IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2247~2259
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2247-2259

Stacking ensemble techniques for automated peripheral blood cell classification using Inception v3 features

Marwa Mawfaq Mohamedsheet Al-Hatab[1], Maysaloon Abed Qasim[2], Nawar A. Sultan[2]
[1]Technical Engineering College, Northern Technical University, Mosul, Iraq
[2]Technical Engineering College for Computer and Artificial Intelligence, Northern Technical University, Mosul, Iraq

Article Info

Article history:

Received Sep 22, 2025
Revised Feb 21, 2026
Accepted Apr 22, 2026

Keywords:

Cross validation
Ensemble learning
Inception v3
Machine learning
Principal component analysis

ABSTRACT

Robust distinction of blood cells is crucial in clinical evaluation. Manual examination is slow and exposed to errors. This work investigates using machine learning (ML) techniques for automated classification of eight categories of peripheral blood cell types from multi-color images. The Inception v3 network was used to extract features, a split of 66%/34% were used to evaluate the model along with 20-fold cross-validation. To reduce computational complexity, principal component analysis (PCA) was used to reduce the 2048-dimensional feature vectors to 100 components. Among all classifiers used, the highest performance without using PCA was achieved using the support vector machine (SVM) with an accuracy equal to 93.4% and an area under the curve (AUC) of 0.996. Using PCA, affected monocytes and immature granulocytes most due to the slight reduction in the accuracy and AUC which became 90.1% and to 0.991 respectively. Results were further enhanced when a stacked ensemble of neural network (NN), logistic regression (LR), and SVM were used, achieving an accuracy of 95.2% and an AUC of 0.998. The obtained findings confirmed the effectiveness of using stacked ensembles in providing a robust, high-accuracy framework for automated blood cell classification, while PCA efficiently reduced dimensions with minimal performance loss.

Corresponding Author:

Marwa Mawfaq Mohamedsheet Al-Hatab
Technical Engineering College, Northern Technical University
Mosul, Iraq
Email: marwa.alhatab@ntu.edu.iq

1. INTRODUCTION

Blood cell analysis is an essential issue in clinical diagnostics because it provides significant information about immune status, hematological disorders, and disease progression [1]. However, although that the traditional manual test which depend on microscope widespread, it suffers from inter-observer variability, subjectivity, and low throughput, especially in resource-constrained environments. These disadvantages led to the development of automated and semi-automated image-based systems that use computer vision and machine learning (ML) to improve reproducibility, speed, and diagnostic accuracy [2]. Early computational methods used the combination of handcrafted features such as shape, texture, and color and classical machine-learning algorithms such as support vector machines (SVMs) or random forests (RFs). These techniques were very successful in simple classification tasks, but they failed in complex morphological variations, overlapping cells, and staining inconsistencies [3], [4].

In contrast, modern advances in deep learning, helped the classification systems and have become the main paradigm for identifying red blood cells (RBCs), white blood cells (WBCs), and platelets [5]. Moreover, using the combination of ensemble learning strategies, for example those using multiple

convolutional neural network (CNN) architectures in classification systems improved robustness, accuracy, and generalization performance compared to single-model approaches [6]. Dimensionality reduction and selection approaches such as principal component analysis (PCA), particle swarm optimization (PSO), entropy minimization, and memetic algorithms are other techniques applied to deep feature embeddings in order to decrease redundancy and keep discriminative information [7]. In addition to these methods, domain adaptation and data augmentation techniques ranging from stain normalization to semi-supervised learning played a big role in performance improvement, especially for rare cell types and datasets from different sources [8]. Multi-class blood-cell classification systems which use high-capacity backbones such as ResNet and EfficientNet together with ensemble methods and advanced feature analysis achieved classification accuracies exceeding 99% [9]. Despite all of the above blood cell classification systems have man challenges nowadays , such as handling image noise, ensuring real-time deployment, enhancing interpretability in clinical contexts, and achieving model generalization across laboratories and imaging protocols [10]. In this work a novel framework combining advanced feature dimensionality analysis with ensemble learning to obtain high accuracy, compact feature representations, and practical value in real-world clinical settings has been proposed.

Recently deep learning, ensemble methods, and feature optimization techniques have been used widely in modern blood cell classification to achieve robustness, and high accuracy. Parayil *et al*. [11] proposed a hierarchical artificial intelligence (AI) pipeline for multi-class RBC morphology analysis, their method illustrates the potential of staged classification models. Banerjee and Chaudhuri introduced a framework based on applying two-stage feature selection based on total contribution score and fuzzy entropy, their work achieved high accuracy and reduced feature dimensionality.

Using ensemble learning in classification system further enhanced classification performance. Ghosh *et al.* [13] achieved 96.67% accuracy for leukocyte classification b integrating DenseNet121, ResNet50, and a custom CNN via a least entropy combiner. Ichim *et al*. [14] proposed a method o enhance he robustness across eight blood cell classes using decision fusion across multiple networks (VGG16, Xception, ResNet50, and NasNetLarge). Tenguam *et al*. [15] introduced a model which integrate deep and handcrafted features with ensemble learning for H&E histopathology images. ReRNet and Routt *et al.* [16] used CNN ensembles and segmentation-based tracking for RBC morphology analysis, they obtained 99.97% accuracy.

Feature optimization using meta-heuristics, entropy control, PCA, and PSO are other techniques which proved effectiveness. Ahmad *et al.* [17] proposed a WBC classification using significant feature reduction, achieving 99.9% accuracy. Awais *et al*. [18] introduced an acute lymphoblastic leukemia (ALL) subtype using memetic deep feature optimization, heir method obtained approximately 99% accuracy. Cai *et al.* [19] proposed a framework combining low-rank adaption (LoRA)-adapted, segment anything models (SAM) with unsupervised autoencoders for cross-domain single-cell image classification.

Sharma [20] developed a CNN model for multi-class blood cell identification across benign and malignant categories. His model recorded more than 98% accuracy. Tepakhan *et al.* [21] proposed a system to differentiate between iron deficiency anemia and thalassemia, they applied ensemble stacking with RF and gradient boosting algorithms. Their system demonstrated high classification performance and the effectiveness of ensemble-based methods for complex clinical datasets. Uzma and Afsar [22] proposed a semi-supervised framework to improve WBC subtype classification with minimal labeled data.

Despite these advances, balancing computational efficiency, feature dimensionality, and class-level discriminative power still challenges. This study addressed these difficulties by combining InceptionV3 feature extraction with a stacked ensemble of neural network (NN), logistic regression (LR), and SVM. The effectiveness of applying PCA on classification performance was systematically evaluated. Compared to the early studies, this approach obtains higher accuracy and robustness, highlighting the advantages of integrating deep-feature and ensemble learning in blood cell classification.

