importance, providing an aggregated measure of their significance in the classification task. This prioritization becomes critical when

handling a diverse set of features, ensuring the model focuses on the most relevant attributes. RFE, on the other hand, is a recursive method that eliminates less important features step by step, thereby refining the feature set.

1. Holistic Data Representation [37]. Given the fusion of diverse features, there's a mix of linear and non-linear data patterns. RF's nature to cater to both ensures that no crucial data pattern is overlooked.

2.Redundancy Reduction [38]. The fusion of multiple feature sets often leads to overlapping information. RF's intrinsic feature ranking, combined with the iterative removal process of RFE, ensures that redundant features are systematically eliminated.

3. Interpretability. Medical diagnostics requires not just accuracy but also the ability to understand the decision-making process. RF offers insight into feature importance, aiding researchers and medical practitioners in discerning the key features driving predictions.

4. Optimal Performance. RF's ensemble nature, utilizing multiple decision trees, ensures a balance between bias and variance, leading to robust and stable predictions. When combined with the refined feature set from RFE, it results in enhanced model performance.

5. Efficiency in Training. By focusing only on the most significant features, the computational burden during model training is reduced, leading to faster and more efficient model training without compromising accuracy.

Optimal feature selection using RF-RFE involves a synergistic approach to refine a fusion of diverse features - keypoints and key descriptors, morphological features, color characteristics, texture patterns, spatial relationship attributes, boundary and contour details, and the Nucleus to Cytoplasm Ratio. This amalgamation offers a comprehensive representation of data, capturing both global and local nuances. RF-RFE stands out in this context due to its inherent ability to rank features based on their ensemble importance. By systematically and recursively eliminating less impactful features, RF-RFE ensures the retention of only the most significant ones, enhancing model performance. This methodology leverages the strengths of Random Forest, such as handling non-linear patterns and feature interactions, to provide a robust and justified selection of optimal features from the intricate fusion.

Filter techniques prioritize features using individual statistical metrics, often overlooking their interactions or their relevance to the target variable. They might consider measures like variance or outcome correlation. While efficient, they can miss intricate relationships in a diverse feature set. RF-RFE, leveraging Random Forest, excels in recognizing inter-feature relationships, giving a richer assessment. With each tree evaluating features across different scenarios, RF-RFE adeptly navigates complex and non-linear relationships, proving more suitable for critical ALL prediction features.

Wrapper techniques, encompassing methods like backward elimination, depend on specific classifiers to assess feature subsets. They take into account feature interplays and can produce classifier-specific feature sets. However, they're resource-heavy, especially for vast feature sets derived from image fusion. RF-RFE smartly amalgamates wrapper and embedded strengths. Random Forest's ranking captures intricate patterns, and RFE's recursive procedure facilitates streamlined, efficient feature pruning. This blend enhances scalability and adaptability, especially beneficial for the nuanced feature set in ALL prediction.

Techniques like LASSO merge feature selection with model training. While efficient and often clear-cut, they may not always grasp complex relationships, especially with a broad fusion of features from blood smear images. RF-RFE, utilizing Random Forest's ensemble strength, delivers a robust feature significance assessment. Paired with RFE's methodical approach, it ensures a thorough yet focused feature exploration, making RF-RFE particularly suitable for the multifaceted feature landscape of ALL prediction.

Theoretical foundation. Random Forest-Recursive Feature Elimination (RF-RFE) combines the robust classification capabilities of Random Forest (RF) [39] with the systematic feature pruning of Recursive Feature Elimination (RFE) [40]. RF builds multiple decision trees on varied data subsets and averages their predictions, offering reduced overfitting and high interpretability. Each tree's construction uses a random subset of features, emphasizing different attributes across trees. RFE, on the other hand, iteratively trains the model, ranks features by their importance, and removes the least significant ones. When fused, RF-RFE leverages RF's feature importance metrics to efficiently and recursively prune irrelevant features, optimizing the model for both performance and interpretability, especially vital in intricate tasks like medical diagnostics.

Random Forest (RF). The strength of RF comes from aggregating (or "bagging") the results of numerous decision trees, each trained on a subset of the data. The variability among trees decreases the model's variance, reducing the likelihood of overfitting.

Let's denotes:
D: The original dataset.
D_i: A bootstrap ample of D
F: The bull set of features.
F_j: A random subset of features at node split j.

Lemma 1. Every tree in the forest is built on a bootstrap sample (a random sample with replacement) from the original data. This bootstrap sampling introduces variability among the trees:

$$D_i = (x_1^*, y_1^*), (x_2^* y_2^*), ..., (x_n^* y_n^*) \tag{3.12}$$

where each $(x_k^* y_k^*)$ is a random sample with replacement from D.

Lemma 2. At each node split, only a random subset of features is considered, further introducing variability among the trees. This randomness ensures that the trees are uncorrelated, making the averaging process more effective at reducing variance. For each node split j: $F_j \subset F$ where F_j is a randomly selected subset of features from F at that node.

The forest's final prediction, Y_{RF} for regression can be an average of the individual trees' predictions, and for classification, it can be a majority vote. If T represents the total number of trees:

$$Y_{RF} = \frac{1}{T} \sum_{i=1}^{T} Y_{\text{tree}_i}$$

where Y_{tree_i} is the prediction of the i^{th} tree.

Recursive Feature Elimination (RFE). RFE is a wrapper-based feature selection algorithm that fits the model multiple times, each time eliminating the least important features.

Let's denote:
Φ: A function which ranks features based on importance after training the model.
F_k: Set of features retained in the k^{th} iteration.

Lemma 3:. At each iteration, after the model (in this case, RF) is trained, features are ranked based on their importance. The least important features are more likely to add noise than provide value. After training on F_k features:

$$(F_k) = {}_1, {}_2, ..., {}_k \tag{3.13}$$

where${}_1$ is the most important and ${}_k$ is the least important.

Lemma 4. By recursively training the model and eliminating the least important features at each step, the model becomes more focused on the most significant features. This stepwise refinement ensures that the final feature subset is optimal or near-optimal for model performance.

Given a step size after each iteration:

$$F_{(k+1)} = F_k - {}_k, {}_{(k-1)}, ..., {}_{(k-+1)} \tag{3.14}$$

That is, the least important δ features from F_k are removed to form F_{k+1}. This recursive process continues until a desired number of features is retained, or until model performance meets a specified criterion.

3.4. RF-RFE Algorithm.

Initialization: Start with the full dataset D and the complete feature set F.
Set T as the number of trees for the RF model.

Set δ as the number of features to remove in each iteration of RFE.
Set $F_{\text{current}} = F$.

Random Forest Training: For $i = 1$ to T
(a) Bootstrap Sampling:
$D_i = \{(x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_n^*, y_n^*)\}$
Where each (x_k^*, y_k^*) is a random sample with replacement from D.
(b) Construct Tree: For each node split j:
Select a random subset of features:
$F_j \subset F_{\text{current}}$
Split the node using the best feature in F_j based on an impurity criterion (e.g., Gini impurity or entropy).

Feature Importance Evaluation: After training the RF model on F_{current}:
$\Phi(F_{\text{current}}) = \{\phi_1, \phi_2, ..., \phi_m\}$
where ϕ_1 is the most important feature, and ϕ_m is the least important, and m is the size of F_{current}.

