EfficientSwin: A Hybrid Model for Blood Cell Classification with Saliency Maps Visualization

Tanviben Patel, Hoda El-Sayed, Md Kamruzzaman Sarker
Bowie State University, Bowie, MD, USA
patelt0902@student.bowiestate.edu, helsayed@bowiestate.edu, ksarker@bowiestate.edu

***Abstract*—Blood cell (BC) classification holds significant importance in medical diagnostics as it enables the identification and differentiation of various types of BCs, which is crucial for detecting specific infections, disorders, or conditions, and guiding appropriate treatment decisions. Accurate BC classification simplifies the evaluation of immune system performance and the diagnosis of various ailments such as infections, leukemia, and other hematological disorders. Deep learning algorithms perform excellently in the automated identification and differentiation of various types of BCs. One of the advanced deep learning models, EfficientNet has shown remarkable performance with limited datasets, another model Swin Transformer's capability to capture intricate patterns and features makes it more accurate, albeit with limitations due to its large number of parameters. However, medical image datasets are often limited, necessitating a solution that balances accuracy and efficiency. To address this, we propose a novel hybrid model, by combining the strengths of these two models. We first fine-tuned the Swin Transformer on a blood cell dataset comprising wihite blood cells, red blood cells and platelets, achieving promising outcomes. Subsequently, our hybrid model, EfficientSwin, outperformed the standalone Swin Transformer, achieving an impressive 98.14% accuracy in BCs classification. Furthermore, we compared our approach with previous research on white blood cell datasets, showcasing the superiority of EfficientSwin in accurately classifying blood cells. We also employed saliency maps for a visual representation of our classification results, further illustrating the efficacy of our approach.**

I. Introduction

Blood cell analysis constitutes a fundamental component within the medical field, necessitating intricate systems, costly chemical reagents, time-intensive protocols, and personnel with specialized training for its execution. [1], [2] This type of analysis is essential for diagnosing a wide range of diseases, such as anemia, leukemia, malaria, various infections, and blood cancers. According to Huang, Le-Tian, et al. [3], variations in the profiles of peripheral blood cells, especially changes in leukocytes, lymphocytes, neutrophils, and overall cell counts, have been linked to Alzheimer's Disease. Traditional methods of blood cell analysis rely heavily on specialized expertise and substantial resources. [4]

CAD has become one of the major research subjects in medical imaging and diagnostic radiology. [5]–[7] A variety of automated methods have been developed for segmenting, classifying, and detecting blood cell images within CAD systems, tackling the inherent challenges of microscope image analysis. These methods span across image and signal processing, machine learning, and deep learning techniques. [8], [9] Notably, deep learning semantic segmentation, a state-of-the-art approach, has been utilized to segment red blood cells (RBCs) and white blood cells (WBCs) in blood smear images, [10] achieving an accuracy rate of 89.45%. However, certain cells were not segmented properly due to overlapping. Convolutional Neural Network (CNN) models, a subset of deep learning, have been applied for the automated diagnosis and prognosis of blood cells. Despite their capacity to detect fine details and patterns, thus aiding in the accurate identification of anomalies and medical conditions, these models encounter difficulties in feature extraction and pattern recognition when processing microscopic blood cell images.

Yu, Kaixin, et al., [11] highlighted the effectiveness of the Swin Transformer model in deep feature extraction, noting its superiority in processing large datasets, which underpins its capability in feature extraction. Conversely, as Yao, Wenjian, et al., [12] point out in their review, Swin Transformer models require combination with CNNs to yield accurate results when working with small datasets, indicating the limitations of relying solely on transformer models for feature extraction in smaller datasets.

In addressing the challenges faced by medical imaging classification models, we conducted a comparative analysis of two distinct approaches: the Swin Transformer, which is based on transformer architectures, and EfficientNet, which stems from the Convolutional Neural Network (CNN) framework. In the related work section, we discuss previous studies, detailing their methodologies and findings on blood cell (BC) datasets. The architecture of our proposed model is outlined in the methodology section, where we also provide a step-by-step explanation of our approach. We conducted a thorough comparison of our method's results against those of previous studies and baseline models, offering a comprehensive overview of the enhancements and benefits our method introduces. This comparative analysis, which includes an examination of saliency maps from our proposed model, is elaborated upon in the results section. Our study covers a total of eight types of blood cells.