2. METHOD AND MATERIALS

This study proposed blood cell classification system based on ML, using Inception v3 for feature extraction. The system evaluates the effectiveness of PCA for dimensionality reduction. Four classifiers: LR, k-nearest neighbors (KNN), NN, and SVM were used with a 66%/34% training - testing division and 20-fold cross-validation. Two different tests: paired t-tests and Wilcoxon signed-rank were used to assess the impact of PCA on the performance of the model. Finally, a stacked ensemble of the best performance classifiers was implemented to improve the accuracy. Figure 1 illustrates the complete workflow.

2.1. Image acquisition

To perform the data set, a total of 11,092 peripheral blood cell images (360×363 pixels, JPG) were collected using the CellaVision DM96 analyzer at Barcelona Hospital Clinic. The images categorized eight

classes: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, erythroblasts, and platelets [23]. The cell size distributions show that platelets exhibit the most uniform morphology, whereas monocytes and immature granulocytes display the greatest variability.

2.2. Inception v3 classification

Inception v3 is deep CNN with 42-layer developed for image recognition by Google Research [24], and is pretrained on ImageNet for 1,000 classes. It improves classification, detection [25], and segmentation performance [26]. Its architecture consists of three main Inception modules: A, B, and C with parallel 1×1, 3×3, and 5×5 convolutions, supported by convolution, pooling, and fully connected layers with an auxiliary classifier used for regularization. Figure 2 shows the Inception v3 architecture. Factorized and asymmetric convolutions are used to decrease parameters and enhance feature extraction [27], while rectified linear unit (ReLU) activation, and softmax layer are used by input images to produce probabilities for eight target classes.

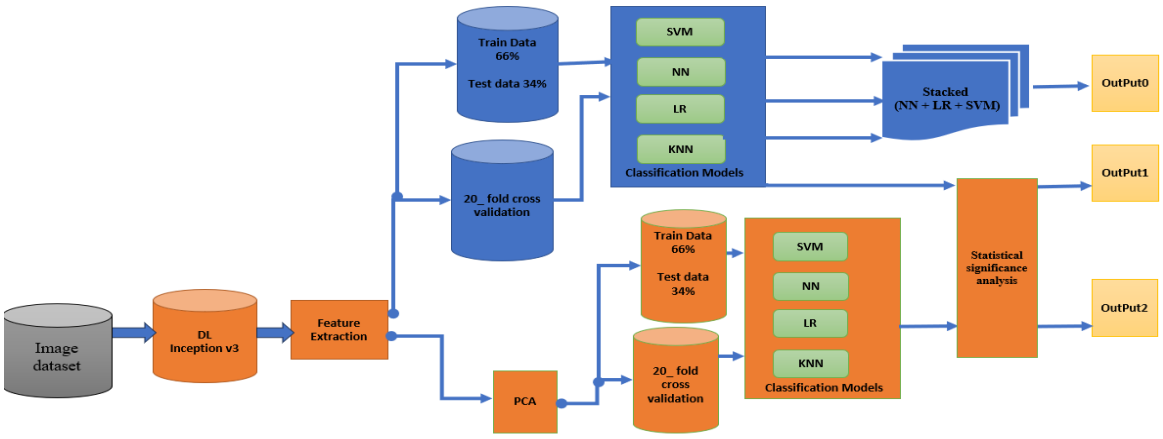


Figure 1. Proposed system for blood cell classification

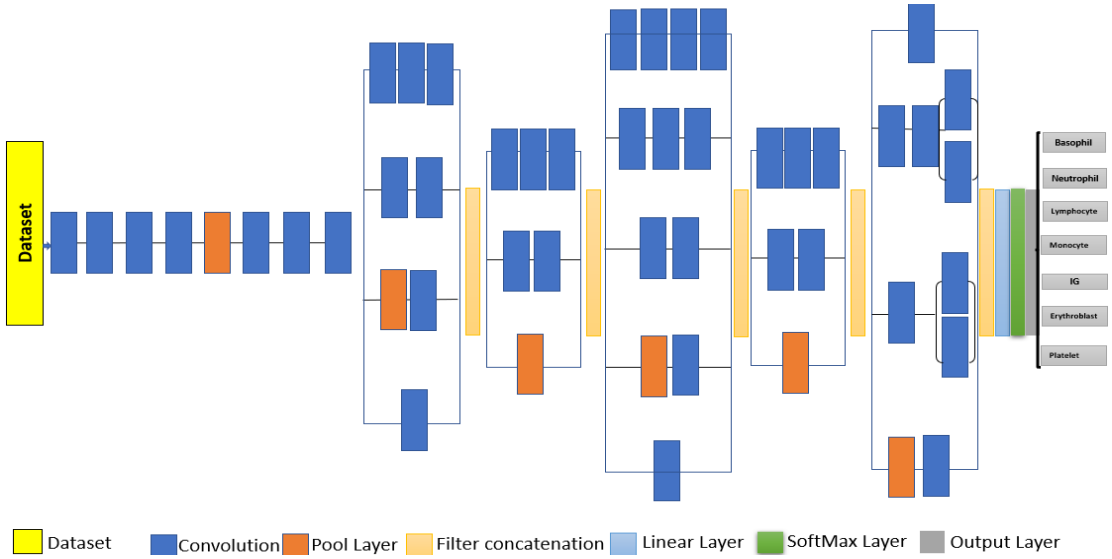


Figure 2. Inception V3 architecture

2.3. Classification algorithm in machine learning

Four ML classifiers were evaluated: LR, NN, SVM, and KNN. LR uses the logistic function (1) to models the probability of a categorical result [28], [29].

$$p(x) = \frac{e^{(b_0+b_1^* X)}}{1+e^{(b_0+b_1^* X)}} \quad (1)$$

Where $p(x)$ is predicted output, b_0 is intercept term, b_1^* coefficient for the single input value (x).

The NN classifier is a multilayer perceptron containing a fully connected hidden layers in addition to tunable parameters such as activation functions, solver type, and training epochs [30], [31]. SVM model, classifies samples by creating a maximum-margin hyperplane in a high-dimensional feature space [32], [33], using (2).

$$f(x) = w.x + b \tag{2}$$

Where x is the input feature, w is weight vector, and b is the bias term. KNN defines class labels depending on KNNs using Euclidean distance (3) [34], [35].

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

Where x, y is the feature vectors of two data points and n is the number of features.