Feature Elimination: Remove the least important δ features:
$F_{\text{next}} = F_{\text{current}} - \{\phi_m, \phi_{m-1}, ..., \phi_{m-\delta+1}\}$
Set $F_{\text{current}} = F_{\text{next}}$

Recursive Iteration: Repeat the above 4 steps from initialization to feature elimination until the desired number of features is retained, or another stopping criterion such as model performance on a validation set reaches a threshold is met.

Final Model Training: Train the RF model on the dataset D using the final selected feature subset from F_{current}.
Model Evaluation and Prediction: Evaluate the model's performance using out-of-bag samples. Out-of-bag (OOB) [41] error estimation is a unique property of the bootstrap aggregating (bagging) procedure, which is central to the Random Forest algorithm. When a specific data instance is not used for building a particular tree during bootstrap sampling, it becomes an OOB sample for that tree. Given the nature of bootstrapping, roughly one-third of the data are left out of the bootstrap sample and not used in the construction of the k^{th}tree.

- Let D be the dataset of size N.
- Let T be the number of trees in the random forest.
- For each instance x_i in D, let Trees(x_i) be the set of trees for which x_i is an OOB sample.
- Let $\text{Pred}_{\text{tree}_j}(x_i)$ be the prediction of the j^{th} tree for the instance x_i:

OOB Prediction for a Data Instance: For each instance x_i, the OOB prediction, $\text{Pred}_{\text{OOB}}(x_i)$, is given by the majority vote (classification) or average (regression) of the predictions of the trees for which x_i is an OOB sample. $\text{Pred}_{\text{OOB}}(x_i) = \text{MajorityVote}(\{\text{Pred}_{\text{tree}_j}(x_i) | \text{tree}_j \in \text{Trees}(x_i)\})$ (For classification)
$\text{Pred}_{\text{OOB}}(x_i) = \frac{1}{|\text{Trees}(x_i)|} \sum_{\text{tree}_j \in \text{Trees}(x_i)} \text{Pred}_{\text{tree}_j}(x_i)$ (For regression)

OOB Error: The OOB error is the proportion of instances that are misclassified (for classification) or the mean squared error (for regression) based on the OOB predictions:

$$Err_OOB = 1/N_{(i=1)}^{N} I[Pred_{OOB}(x_i) y_i] \tag{3.15}$$

(For classification, where is the indicator function, which is 1 if the condition is true and 0 otherwise) $Err_{OOB} = 1/N_{(i=1)}^{N} (Pred_{OOB}(x_i) - y_i)^2$ (For regression)

The Err_{OOB} provides an unbiased estimate of the rest error without the need for cross-validation or a separate rest set, making it highly efficient for model evaluation in bagging-based methods like random forest.

3.5. Model Building with XGBoost. Utilizing XGBoost with features selected by RF-RFE addresses the complexity and high dimensionality inherent in biological data, such as blood smear images for ALL prediction. By combining Random Forest's capacity to discern feature importance with XGBoost's gradient-boosting mechanism, this approach offers an enhanced predictive accuracy and efficiency. The synergy between the ensemble techniques of RF-RFE and XGBoost together ensures robust feature selection, reduced overfitting, and a model fine-tuned for performance, making it particularly vital for the precise and critical domain of medical diagnoses like ALL.

Let D be the dataset of blood smear images.
Let F be the fusion of features extracted from D, where:
$F = \{$keypoints and key descriptors,
morphological features, color characteristics,
texture patterns,
spatial relationship attributes,
boundary and contour details,
Nucleus to Cytoplasm Ratio$\}$

Step 1: Feature Extraction and Fusion

- For each image i in D:
- Extract each feature set f in F
- Create a combined feature vector v_i for image i

Step 2: RF-RFE for Optimal Feature Selection

- Train a Random Forest classifier on D with all features in F.
- Use the feature importance scores provided by the RF classifier to rank the features.
- Initialize a subset S with all features from F.
- While S has more than one feature:
- Remove the least important feature (based on RF's ranking) from S.
- Retrain the RF classifier with the reduced feature set S.
- Update the feature ranking based on the retrained RF classifier.
- The final feature subset S^* is the one that achieves the highest classification performance on a validation set.

Step 3: Model Building with XGBoost

- Let $L(y, \hat{y})$ be the logistic loss function where y is the true label and $\hat{y}$ is the predicted probability.
- Initialize model with: $\hat{y}_i^{(0)} = \frac{1}{2} \ln \left(\frac{\sum_{y_i=1} \omega_i}{\sum_{y_i=0} \omega_i} \right)$ where ω_i is the instance weight.
  For each boosting round $t = 1$ to T:
  Compute the gradient and hessian for each instance i:
  $g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$
  $h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$
- Build a regression tree to predict the gradients using feature from S^*
- For each leaf j of the tree, compute: $\omega_j = -\frac{\sum_{i \in \text{leaf } j} g_i}{\sum_{i \in \text{leaf } j} (h_i + \lambda)}$ where λ is a regularization parameter.
- Update the predictions for each instance i: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \omega_j$ where η is the learning rate and i belongs to leaf j.

3.6. Model Tuning and Re-evaluation. Stagnation isn't acceptable. The model, post its initial training, enters a phase of continuous evaluation. Through regular hyper-parameter tuning and occasional feature re-selection, it ensures its predictions remain sharp and relevant.

Let D_{train} be the training dataset of blood smear images, and D_{al} be the validation dataset.
Let Prepresent hyperparameters for XGBoost, including:

- Learning rate
- Maximum tree depth$D$

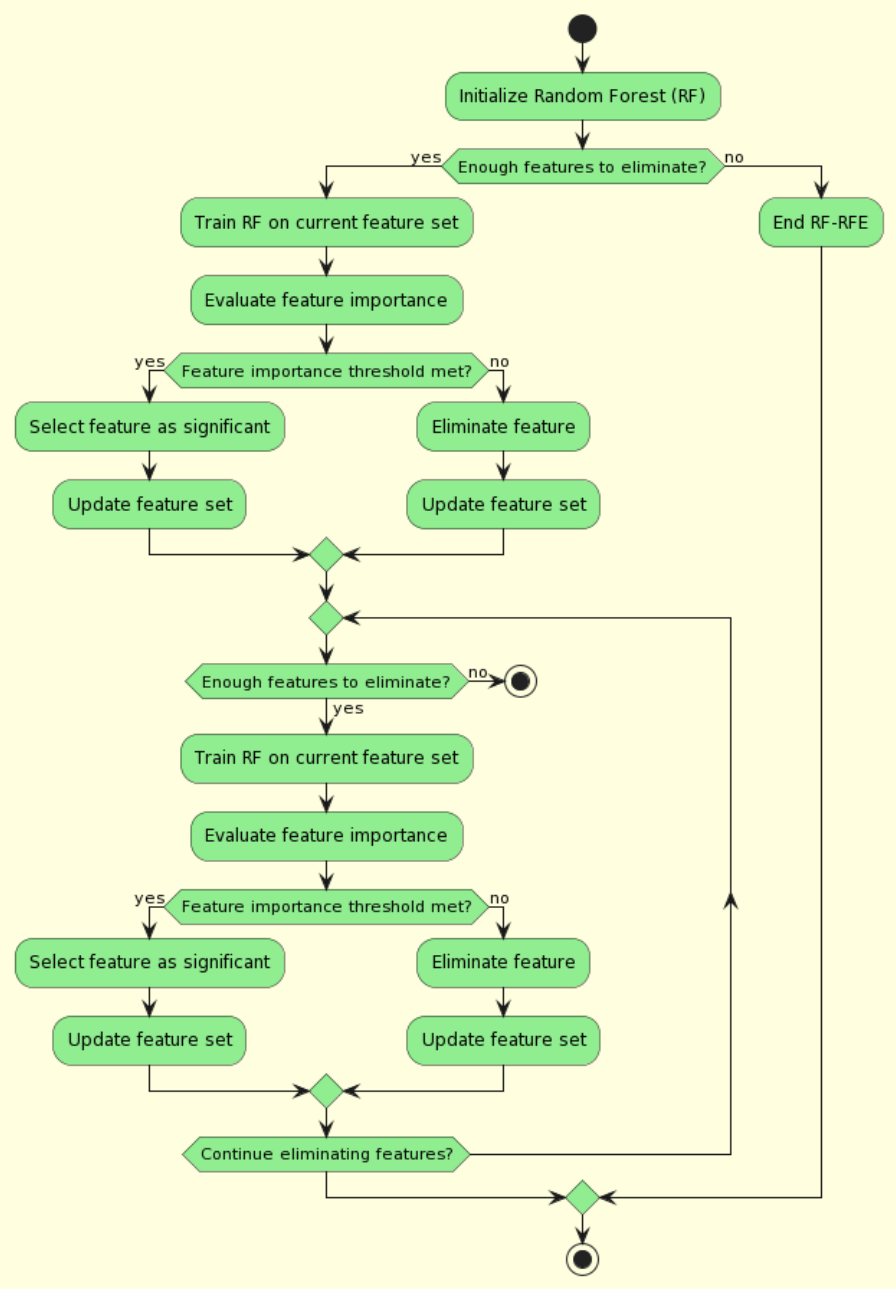

Fig. 3.2: Flow Diagram of the RF-RFE

- Minimum child weight$_{min}$
- Subsample ratio
- Column (feature) sample rate
- Regularization term

Initial Training and Evaluation. Train the XGBoost model on D_{train} using features selected by RF-RFE and initial hyperparameters P_0. Evaluate the model on D_{al} to obtain performance metric M_0.

Hyper-parameter Tuning. For each hyperparameter pin$P$:

Table 4.1: List of Assumptions related to 10 fold cross validation perform

Assumption	Description
Data Source and Quality (C_NMC_2019 dataset)	The C_NMC_2019 dataset [43] is assumed to be a reliable and representative source of blood smear images for Acute Lymphoblastic Leukemia (ALL) diagnosis. It is assumed that the dataset has been carefully curated and contains images indicative of various disease stages.
10-Fold Cross-Validation Methodology	The experimental study assumes the use of a 10-fold cross-validation methodology, which involves dividing the dataset into 10 subsets (folds) for training and testing. This methodology ensures robust model evaluation by exposing it to different training-test splits.
Model Comparison (RF-RFE, GWO, RESRANDSVM)	The study assumes the comparison of the RF-RFE model's performance with that of two contemporary models, GWO [9] and RESRANDSVM [16]. It is assumed that these models are suitable benchmarks for evaluating the effectiveness of RF-RFE in leukemia detection.
Performance Metrics	Multiple performance metrics, including precision, recall, specificity, accuracy, f-measure, and ROC (Receiver Operating Characteristic), are assumed to be used for model evaluation. These metrics provide a comprehensive understanding of the models' capabilities.
Goal of Comprehensive Evaluation	The ultimate goal of the experimental study is to assess the true predictive power of the models and their ability to balance false positives and false negatives. This assessment aims to determine the potential real-world applicability of the models in leukemia detection.

1. Adjust p within a predefined range or set.
2. Retrain XGBoost on D_{train} using the updated hyperparameters.
3. Evaluate the model on $D_{al}$to Obtain performance matricM_p .
4. If M_{pis} better (e.g. higher accuracy or AUC, lower loss) than the best metric so far, update P_{best} with the current set of hyperparameters.

Model Re-evaluation with Updated Hyper-parameters. Train the XGBoost model on D_{train}using features selected by RF-RFE and P_{best}.

Evaluate the model on $D_{al}$to confirm performance improvement.

The flow diagram shown in figure 3.2 illustrates the process of Random Forest - Recursive Feature Elimination (RF-RFE), a feature selection technique used in machine learning. The diagram starts with the initialization of the Random Forest. It then enters a loop where it evaluates the importance of features in the current feature set. If a feature meets the importance threshold, it is selected as significant, and the feature set is updated. If not, the feature is eliminated from the set, and the feature set is also updated. This loop continues until there are no more features to eliminate or until a predefined stopping condition is met. Once the loop ends, the RF-RFE process concludes, and the diagram depicts the end of the process. RF-RFE is a systematic approach to select the most relevant features for model training, reducing complexity and improving model performance.

4. Experimental Study. In order to assess the efficacy of the RF-RFE model in the diagnosis of Acute Lymphoblastic Leukemia (ALL) [42] using the C_NMC_2019 dataset [43], an experimental study was carefully planned. The study applied a rigorous 10-fold cross-validation methodology while utilizing the dataset's richness, which offers a wide variety of blood smear images indicative of different stages of the disease. This improved the reliability of the performance evaluation by ensuring that the model was exposed to a variety of training-test splits. The assumptions have been listed in table 4.1 The performance of the RF-RFE model was then compared to that of the contemporary models GWO [9] and RESRANDSVM [16]. Precision, recall, specificity, accuracy, f-measure, and ROC were just a few of the metrics used to provide a thorough understanding of the models' capabilities. The goal of this comprehensive evaluation strategy was to reveal the models' true predictive power as well as their capacity to balance false positives and false negatives, thereby capturing their potential real-world applicability in the crucial field of leukemia detection.

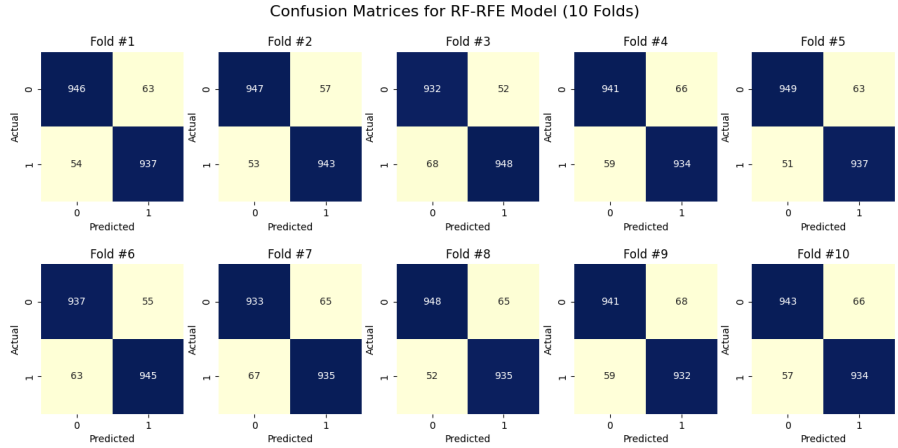

Fig. 4.1: Confusion matrices of 10-fold cross validation performed on proposed model RF-RFE

4.1. The Data. The C_NMC_2019 (Children's Leukemia Data Challenge 2019) dataset is a valuable resource for the development of machine learning models aimed at pediatric leukemia diagnosis, encompassing a total of 12,528 cell images. Among these images, 8,491 represent cases of Acute Lymphoblastic Leukemia (ALL), a critical cancer subtype, while 4,037 images depict normal cell samples. This dataset's significant size and the balanced distribution of cancer and normal cell images make it an ideal choice for robust and comprehensive model training and evaluation in the domain of pediatric leukemia diagnosis.