2.4. Principal component analysis

PCA is a linear dimensionality-reduction technique, coverts high-dimensional data to a smaller set of uncorrelated components maintaining the most informative features. In this study, PCA reduced 2,048-dimensional feature vectors extracted by Inception v3 to 100 principal components called feature blood cells (FBCs) [36]. Features were first standardized using (4).

$$Z = \frac{X-\mu}{\sigma} \tag{4}$$

Then covariance between feature pairs computed using (5).

$$Con(x1, x2) = \frac{\sum_{i=1}^{n}(x1_i - \grave{x1})(x2_i - \grave{x2})}{n} \tag{5}$$

Where n is the total number of samples and $\grave{x1}$ and $\grave{x2}$ is the mean value. Finally, the principal components obtained using the eigenvalue (6).

$$AX = \lambda X \tag{6}$$

Where A is elements of features as square matrix.

2.5. K-fold cross validation method

K-fold cross-validation divides the dataset into n folds; each fold is used once as validation set while the remaining are used as training data. This process is repeated till all folds have been used for validation. By training the model to different data distributions, K-fold cross-validation reduces the probability of overfitting, enhances generalization, and provides a more stable estimation of model performance based on the average accuracy across all folds [37].

2.6. Performance evaluation metrics

Standard metrics were used to evaluate the performance of the proposed blood cell classification. Accuracy as in (7) determines the overall correctness of predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

Recall or sensitivity as in (8) refers to the proportion of actual positives correctly identified.

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

Precision as in (9) quantifies the accuracy of positive predictions.

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

F1-score as in (10) is the harmonic mean of precision and recall. It is useful for imbalanced datasets.

$$F1 - score = \frac{2 \times precision \times Recall}{Precision + Recall} \quad (10)$$

Area under the curve (AUC)-receiver operating characteristic (ROC) evaluates the ability of model to distinguish between positive and negative classes. The values of AUC ranging from 0.5 which represents random classification to 1 referring perfect classification, while the ROC curve plot of true positive rate against the false positive rate across different thresholds [34].

2.7. Statistical significance analysis

To assess whether the observed improvements in classification performance after applying PCA were statistically significant, both paired t-test and Wilcoxon signed-rank test were conducted. These tests compare the performance of models before and after PCA across multiple cross-validation folds, determining whether performance gains were due to genuine improvements rather than random variation. For each model, classification accuracy from several independent folds was used as paired samples:

i) Paired t-test: the paired t-test evaluates the mean difference between two paired samples under the assumption of normal distribution. The test statistic is computed as (11).

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (11)$$

Where $\bar{d}$ is the mean difference between paired observations, is computed as (12).

$$\bar{d} = \frac{1}{n} \sum_{1=1.}^{n} (x_i - y_i) \quad (12)$$

The stander deviation of differences represents by s_d is computed as (13).

$$s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} ((x_i - y_i) - \bar{d})^2} \quad (13)$$

Where n is the number of paired observations (folds), x_i, y_i are the accuracies before and after PCA, respectively. A small p-value ($p < 0.05$) indicates that the improvement after PCA is statistically significant [38].

ii) Wilcoxon signed-rank test: the Wilcoxon signed-rank test is a non-parametric alternative to the paired t-test, which does not assume normality. It evaluates whether the median of differences between paired samples is zero. The test statistic w is calculated as (14).

$$w = \sum_{i=1}^{n} R^+{}_i \quad (14)$$

Where $d_i = x_i - y_i$ is the difference between paired samples, $R^+{}_i$ is the rank of the absolute differences for positive differences ($d_i > 0$), ties are handled by assigning average ranks. A small p-value ($p < 0.05$) indicates that the differences are significant, supporting that PCA improved performance [39].

2.8. Stacking ensemble method

Stacking is an ensemble learning approach that improves the classification performance by integrating the outputs of multiple predictive models. Its architecture comprises of two layers. In the first one, different base models are trained separately using the same dataset, each model capture different characteristics such as linear relationships, non-linear decision boundaries, or complex feature interactions. Then the classifications which generated by these models are forwarded as input to the second layer [40].

The second layer consists of a meta-model which learns how to optimally integrate the base-model outputs. This model collects the advantages of the base classifiers while treating their individual weaknesses. To ensure generalization and prevent overfitting, the cross-validated predictions obtained from the base models are used to train the meta-model on unseen folds [41].

Mathematically, the stacking process can be expressed as (15).

$$P_1 = M_1(x), P_2 = M_2(x), \ \ \ldots, P_n = M_n(x) \quad (15)$$

Where x be the input features, and M_1, M_2, …, M_n represents the n base models. Each base model generates predictions. These predictions are concatenated to form a new feature set P (16).

$$P = [P_{1,} P_2, \ldots, P_n] \quad (16)$$

A meta-model M_{meta} is then trained on P to produce the final prediction (17).

$$\hat{Y} = M_{meta}(P) \quad (17)$$

Stacking techniques enhances predictive accuracy by taking advantage of various learning mechanisms and integrated them into a unified model, providing robustness, reduce overfitting, and improve generalization through cross-validation, and they usually outperform individual classifiers.

3. RESULTS AND DISCUSSION

The results illustrate performance differences between models before and after applying PCA. The highest accuracy was achieved by SVM across most cell types, and the stacked ensemble further improved overall performance. Detailed findings are summarized.

3.1. Impact of feature dimensionality on model performance

The performance of the ML classifiers was analyzed through two feature conditions: full deep feature with 2,048-D and, PCA reduced features with 100-D. Results were assessed using two validation strategies: a 66%/34% split. In addition, 20-fold cross-validation was applied for each feature condition.

3.1.1. Performance without PCA

Table 1 illustrates the results achieved using both evaluation strategies without PCA. For 66/34 split, the highest performance was consistently achieved by SVM with 92.6% accuracy and an AUC of 0.995, followed by NN and LR with 92.2% and 92.4% accuracy respectively and an AUC of 0.994 for each classifier. In contrast, KNN obtained the lowest performance with 73.4% accuracy, 0.935 AUC due to high-dimensional noise. Similar trends were observed using 20-fold cross-validation, where SVM achieved 93.4% accuracy with an AUC of 0. 996. Results confirmed that SVM can effectively handles high-dimensional, multi-class features, while KNN struggles in this issue.

3.1.2. Performance with PCA

PCA reduced the feature space from 2,048 to 100 components, resulting in slightly decreased in accuracy of the top performance classifiers while enhancing computational efficiency. For 66/34 split, SVM achieved 89.3% accuracy and 0.990 AUC, NN 88.0% and 0.987 AUC, LR 87.8% and 0.984 AUC, and KNN 72.9% and 0.920 AUC. In 20-fold cross-validation, SVM achieved 90.1% accuracy and 0.991 AUC, while KNN slightly improved to 74.2% accuracy due to noise reduction in the distance-based feature space. Despite the slight decrease, SVM remained the top-performing classifier across all PCA-reduced settings, confirming its reliability for blood cell classification even with reduced dimensionality. Table 2 summarizes the complete evaluation using two evaluated strategies.

Table 1. ML model performance without PCA (66%/34% split vs. 20-fold CV)

Model	Accuracy 6/34 (%)	Accuracy cross-validation (%)	AUC 66/34	AUC cross-validation %	F1-score 66/34	F1-score cross-validation %	Precision 66/34 (%)	Precision cross-validation %	Recall 66/34 (%)	Recall cross-validation %
NN	92.2	93.1	0.994	0.995	92.2	93.1	92.2	93.1	92.2	93.1
LR	92.4	92.9	0.994	0.995	92.4	92.9	92.4	93.0	92.4	92.9
SVM	92.6	93.4	0.995	0.996	92.6	93.4	92.7	93.4	92.6	93.4
KNN	73.4	73.4	0.935	0.925	72.9	73.1	74.2	74.0	73.4	73.4

Table 2. ML model performance with PCA (66%/34% split vs. 20-fold CV)

Model	Accuracy 66/34 (%)	Accuracy cross-validation (%)	AUC 66/34	AUC cross-validation %	F1-score 66/34	F1-score cross-validation %	Precision 66/34 (%)	Precision cross-validation %	Recall 66/34 (%)	Recall cross-validation %
NN	88.0	88.1	0.987	0.989	87.9	88.0	88.0	88.1	88.0	88.1
LR	87.8	88.1	0.987	0.985	87.7	88.1	87.7	88.1	87.8	88.1
SVM	89.3	90.1	0.990	0.991	89.2	90.0	89.3	90.1	89.3	90.1
KNN	72.9	74.2	0.920	0.928	72.5	73.9	73.3	74.5	72.9	74.2

The screen plot and cumulative variance curve of the PCA components are shown in figure (3), the most of the variance are concentrated in the first few principal components as shown in figure 3(a), forming a clear elbow point. Figure 3(b) illustrated that the first 100 components preserved approximately 95-98% of the total variance which reflects the effectiveness of PCA in retaining the most meaningful information while removing redundant or noisy features.

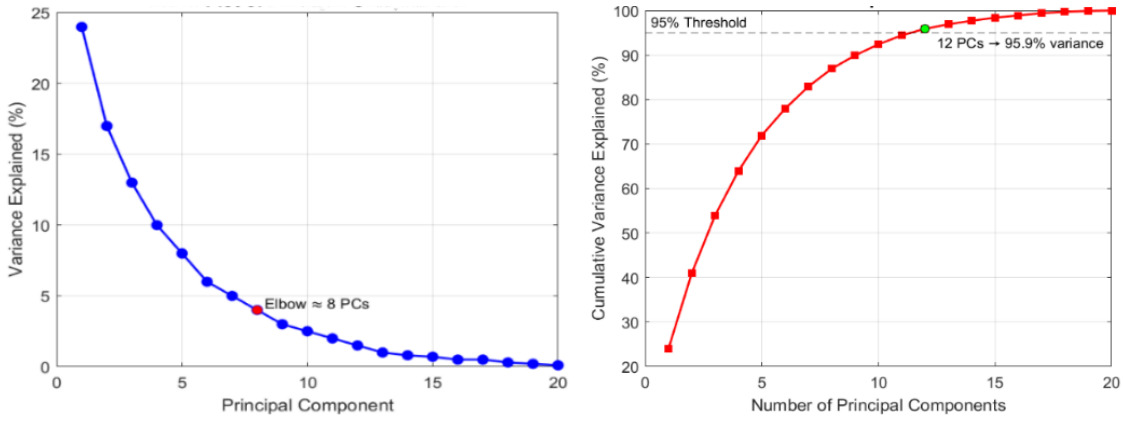


Figure 3. PCA analysis for (a) scree plot of principal components and (b) cumulative variance plot showing ~95–98% of variance retained

3.2. Class-wise analysis and probability thresholds

Figure 4 and Table 3 show the class-wise probability thresholds of SVM for all blood cell types before and after PCA. The results revealed that morphologically similar cells were affected by PCA more than distinct types. For example, probability thresholds for neutrophils, eosinophils, and erythroblasts were slightly reduced with PCA (0.975 vs. 0.987, 0.976 vs. 0.994, 0.969 vs. 0.983, respectively). Monocytes and immature granulocytes recorded the most notable drops with (0.862 vs. 0.950) and (0.913 vs. 0.955) respectively, due to their morphological similarity to other WBCs. These results indicate PCA capability to preserve the majority of discriminative features but it may discard subtle details for similar cell types.

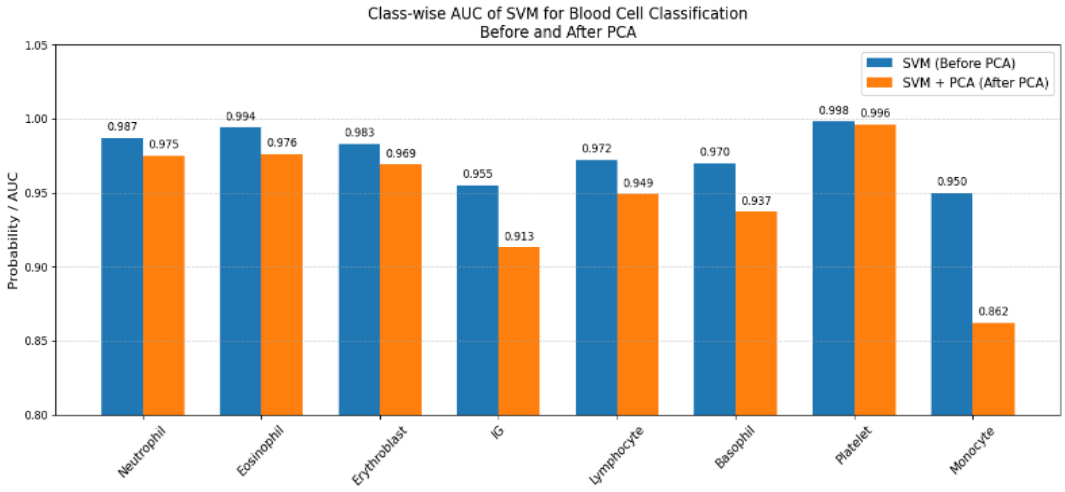


Figure 4. The class-wise predicted probabilities of the SVM model before and after PCA

Table 3. Probability of SVM model with all classes

Classes	Neutrophil	Eosinophil	Erythroblast	Immature granulocytes	Lymphocyte	Basophil	Platelet	Monocyte
SVM+PCA	0.975	0.976	0.969	0.913	0.949	0.937	0.996	0.862
SVM	0.987	0.994	0.983	0.955	0.972	0.97	0.998	0.95

Figure 5 illustrates class-wise ROC curves and AUC values for SVM before and after PCA using 20-fold cross-validation. Figure 5(a) corresponds to erythroblast, Figure 5(b) to neutrophil, Figure 5(c) to eosinophil, Figure 5(d) to immature granulocytes, Figure 5(e) to lymphocyte, Figure 5(f) to basophil, Figure 5(g) to platelet, and Figure 5(h) to monocyte. This figure highlights the ability of the model to discriminate between blood cell type and confirms that PCA slightly reduces performance for morphologically similar cells.