For a precise diagnosis, the integrity of medical images, particularly blood smear images, is crucial. However, in actual situations, a number of factors may add noise to these images. The C_NMC_2019 dataset intentionally include noise to simulate these imperfect conditions, testing the robustness of the diagnostic algorithms. Intentionally reducing the specificity and sensitivity of the features extracted from the blood smear images allows for a more thorough assessment of the algorithms being tested. A total of 20,000 cell segments have been meticulously selected from the source microscopic images of blood smears, comprising an equal distribution of 10,000 cells from leukemia-infected blood smear images and an additional 10,000 cells from the source images of normal blood smears. This balanced and comprehensive dataset ensures a diverse representation of both pathological and healthy cell samples, offering a robust foundation for subsequent analyses and research endeavors.

4.2. Performance analysis. The RF-RFE model, evaluated through 10-fold cross-validation on a dataset with balanced positives and negatives, consistently demonstrated exceptional performance in leukemia detection. It achieved high true positives and true negatives across all folds, indicating its proficiency in accurately classifying both leukemia-infected and normal cells that shown in figure 4.1. With precision values ranging from 0.9326 to 0.9472 and sensitivity values between 0.932 and 0.949, the model showcased its ability to minimize false positives while effectively identifying positive instances. Furthermore, its specificity remained consistently high, varying from 0.932 to 0.948, ensuring reliable negative classifications. The model's overall accuracy ranged from 0.934 to 0.945, highlighting its capacity for accurate predictions. The F-measure, between 0.9323 and 0.9476, struck a balance between precision and recall, while the false alarming rate remained impressively low at 0.055 to 0.066. With Matthews correlation coefficients ranging from 0.868 to 0.890 and a false positive rate varying from 0.052 to 0.068, the RF-RFE model consistently exhibited robust and reliable leukemia detection capabilities across different folds, making it a promising tool for accurate disease diagnosis. According to the confusion matrices visualized in figure 4.2, the Grey Wolf Optimization (GWO) model demonstrated robust performance across the ten-fold cross-validation, showcasing its effectiveness in distinguishing between leukemia-infected and normal blood smear images. With an average accuracy of 92.35%, GWO exhibited a strong ability to correctly classify instances, supported by high precision (93.95%) and sensitivity

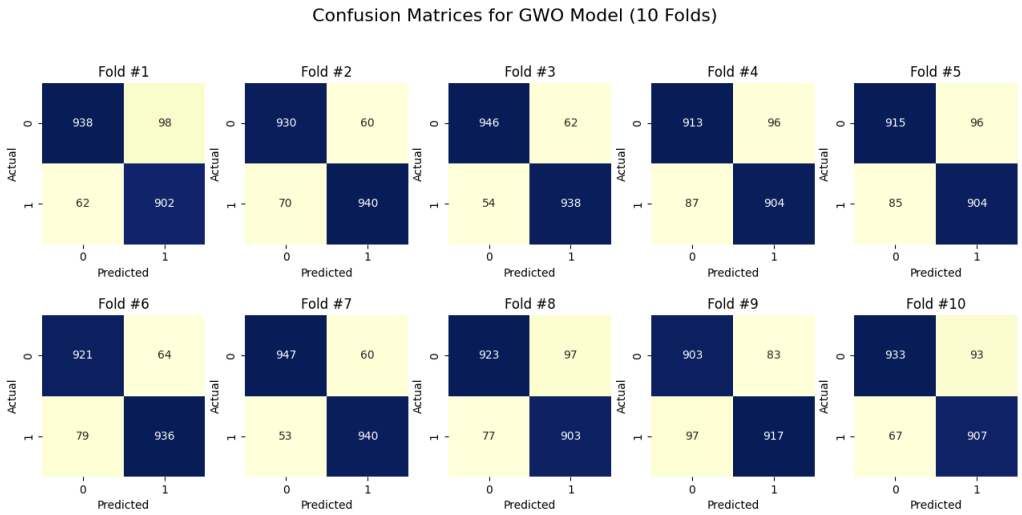

Fig. 4.2: Confusion matrices of 10 fold cross validation performed on contemporary model GWO

(93.05%). The model maintained a well-balanced trade-off between specificity (93.75%) and false positive rates, indicating its competence in avoiding misclassifications. Additionally, the F1-score of 93.82% highlights the model's capability to achieve a harmonious balance between precision and recall. The Matthews Correlation Coefficient (MCC) of 0.8692 further affirmed its performance. Overall, the GWO model exhibited promising potential in the task of leukemia prediction, demonstrating consistent and reliable results across different folds of the dataset. The RESRANDSVM model demonstrates consistent and performance across the 10-fold cross-validation experiments that visualized as confusion matrices of all 10-folds of the cross validation in figure 4.3. It exhibits good precision, specificity, and sensitivity, with values consistently above 0.88, indicating a strong ability to correctly classify both positive and negative cases. The model maintains a high accuracy ranging between 0.88 and 0.91, demonstrating its effectiveness in overall classification. Furthermore, the F-measure, a harmonic mean of precision and sensitivity, consistently exceeds 0.88, indicating a balanced trade-off between precision and recall. The false alarm rate is acceptably low, with values around 0.10, indicating a relatively low rate of misclassification. The Matthews correlation coefficient (MCC) values range between 0.76 and 0.82, signifying a moderate to substantial degree of correlation between predicted and actual classifications. Overall, the RESRANDSVM model showcases a commendable performance in binary classification tasks across various folds, highlighting its reliability and suitability for the given dataset and problem domain.

4.3. Comparative Study. Precision is an imperative metric that gauges the capability of a classification model to identify only the relevant data points accurately. High precision suggests that false positives (incorrectly identified positives) are minimal. According to the results visualized in figure 4.4, RF-RFE emerges as a consistent performer with its precision scores maintaining a tight range around the 0.93 to 0.94 mark across all ten folds. This suggests that its predictions are both accurate and reliable. In contrast, GWO showcases a slightly broader range of fluctuation. Although its precision peaks around 0.940 in a couple of folds, some dips to 0.904 highlight pockets of inconsistency. The RESRANDSVM method exhibits the most variability, with precision scores hovering between 0.87 and 0.90. This indicates a higher propensity to misclassify positive instances compared to the other two methods.