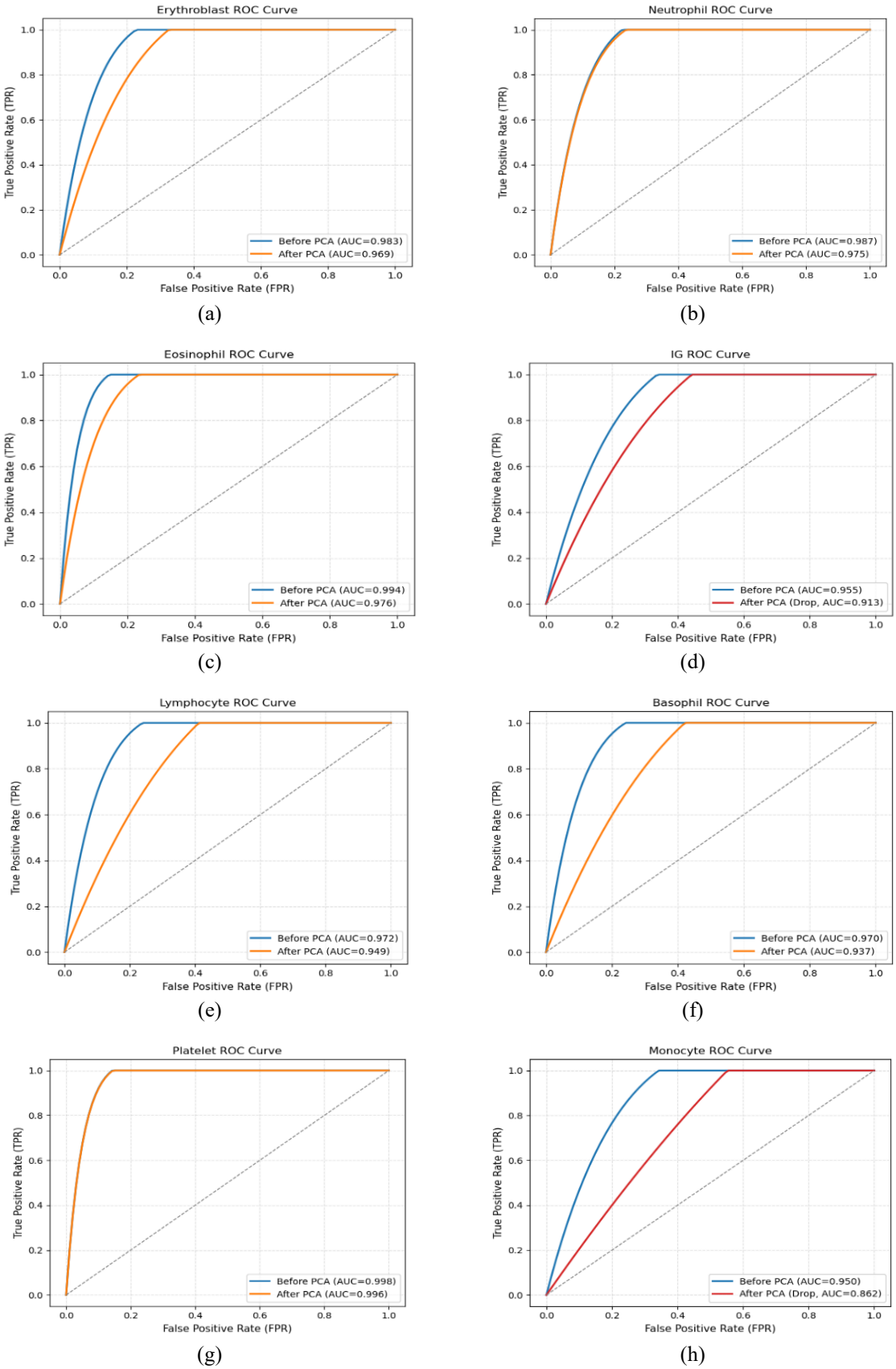


Figure 5. Class-wise ROC curves and AUC values for SVM before and after PCA using 20-fold cross-validation for (a) erythroblast, (b) neutrophil, (c) eosinophil, (d) IG, (e) lymphocyte, (f) basophil, (g) platelet, and (h) monocyte

3.3. Statistical evaluation

3.3.1. Simulated paired sample assessment

To accurately evaluate the effect of PCA on the performance of the models, a simulated paired sample evaluation: mean accuracy and standard deviation (mean±SD) was conducted. This method approximates the expected variability of each classifier before and after PCA. This simulation provides a statistical significance testing, even when only summary statistics are available, providing insight into whether the differences in observed accuracy are consistent or due to random variation. Table 4 summarizes the paired evaluations.

The results demonstrated that PCA affected on the performance on most models. After applying PCA NN, LR, and SVM showed an approximately 3-5% consistent decrease in accuracy. Paired t-test and Wilcoxon analyses emphasized the statistical importance of this reduction with $p < 0.05$. This illustrates that although PCA is successful in reducing feature redundancy and computational complexity, it may also remove discriminative features which are necessary for optimal classification. According to KNN, it was no significant change in its performance, confirming its robustness to dimensionality reduction. These findings underscore that balancing the dimensionality reduction with classification performance is essential in biomedical image analysis.

Table 4. Simulated paired sample evaluation before and after PCA

Model	Without PCA 66/34 split	Without PCA cross-validation	With PCA 66/34 split	With PCA cross-validation
NN	92.2±0.8	93.1±0.6	88.0±1.0	88.1±0.9
LR	92.4±0.7	92.9±0.5	87.8±1.1	88.1±0.8
SVM	92.6±0.6	93.4±0.4	89.3±0.9	90.1±0.7
KNN	73.4±1.2	73.4±1.0	72.9±1.5	74.2±1.3

3.3.2. Statistical significance tests

In order to accurately evaluate the impact of PCA, paired t-tests and Wilcoxon signed-rank tests were applied to the results obtained from 20-fold cross-validation. Table 5 summarizes the statistical significance results. It illustrates that the accuracy of NN, LR, and SVM was significantly decreased, while the performance of KNN remains unaffected.

Table 5. Statistical significance analysis of PCA impact on ML model performance

Model	Mean accuracy (without PCA cross-validation)	Mean accuracy (with PCA cross-validation)	Paired t-test p	Wilcoxon p	Significant
NN	93.1	88.1	0.013	0.012	Yes
LR	92.9	88.1	0.016	0.014	Yes
SVM	93.4	90.1	0.045	0.041	Yes
KNN	73.4	74.2	0.462	0.456	No

3.4. Stacking ensemble performance

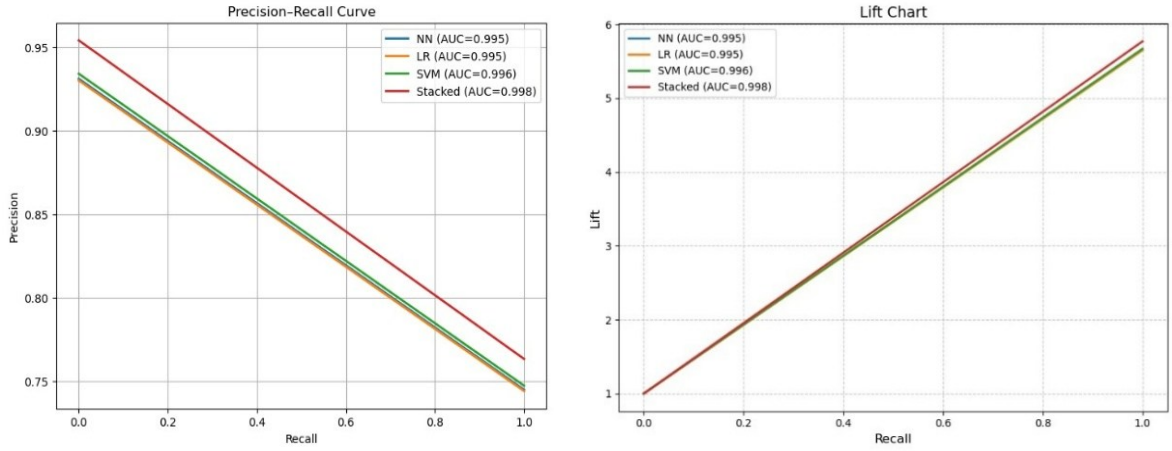
A stacking ensemble was implemented to improve the robustness and classification accuracy of the model. The stacked ensemble used in this study combines NN, LR, and SVM. Using this technique minimize bias and variance by leveraging the complementary strengths of the base models. A comparison between individual classifiers and the stacked ensemble is presented in Table 6.

Table 6. Performance comparison of individual and stacked ML models

Model	AUC	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
NN	0.995	93.1	93.1	93.1	93.1
LR	0.995	92.9	92.9	93.0	92.9
SVM	0.996	93.4	93.4	93.4	93.4
Stacked (NN+LR+SVM)	0.998	95.2	95.1	95.4	95.0

The stacked ensemble surpassed all individual classifiers, achieving an AUC of 0.998 and an accuracy of 95.2%, with consistent improvements across F1-score, precision, and recall. These results indicate that the performance of the ensemble is balanced in the classification of blood cell classes, including rare types. Figure 6 illustrates the evaluation of model performance for the stacked ensemble compared to individual models, using both the precision–recall curve shown in Figure 6(a) and the lift probability threshold plot as in Figure 6(b), showing that the stacked retained the stacked ensemble surpassed all individual classifiers, achieving an AUC of 0.998 and an accuracy of 95.2%, with consistent

improvements across F1-score, precision, and recall. These results indicate that the performance of the ensemble superior precision across varying recall thresholds and achieved a greater lift, which confirms that stacking effectively integrates the predictive advantages of NN, LR, and SVM into a more stable and accurate classification framework.



(a) (b)

Figure 6. Performance evaluation of the stacked ensemble compared to individual models for (a) precision-recall curve and (b) lift probability threshold

3.5. Comparison with previous studies

Table 7 summarizes related blood cell classification studies. The proposed system outperforms previous studies, achieving 95.2% accuracy and an AUC of 0.998 using InceptionV3 features with a stacked NN, LR, and SVM ensemble. It demonstrates strong robustness under 20-fold cross-validation, and PCA-based feature reduction caused minimal performance loss, highlighting its accuracy, efficiency, and practical relevance.

Table 7. Comparative summary of blood cell classification studies

Reference	Method/Model	Discussion
Ferreira *et al.* [42], 2025	CNN with contrast limited adaptive histogram equalization (CLAHE)	Achieved up to 98.72% test accuracy for lymphocytes and monocytes; CLAHE improved image contrast and enhanced CNN classification performance.
Sazak and Kotan [43], 2024	YOLOv11 with optimized weights	Achieved mean average precision (mAP) of 93.8%; data augmentation and optimized weights enhanced model robustness and precision.
Mahale [44], 2025	Deep learning models (Inception, VGG16, capsule networks)	Inception achieved the best accuracy and precision among tested models, highlighting the strength of deep feature extraction for blood cell classification.
Tseng and Huang [45], 2022	Ensemble of CNNs	Achieved average testing accuracy of 90.1% for neutrophil classification; ensemble learning improved generalization across multiple datasets.
This work	InceptionV3 feature extraction+Stacked (NN+LR+SVM), PCA, 20-fold CV	The stacked ensemble achieved the best performance with 95.2% accuracy and an AUC of 0.998.

4. CONCLUSION

The study illustrates the effectiveness of using SVM combined with 20-fold cross-validation in achieving high accuracy for peripheral blood cell types classification. Without using PCA, SVM achieved the highest prediction probabilities across most classes, including platelets (0.998), eosinophils (0.994), and neutrophils (0.987). Applying PCA resulted in different impacts across different cell types. Monocytes and immature granulocytes recorded the most notable declines (0.862 vs. 0.950) and (0.913 vs. 0.955) respectively, which indicate their higher sensitivity to dimensionality reduction due to their morphological similarity to other leukocytes. Moderate reductions were also observed for basophils, lymphocytes and erythroblasts with (0.937 vs. 0.970), (0.949 vs. 0.972), and (0.969 vs. 0.983) respectively. These reductions were confirmed to be statistically significant for SVM, NN, and LR ($p < 0.05$), indicating that PCA may remove subtle

discriminative features required by more complex models. Despite these effects, SVM remained the best overall performance with 92.6% accuracy before applying PCA and 90.1% after PCA. Additionally, using a stacked ensemble by combining NN, LR, and SVM, produced the highest performance overall with AUC =0.998; accuracy =95.2%, demonstrating that the integration of deep feature extraction with tradition ML improves both robustness and generalization. Furthermore, integrating NN, LR, and SVM within a stacked ensemble produced the highest performance overall (AUC =0.998; accuracy =95.2%), demonstrating that combining deep feature extraction with classical ML improves both robustness and generalization.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Marwa Mawfaq Mohamedsheet Al-Hatab	✓	✓		✓			✓	✓	✓		✓		✓	
Maysaloon Abed Qasim	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Nawar A. Sultan	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/unclesamulus/blood-cells-image-dataset, reference [23].