Specificity is a crucial metric, particularly when the cost of false positives is high. It gauges the accuracy of a model in identifying negative outcomes. As visualized in figure 4.4, once again, RF-RFE stands out with its specificity values mirroring its precision scores, ranging mostly between 0.93 and 0.94. Its consistent performance across both metrics emphasizes its balanced and effective classification capabilities. GWO, on the other hand, portrays a pattern akin to its precision values. While it achieves commendable specificity scores upwards of

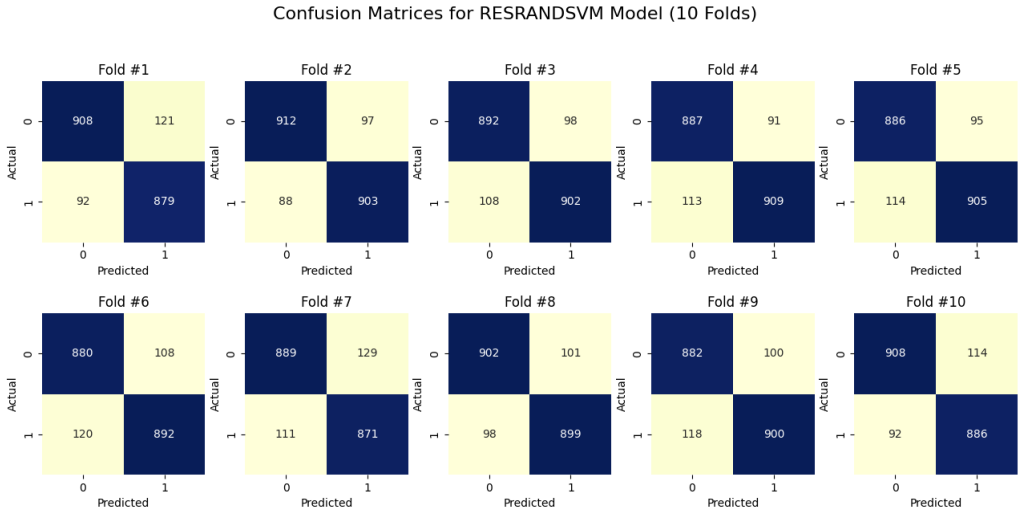

Fig. 4.3: Confusion matrices of 10-fold cross validation performed on contemporary model RESRANDSVM

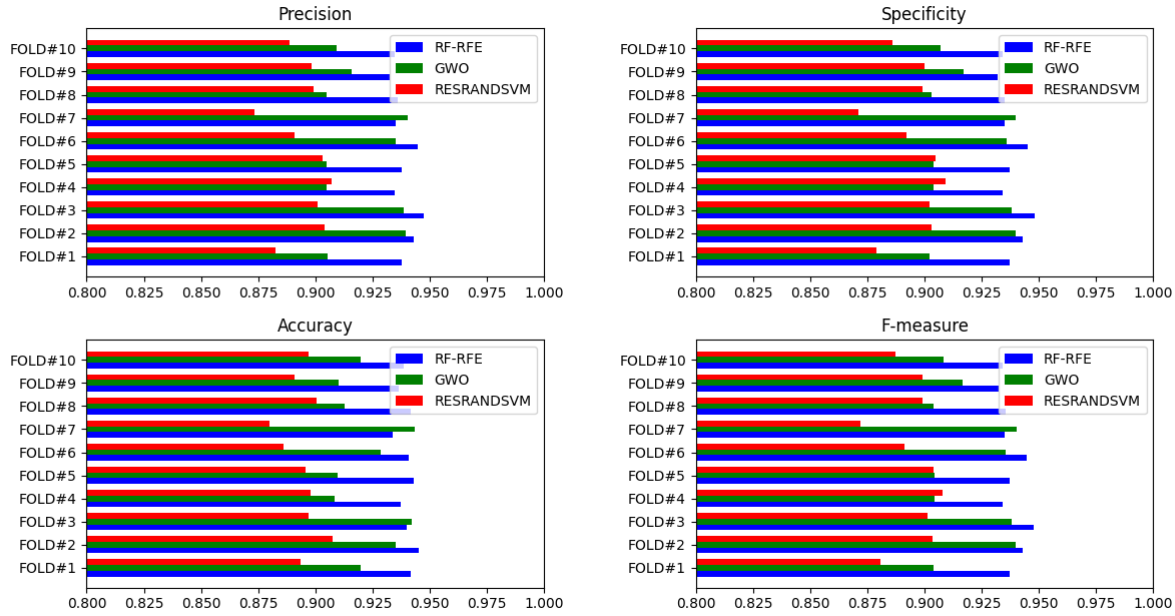

Fig. 4.4: graphs representing the performance metrics precision, specificity, accuracy and f-measure of RF-RFE, GWO, and RESRANDSVM obtained from 10-fold cross validation

0.940 in certain folds, occasional dips towards 0.902 suggest occasional inconsistencies. RESRANDSVM remains the least consistent across the board, with most of its specificity scores settled between 0.87 and 0.90.

Accuracy, perhaps one of the most intuitive performance metrics, offers a comprehensive overview of a model's classification prowess by accounting for both true positives and true negatives. Based in the figure 4.4, RF-RFE consistently achieves the pinnacle of accuracy among the methods, oscillating mainly between

Table 4.2: The presents a Cross Validation of three Methods: RF-RFE, GWO, and RESRANDSVM, across four metrics: Precision, Specificity, Accuracy, and F-measure

Metric	**RF-RFE**	**GWO**	**RESRANDSVM**
Precision	0.9398 ± 0.0048	0.9199 ± 0.0145	0.8948 ± 0.0099
Specificity	0.9390 ± 0.0048	0.9186 ± 0.0146	0.8946 ± 0.0101
Accuracy	0.9401 ± 0.0036	0.9230 ± 0.0133	0.8947 ± 0.0074
F-measure	0.9397 ± 0.0048	0.9195 ± 0.0143	0.8946 ± 0.0096

0.934 and 0.945 across the folds. This affirms its ability to make correct predictions reliably. GWO presents a mixed bag, with accuracy values that diverge notably from one fold to another, spanning from 0.9085 to 0.9435. This variability suggests that its performance might be context-dependent. RESRANDSVM continues its trend of trailing the pack, managing accuracy primarily in the 0.88 to 0.90 range, indicating potential areas for improvement.

The F-measure is a composite metric that strikes a balance between precision and recall, offering a more holistic view of a model's performance, especially when classes are imbalanced. The consistent brilliance of RF-RFE is evident once more from the figure 4.4, as its F-measure scores are closely aligned with its precision, averaging around 0.93 to 0.94. This indicates a harmonious balance between its precision and recall capabilities. GWO displays scores ranging from 0.9037 to 0.9402, reinforcing the narrative of its slightly fluctuating performance. Finally, RESRANDSVM hovers in the lower spectrum with F-measure values mostly between 0.87 and 0.90, reinforcing the notion that it might not be as effective as the other two in the tested scenarios.

Table 4.2 presents a comparative analysis of three methods: RF-RFE, GWO, and RESRANDSVM, across four metrics: Precision, Specificity, Accuracy, and F-measure. Each entry is represented by its average value followed by a deviation. Among the methods, RF-RFE consistently showcases the highest values across all metrics, with Precision at 0.9398 ± 0.0048, Specificity at 0.9390 ± 0.0048, Accuracy at 0.9401 ± 0.0036, and F-measure at 0.9397 ± 0.0048. GWO follows closely, while RESRANDSVM tends to have the lowest values in each category. The deviations also highlight the consistency in the results, with RF-RFE having the smallest variations, indicating its robust performance.