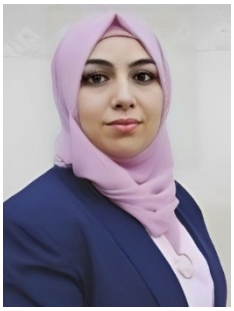
REFERENCES

[1] M. Shahzad *et al.*, "Blood cell image segmentation and classification: a systematic review," *PeerJ Computer Science*, vol. 10, Feb. 2024, doi: 10.7717/peerj-cs.1813.
[2] M. Zolfaghari and H. Sajedi, "A survey on automated detection and classification of acute leukemia and WBCs in microscopic blood cells," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6723–6753, Feb. 2022, doi: 10.1007/s11042-022-12108-7.
[3] D. K. Baruah, "Application of CNN in blood smeared images: a review," *Revista Electronica De Veterinaria*, vol. 25, no. 1, pp. 3458–3465, Dec. 2024, doi: 10.69980/redvet.v25i1.1595.
[4] S. Khan, M. Sajjad, T. Hussain, A. Ullah, and A. S. Imran, "A review on traditional machine learning and deep learning models for WBCs classification in blood smear images," *IEEE Access*, vol. 9, pp. 10657–10673, 2021, doi: 10.1109/ACCESS.2020.3048172.
[5] H. Chen *et al.*, "Accurate classification of white blood cells by coupling pre-trained ResNet and DenseNet with SCAM mechanism," *BMC Bioinformatics*, vol. 23, no. 1, 2022, doi: 10.1186/s12859-022-04824-6.
[6] E. H. Houssein *et al.*, "Using deep DenseNet with cyclical learning rate to classify leukocytes for leukemia identification," *Frontiers in Oncology*, vol. 13, Sep. 2023, doi: 10.3389/fonc.2023.1230434.
[7] G. Atteia, R. Alnashwan, and M. Hassan, "Hybrid feature learning based PSO PCA feature engineering approach for blood cancer classification," *Diagnostics*, vol. 13, no. 16, Aug. 2023, doi: 10.3390/diagnostics13162672.
[8] D. Tellez *et al.*, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, Dec. 2019, doi: 10.1016/j.media.2019.101544.
[9] M. Hussein and F. A. E.-S. Z. El-Mougi, "Integrating deep learning and transfer learning: optimizing white blood cells classification in medical educational institutions," *Journal of Big Data*, vol. 12, no. 1, Jul. 2025, doi: 10.1186/s40537-025-01235-1.
[10] H. Üzen and H. Fırat, "A hybrid approach based on multipath Swin transformer and ConvMixer for white blood cells classification," *Health Information Science and Systems*, vol. 12, no. 1, Apr. 2024, doi: 10.1007/s13755-024-00291-w.
[11] S. Parayil, A. Debnath, V. Perumal, A. S. Narayanan, B. K. Natarajan, and S. S. N. Bose, "Digital red blood cell morphology analysis and multi-class classification powered by hierarchical AI," in *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Jul. 2024, pp. 1–6, doi: 10.1109/CONECCT62155.2024.10677056.

[12] S. Banerjee and S. S. Chaudhuri, "Total contribution score and fuzzy entropy based two-stage selection of FC, ReLU and inverseReLU features of multiple convolution neural networks for erythrocytes detection," *IET Computer Vision*, vol. 13, no. 7, pp. 640–650, Oct. 2019, doi: 10.1049/iet-cvi.2018.5545.
[13] S. Ghosh, M. Majumder, and A. Kudeshia, "LeukoX: leukocyte classification using least entropy combiner (LEC) for ensemble learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 8, pp. 2977–2981, Aug. 2021, doi: 10.1109/TCSII.2021.3064389.
[14] L. Ichim, C.-A. Iordan, and D. Popescu, "Multi network blood cell classification system based on decision fusion," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Nov. 2022, pp. 1–6, doi: 10.1109/ICECCME55909.2022.9987906.
[15] J. J. Tenguam, L. H. D. C. Longo, A. B. Silva, P. R. De Faria, M. Z. Do Nascimento, and L. A. Neves, "Classification of H&E images exploring ensemble learning with two-stage feature selection," in *2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jun. 2022, pp. 1–4, doi: 10.1109/IWSSIP55020.2022.9854418.
[16] A. H. Routt, N. Yang, N. Z. Piety, M. Lu, and S. S. Shevkoplyas, "Deep ensemble learning enables highly accurate classification of stored red blood cell morphology," *Scientific Reports*, vol. 13, no. 1, Feb. 2023, doi: 10.1038/s41598-023-30214-w.
[17] R. Ahmad, M. Awais, N. Kausar, and T. Akram, "White blood cells classification using entropy-controlled deep features optimization," *Diagnostics*, vol. 13, no. 3, Jan. 2023, doi: 10.3390/diagnostics13030352.
[18] M. Awais, R. Ahmad, N. Kausar, A. I. Alzahrani, N. Alalwan, and A. Masood, "ALL classification using neural ensemble and memetic deep feature optimization," *Frontiers in Artificial Intelligence*, vol. 7, Apr. 2024, doi: 10.3389/frai.2024.1351942.
[19] L. Cai *et al.*, "Towards cross-domain single blood cell image classification via large-scale LoRA-based segment anything model," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2024, vol. 13, no. 3, pp. 1–5, doi: 10.1109/ISBI56570.2024.10635629.
[20] J. Sharma, "Enhanced blood cell classification using convolutional neural networks for accurate diagnosis of haematological conditions," in *2024 Global Conference on Communications and Information Technologies (GCCIT)*, Oct. 2024, pp. 1–5, doi: 10.1109/GCCIT63234.2024.10862486.
[21] W. Tepakhan, W. Srisintorn, T. Penglong, and P. Saelue, "Machine learning approach for differentiating iron deficiency anemia and thalassemia using random forest and gradient boosting algorithms," *Scientific Reports*, vol. 15, no. 1, May 2025, doi: 10.1038/s41598-025-01458-5.
[22] Z. Uzma and R. Afsar, "Improving white blood cell classification with minimal labels: a semi-supervised learning framework," *International Journal of Engineering Research and Science & Technology*, vol. 21, no. 3, pp. 1523–1528, Aug. 2025, doi: 10.62643/ijerst.v21.n3(1).pp1523-1528.
[23] U. Samulus, "Blood cells image dataset," *Kaggle*. 2022. Accessed: Jan. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/unclesamulus/blood-cells-image-dataset
[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
[25] M. Shoaib and N. Sayed, "YOLO object detector and Inception-V3 convolutional neural network for improved brain tumor segmentation," *Traitement du Signal*, vol. 39, no. 1, pp. 371–380, Feb. 2022, doi: 10.18280/ts.390139.
[26] S. Likhitha and R. Baskar, "Skin cancer segmentation using R-CNN comparing with Inception V3 for better accuracy," in *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Dec. 2022, pp. 1293–1297, doi: 10.1109/SMART55829.2022.10047686.
[27] O. I. -Villanueva, V. G. -Ponce, O. R. Paredes, F. S. -Liñan, J. Z. -Paulini, and M. C. -Carbonell, "Convolutional neural networks with transfer learning for pneumonia detection," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 544–551, 2022, doi: 10.14569/IJACSA.2022.0130963.
[28] D. Saini, T. Chand, D. K. Chouhan, and M. Prakash, "A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 419–444, Apr. 2021, doi: 10.1016/j.bbe.2021.03.002.
[29] R. H. M. Ameen, N. M. Basheer, and A. K. Younis, "A survey: breast cancer classification by using machine learning techniques," *NTU Journal of Engineering and Technology*, vol. 2, no. 1, May 2023, doi: 10.56286/ntujet.v2i1.367.
[30] E. M. -Tzanakou, H. Sheikh, and B. Zhu, "Neural networks and blood cell identification," *Journal of Medical Systems*, vol. 21, no. 4, pp. 201–210, Aug. 1997, doi: 10.1023/A:1022899519704.
[31] P. Tabesh, G. Lim, S. Khator, and C. Dacso, "A support vector machine approach for predicting heart conditions," in *IIE Annual Conference and Expo 2010 Proceedings*, 2010, pp. 916–921.
[32] S. Q. Hasan, "Shallow model and deep learning model for features extraction of images," *NTU Journal of Engineering and Technology*, vol. 2, no. 3, Nov. 2023, doi: 10.56286/ntujet.v2i3.449.
[33] S. H. Shetty, S. Shetty, C. Singh, and A. Rao, "Supervised machine learning: algorithms and applications," in *Fundamentals and Methods of Machine and Deep Learning*, Wiley, 2022, pp. 1–16, doi: 10.1002/9781119821908.ch1.
[34] L. P.-Coelho, "How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications," *Bioengineering*, vol. 10, no. 12, Dec. 2023, doi: 10.3390/bioengineering10121435.
[35] T. A. J. Ali, S. A. Altutunji, and L. A. Khalaf, "Sonar data classification using deep learning techniques," in *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)*, 2025, pp. 1–7, doi: 10.1109/ICBATS66542.2025.11258274.
[36] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Apr. 2016, doi: 10.1098/rsta.2015.0202.
[37] M. Bhagat and B. Bakariya, "A comprehensive review of cross-validation techniques in machine learning," *International Journal on Science and Technology*, vol. 16, no. 1, Jan. 2025, doi: 10.71097/IJSAT.v16.i1.1305.
[38] T. Rietveld and R. van Hout, "The paired t test and beyond: recommendations for testing the central tendencies of two paired samples in research on speech, language and hearing pathology," *Journal of Communication Disorders*, vol. 69, pp. 44–57, Sep. 2017, doi: 10.1016/j.jcomdis.2017.07.002.
[39] K. Okoye and S. Hosseini, "Wilcoxon statistics in R: signed-rank test and rank-sum test," in *R Programming: Statistical Data Analysis in Research*, 2024, pp. 279–303, doi: 10.1007/978-981-97-3385-9_13.
[40] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning," *Sustainability*, vol. 14, no. 21, Oct. 2022, doi: 10.3390/su142113998.
[41] X. Yin, Q. Liu, Y. Pan, X. Huang, J. Wu, and X. Wang, "Strength of stacking technique of ensemble learning in rockburst prediction with imbalanced data: comparison of eight single and ensemble models," *Natural Resources Research*, vol. 30, pp. 1795–1815, 2021, doi: 10.1007/s11053-020-09787-0.

[42] G. C. Ferreira, L. C. Ishiuchi, L. H. M. Teixeira, E. D. de Oliveira, and W. D. Parreira, "Aprimoramento da classificação de linfócitos e monócitos em imagens médicas: o impacto do CLAHE em redes neurais convolucionais," in *Computer on the Beach*, 2025, pp. 199–212, doi: 10.14210/cotb.v16.p199-212.
[43] H. Sazak and M. Kotan, "Automated blood cell detection and classification in microscopic images using YOLOv11 and optimized weights," *Diagnostics*, vol. 15, no. 1, 2024, doi: 10.3390/diagnostics15010022.
[44] P. P. Mahale, "Enhancing hematological diagnostics: deep learning models for human blood cell classification," *Journal of Information Systems Engineering and Management*, vol. 10, no. 23s, pp. 449–459, Mar. 2025, doi: 10.52783/jisem.v10i23s.3717.
[45] T. Tseng and H. Huang, "Classification of peripheral blood neutrophils using deep learning," *Cytometry Part A*, vol. 103, no. 4, pp. 295–303, 2023, doi: 10.1002/cyto.a.24698.

BIOGRAPHIES OF AUTHORS



Marwa Mawfaq Mohamedsheet Al-Hatab received the B.Sc. degree in Medical Instruments Engineering from Technical Engineering College, Mosul, Northern Technical University, Mosul, Iraq, and the M.Sc. degree in Biomedical Engineering from Universita Politecnica delle Marche. She is a lecturer at Technical Engineering College, Mosul, Northern Technical University, Mosul, Iraq. She has published multiple research paper. She can be contacted at email: marwa.alhatab@ntu.edu.iq.

Dr. Maysaloon Abed Qasim received the B.Sc. degree in Electrical Engineering from Mosul University, Iraq, the M.Sc. degree in Electronic and Communication from Mosul University, Iraq, and the Ph.D. degree Electronic and Computer Engineering from Istanbul, Turkey. She is assistant professor at Technical Engineering College for Computer and Artificial Intelligence, Northern Technical University, Mosul, Iraq. She has published multiple research papers. She can be contacted at email: maysloon.alhashim@ntu.edu.iq.

Dr. Nawar A. Sultan received the M.A. degree in Computer Sciences from the University of Mosul, Iraq, in 2011, and the Ph.D. degree in Computer Sciences from the same university in 2024. He is currently a lecturer and head of the Department of Cyber Security Techniques Engineering at Technical Engineering College for Computer and Artificial Intelligence, Northern Technical University, Mosul, Iraq. His research interests include information retrieval, big data, blockchain, container technology, cloud computing, artificial intelligence, machine learning, and distributed computing. He has published multiple research papers and actively contributes to the international scientific community. He can be contacted at email: nawarabd@ntu.edu.iq.