Sensitivity measures the proportion of actual positives that are correctly identified, which is showcased in figure 4.5. RF-RFE shows commendable sensitivity, predominantly fluctuating in the range of 0.932 to 0.949 across the folds. This consistent performance indicates that RF-RFE is adept at identifying true positive cases. On the other hand, GWO exhibits a broader spread ranging from 0.903 to 0.947. While in some folds it manages to rival RF-RFE, in others, it tends to drop notably. RESRANDSVM lingers mostly in the 0.880 to 0.912 bracket, making it the method with the lowest sensitivity on average. It suggests that of the three methods, RESRANDSVM might miss a higher proportion of positive instances.

The false positive rate quantifies the proportion of negatives that are mistakenly classified as positive. Lower FPR values are desirable. As shown in figure 4.5, RF-RFE showcases impressive control over FPR, with values mainly clustered between 0.052 and 0.068. GWO presents a wider range, oscillating between 0.06 and 0.098, signifying a slightly elevated risk of incorrectly classifying negatives. RESRANDSVM consistently registers the highest FPR among the three, with values spanning from 0.091 to 0.129, highlighting its potential vulnerability in misclassifying negative instances.

Similar in essence to FPR, the false alarm rate gauges the frequency of false alarms that presented in figure 4.5. RF-RFE continues its trend of robust performance with values chiefly contained within the 0.055 to 0.066 bracket. This demonstrates its reliability in curbing false alarms. GWO, while respectable in its performance, exhibits a tad more variability, spanning 0.0565 to 0.0915. RESRANDSVM again lags, recording rates from 0.0925 to 0.120, signifying its increased likelihood to raise false alarms compared to the other two techniques.

MCC is a balanced metric that considers all values in the confusion matrix, with 1 indicating perfect prediction, -1 indicating total disagreement, and 0 denoting no better than random prediction. RF-RFE consistently leads in this metric as shown in figure 4.5, with scores ranging from 0.868 to 0.890, underscoring its

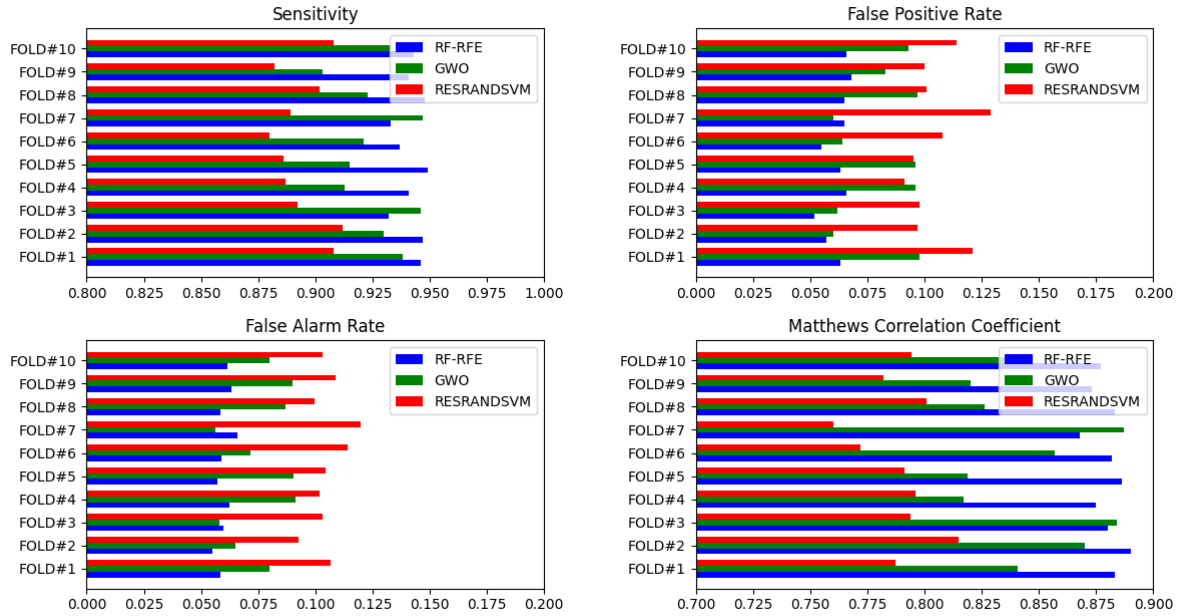

Fig. 4.5: Graphs Representing the Performance Metrics Sensitivity, FPR, FAR, MCC proposed RF-RFE of the compared GWO and RESRANDSVM obtained from 10-fold Cross Validation

Table 4.3: Presents a Cross Validation of three methods: RF-RFE, GWO, and RESRANDSVM, across four metrics: Sensitivity, FPR, FAR, and MCC

Metric	**RF-RFE**	**GWO**	**RESRANDSVM**
Sensitivity	0.9413 ± 0.0061	0.9216 ± 0.0144	0.8961 ± 0.0099
False Positive Rate	0.0615 ± 0.0052	0.0809 ± 0.0151	0.1054 ± 0.0108
False Alarm Rate	0.0602 ± 0.0028	0.0761 ± 0.0133	0.1055 ± 0.0067
Matthews Correlation Coefficient	0.8817 ± 0.0070	0.8461 ± 0.0249	0.7903 ± 0.0158

all-rounded efficacy. GWO follows suit with values mostly between 0.817 and 0.887, suggesting a commendable yet slightly more varied performance. RESRANDSVM, while not too far behind, predominantly hovers in the 0.7601 to 0.815 range. This indicates that, on average, its predictions might be somewhat less correlated with the actual outcomes compared to the other two methods. In the comparative analysis based on the metrics provided in the table 4.3, the RF-RFE method consistently showcased superior performance across all metrics when compared to GWO and RESRANDSVM. Specifically, for Sensitivity, RF-RFE averaged 0.9413, which was higher than GWO's average of 0.9216 and RESRANDSVM's average of 0.8961. Similarly, RF-RFE also exhibited the lowest False Positive Rate and False Alarm Rate among the three methods, indicating a lower likelihood of erroneous classifications. In terms of the Matthews Correlation Coefficient, which measures the quality of binary classifications, RF-RFE again outperformed with an average score of 0.8817. Overall, while all three methods yielded commendable results, RF-RFE stood out as the most effective in this analysis.

4.3.1. Precision-Recall (PR)-Curve. Precision-Recall (PR) curves that presented in figure 4.6, provide an insightful way of examining the performance of classification algorithms, particularly in scenarios where classes are imbalanced. They plot the trade-off between the positive predictive value (precision) and the true positive rate (recall/sensitivity), providing a holistic view of an algorithm's ability to distinguish between classes. In our comparative evaluation of the PR-curves for RF-RFE, GWO, and RESRANDSVM methods, distinct

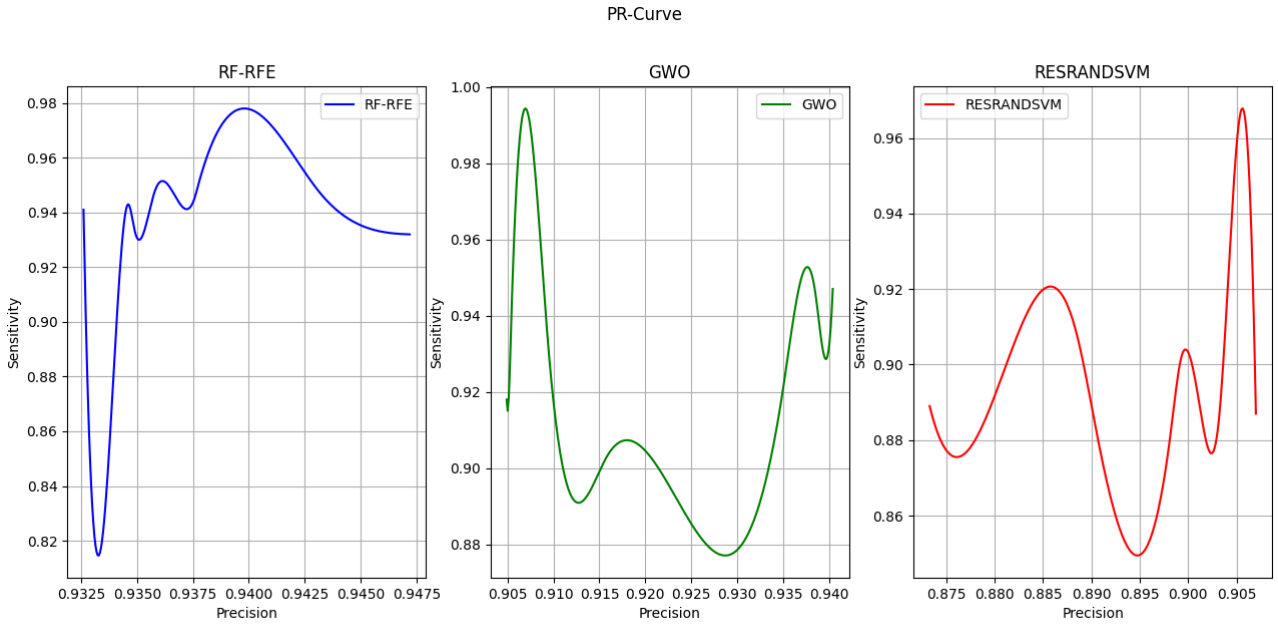

Fig. 4.6: PR-Curve of RF-RFE, GWO, and RESRANDSVM Methods Derived from 10-Fold Cross Validation.

trends are observed. RF-RFE stands out with a consistently superior performance, demonstrated by its steeper ascent in the curve, which implies a robust capability to maintain high precision across diverse sensitivity levels. GWO, though exhibiting some fluctuations suggesting potential variances in precision at different recall intervals, still holds a notable position in the analysis. RESRANDSVM, while displaying a more equilibrated precision-recall trade-off, might not reach the precision peaks of RF-RFE. Analytically, RF-RFE emerges as the top performer in this comparison, suggesting that it's likely to produce fewer false positives for a given recall threshold. However, each method brings its strengths and weaknesses, reinforcing the importance of using PR-curves in understanding the nuances of classifier performance.

4.3.2. Receiver Operating Characteristic (ROC)-Curve. The ROC-Curve that shown in figure 4.7 is a graphical representation of the true positive rate (Sensitivity) against the false positive rate for various threshold values. An ideal method would yield a point in the upper-left corner of the ROC space, representing 100% sensitivity and 0% false positive rate.

From the provided data, RF-RFE demonstrates higher sensitivity across almost all folds compared to the other two methods, especially when false positive rates are low. This means that RF-RFE is potentially better at discriminating between the positive and negative classes. GWO, on the other hand, shows competitive sensitivity values but often at the cost of higher false positive rates. The RESRANDSVM method appears to have the lowest sensitivity values among the three methods in most of the folds, suggesting it might have a lower discriminative ability in this specific context.

5. Conclusion. A sophisticated machine learning and image processing method for blood cancer prediction from blood smear images, the RF-RFE model, was introduced in this study. Blood cancer diagnosis, especially Acute Lymphoblastic Leukemia, has improved with RF-RFE's clarity and precision. Our model greatly improves leukemia detection efficiency and accuracy. Its ability to reduce redundancy in high-dimensional medical imagery data makes RF-RFE unique. The feature set of modern models like GWO [9] and RESRANDSVM [16] is optimised by RF-RFE in detail. Specificity, Accuracy, Precision, and F-measure are a set of performance metrics. It shows the model's dependability. Our extensive analysis showed RF-RFE's precision and low binary classification errors. Further comparisons of Sensitivity, False Positive Rate, False Alarm Rate, and Matthews

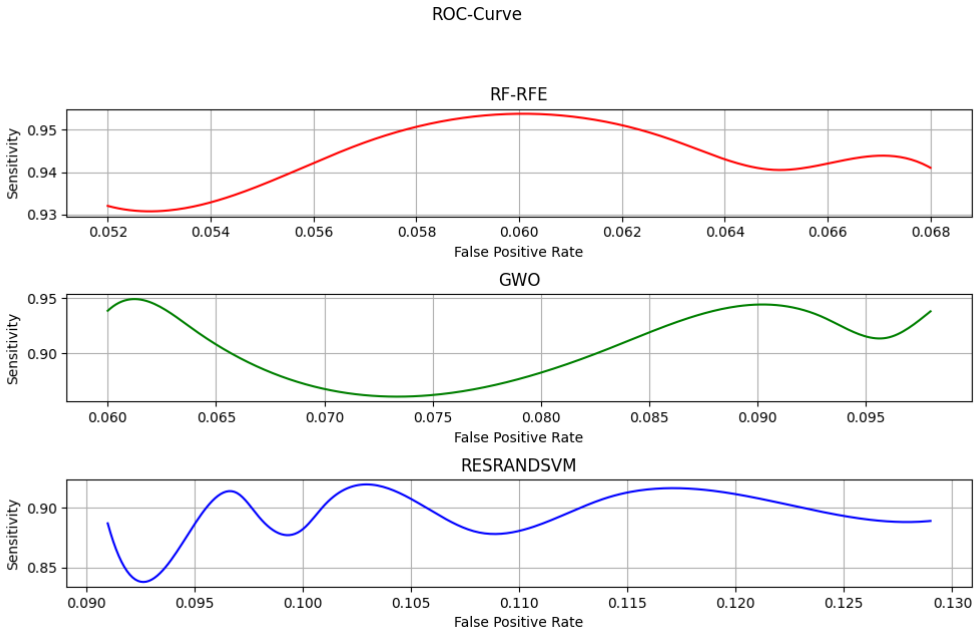

Fig. 4.7: ROC-curve of RF-RFE, GWO, and RESRANDSVM Methods Derived from 10-Fold Cross Validation .

Correlation Coefficient show the model's precision and ability to balance false positives and negatives. Grey Wolf Optimization, Bayesian-based CNNs [14], and GLCM [6] are impressive, but RF-RFE stands out. Strategically integrating the XGBoost algorithm, RF-RFE sets the standard for ALL diagnosis and biological data analysis. Due to its unique approach and optimized feature set, it is a blood cancer diagnostic breakthrough with consistent performance across diverse datasets. Early blood cancer detection, especially for ALL, is transformed by the model's false positive and negative ability. Many medical diagnostics applications and research are promising with RF-RFE. Adapting RF-RFE for multi-class classification will increase its applicability and help us understand subtype-specific treatment approaches for blood cancer. High-dimensional feature values can be handled without affecting predictive quality with advanced dimensionality reduction. These improvements may enhance RF-RFE's medical diagnostic performance, relevance, and applicability.

REFERENCES

[1] Y. M. Alomari, S. N. H. Sheikh Abdullah, R. Z. Azma, and K. Omar, *Automatic detection and quantification of WBCs and RBCs using iterative structured circle detection algorithm*, Computational and mathematical methods in medicine, 2014.

[2] *Blood Smear Test*, Available at: https://www.testing.com/tests/blood-smear/.

[3] S. Fathima, P. Meenatchi, and A. Purushothaman, *Comparison of manual versus automated data collection method for haematological parameters*, Biomedical Journal of Scientific & Technical Research, 15, no. 3 (2019), pp. 11372-11376.

[4] W. Jiao, X. Hao, and C. Qin, *The image classification method with CNN-XGBoost model based on adaptive particle swarm optimization*, Information, 12, no. 4 (2021), p. 156.

[5] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, *A novel image classification method with CNN-XGBoost model*, In: Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, Proceedings 16, Springer International Publishing, 2017, pp. 378-390.

[6] A. Muntasa, and M. Yusuf, *Multi Distance and Angle Models of the Gray Level Co-occurrence Matrix (GLCM) to Extract the Acute Lymphoblastic Leukemia (ALL) Images*, International Journal of Intelligent Engineering & Systems, 14, no. 6 (2021).

[7] *Classification of Acute Lymphoblastic Leukemia on White Blood Cell Microscopy Images Based on Instance Segmentation Using Mask R-CNN.*

[8] Z. Moshavash, H. Danyali, and M. S. Helfroush, *An automatic and robust decision support system for accurate acute leukemia diagnosis from blood microscopic images*, Journal of digital imaging, 31 (2018), pp. 702-717.
[9] N. M. Sallam, A. I. Saleh, H. A. Ali, and M. M. Abdelsalam, *An efficient strategy for blood diseases detection based on grey wolf optimization as feature selection and machine learning techniques*, Applied Sciences, 12, no. 21 (2022), p. 10760.
[10] G. E. Atteia, *Latent Space Representational Learning of Deep Features for Acute Lymphoblastic Leukemia Diagnosis*, Computer Systems Science & Engineering, 45, no. 1 (2023).
[11] P. Manescu, et al., *Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning*, Scientific Reports, 13, no. 1 (2023), p. 2562.
[12] S. Praveena, and S. P. Singh, *Sparse-FCM and Deep Convolutional Neural Network for the segmentation and classification of acute lymphoblastic leukaemia*, Biomedical Engineering/Biomedizinische Technik, 65, no. 6 (2020), pp. 759-773.
[13] A. T. Sahlol, A. M. Abdeldaim, and A. E. Hassanien, *Automatic acute lymphoblastic leukemia classification model using social spider optimization algorithm*, Soft Computing, 23 (2019), pp. 6345-6360.
[14] G. Atteia, et al., *Bo-allcnn: Bayesian-based optimized cnn for acute lymphoblastic leukemia detection in microscopic blood smear images*, Sensors, 22, no. 15 (2022), p. 5520.
[15] A. M. Abdeldaim, A. T. Sahlol, M. Elhoseny, and A. E. Hassanien, *Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis*, Advances in Soft Computing and Machine Learning in Image Processing, 2018, pp. 131-147.
[16] A. Sulaiman, et al., *ResRandSVM: Hybrid Approach for Acute Lymphocytic Leukemia Classification in Blood Smear Images*, Diagnostics, 13, no. 12 (2023), p. 2121.
[17] T. Chen, et al., *Xgboost: extreme gradient boosting*, R package version 0.4-2, 1, no. 4 (2015), pp. 1-4.
[18] A. Herod, *Scale*, Routledge, 2010.
[19] B. K. P. Horn, *Relative orientation*, International Journal of Computer Vision, 4, no. 1 (1990), pp. 59-78.
[20] J. J. Gabszewicz, and J.-F. Thisse, *Location*, Handbook of game theory with economic applications, 1 (1992), pp. 281-304.
[21] C. Owsley, *Contrast sensitivity*, Ophthalmology Clinics of North America, 16, no. 2 (2003), pp. 171-177.
[22] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi, *Unsupervised learning of landmarks by descriptor vector exchange*, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6361-6371.
[23] Y. Hou, J. Li, and Y. Pan, *On the Laplacian eigenvalues of signed graphs*, Linear and Multilinear Algebra, 51, no. 1 (2003), pp. 21-30.
[24] R. E. Kirk, *The importance of effect magnitude*, Handbook of research methods in experimental psychology, 2003, pp. 83-105.
[25] P. H. Pelham, and J. G. Dickson, *Physical characteristics*, The wild turkey: biology and management, Stackpole Books, Mechanicsburg, Pennsylvania, USA, 1992, pp. 32-45.
[26] E. Miyamoto, and T. Merryman, *Fast calculation of Haralick texture features*, Human computer interaction institute, Carnegie Mellon University, Pittsburgh, USA, Japanese restaurant office, 2005.
[27] J. R. Movellan, *Tutorial on Gabor filters*, Open source document, 40 (2002), pp. 1-23.
[28] B. Bein, *Entropy*, Best Practice & Research Clinical Anaesthesiology, 20, no. 1 (2006), pp. 101-109.
[29] J. Theiler, *Estimating fractal dimension*, JOSA A, 7, no. 6 (1990), pp. 1055-1073.
[30] T. Lindeberg, *Scale invariant feature transform*, 2012, p. 10491.
[31] A. Bundy and L. Wallen, *Difference of gaussians*, Catalogue of Artificial Intelligence Tools, (1984), p. 30.
[32] U. Ziegler and P. Groscurth, *Morphological features of cell death*, Physiology, 19, no. 3 (2004), pp. 124-128.
[33] S. V. Bino, A. Unnikrishnan, and K. Balakrishnan, *Gray level co-occurrence matrices: generalisation and some new features*, arXiv preprint arXiv:1205.4831, (2012).
[34] J. A. Sebastian, M. J. Moore, E. S. L. Berndl, and M. C. Kolios, *An image-based flow cytometric approach to the assessment of the nucleus-to-cytoplasm ratio*, PLoS One, 16, no. 6 (2021), e0253439.
[35] S. J. Rigatti, *Random forest*, Journal of Insurance Medicine, 47, no. 1 (2017), pp. 31-39.
[36] X.-w. Chen and J. C. Jeong, *Enhanced recursive feature elimination*, In Sixth international conference on machine learning and applications (ICMLA 2007), IEEE, 2007, pp. 429-435.
[37] J. Kim, H. Lee, M. Imani, and Y. Kim, *Efficient Hyperdimensional Learning with Trainable, Quantizable, and Holistic Data Representation*, In 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), IEEE, 2023, pp. 1-6.
[38] H. Barlow, *Redundancy reduction revisited*, Network: computation in neural systems, 12, no. 3 (2001), p. 241.
[39] L. Breiman, *Out-of-bag estimation*, (1996).
[40] M. Onciu, *Acute lymphoblastic leukemia*, Hematology/oncology clinics of North America, 23, no. 4 (2009), pp. 655-674.
[41] *The Cancer Imaging Archive*, Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52758223.

Edited by: S. B. Goyal
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Oct 13, 2023
Accepted: Jan 23, 2024