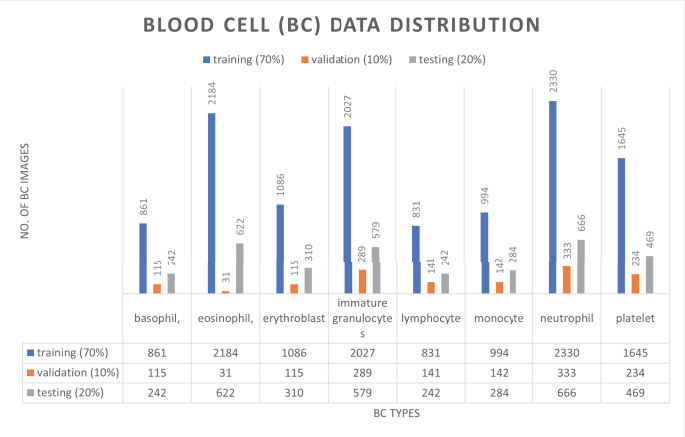
The development of a hybrid neural network model, combining EfficientNet and Swin Transformer, marks a significant advancement in predicting BC types, offering notable contributions in several key areas: 1) Integration of Features: Our research introduces a model that capitalizes on the strengths of both Swin Transformer and EfficientNet for lesion diagnosis. There are three major contribution of this study are 2) The

synergy of the hybrid model markedly elevates prediction accuracy over that of singular models, underscoring the benefit of this combined approach. 3) The model's efficacy and practicality in accurately identifying different BCs types were confirmed through extensive testing on publicly available datasets and comparison with other leading deep learning models. 4) The study employs saliency maps to visually interpret how the model processes image classification, offering insights into its decision-making process.

II. Related work

A. Dataset

In this study, we have used BloodMNIST dataset. [13] It is also accessible through the GitHub link https://github.com/MedMNIST/MedMNIST. The BloodMNIST dataset, sometimes referred to as "bloodmnist," is a set of medical images that show individual normal blood cells. These images were obtained from individuals who were free from infection, hematologic disorders, or oncologic diseases, and had not undergone any pharmacologic treatment at the time of blood collection. There are total 17,092 images which are divided in 8 different classes. These classes correspond to different types of blood cells, including basophils, eosinophils, erythroblasts, immature granulocytes (myelocytes, metamyelocytes, and promyelocytes), lymphocytes, monocytes, neutrophils, and platelets (Showing in Fig. 1).



	basophil,	eosinophil,	erythroblast	immature granulocytes	lymphocyte	monocyte	neutrophil	platelet
training (70%)	861	2184	1086	2027	831	994	2330	1645
validation (10%)	115	31	115	289	141	142	333	234
testing (20%)	242	622	310	579	242	284	666	469

Fig. 1. BloodMINST dataset distribution over eight different classes of the human peripheral blood cell (HPBC) images. The images were randomly split per class into 70% (training), 10% (validation), and 20% (testing).

The dataset is divided into three subsets for testing, validation, and training, with a ratio of 7:1:2. After center-cropping the original images to 3×200×200 pixels and resizing them to a final resolution of 3×28×28 pixels, the 3×360×363 pixel images underwent preprocessing.

B. Literature review of blood cells classification

In order to address a major challenge in medical imaging, the study presents [14] W-Net as inventive CNN-based architecture created for the classification of white blood cells (WBC). Its evaluation on a large-scale dataset with 6562 real WBC images showed that it performed exceptionally well, with accuracy of 97%. In WBC classification, this accuracy significantly outperforms the current convolutional neural network (CNN) and recurrent neural network (RNN) based models. Moreover, a noteworthy development is W-Net's flexibility in transfer learning scenarios, which permits customization for particular tasks or integration with various datasets. Using a Generative Adversarial Network to create synthetic WBC images is another important contribution. In this paper [15], they also introduce a new classification method using CNNs and transfer learning. It solved a significant challenge in hematological diagnostics: the difficulty of morphologically differentiating various normal and abnormal blood cells. They used two CNN, VGG16 and Inceptionv3. One uses the networks as feature extractors for a support vector machine (SVM) classifier, while other fine-tunes networks for end-to-end classification. The overall accuracy of the classification model reached up to 96.2%. The paper's main contribution lies in its development of an end-to-end classifier capable of discriminating between eight cell types, trained on a substantial dataset sourced from clinical practice.

This study [16] explores the sophisticated use of deep learning for the fine-grained classification of leukocytes in the medical domain. The intricacy of leukocyte classification is too complex for conventional machine learning methods, like SVM classifiers, especially when there are up to 40 categories to choose from. In order to overcome these difficulties, the authors build a deep residual neural network (ResNet) that robustly extracts salient features and emulates the cell recognition method used by domain experts. After carefully modifying its architecture in light of previous experience with white blood cell tests, the network is trained on an extensive dataset consisting of almost 100,000 labeled leukocytes in 40 different categories. Combining training strategies allows for improved generalization. The average accuracy on test dataset was 76.84%. These outcomes mainly focus on complexity and variability of leukocyte categories. Sharma et al.'s paper [17] utilized CNN, including a bespoke five-layer CNN known as "LeNet-5", along with established models like "VGGs", "Inception V3", and "Xception", for classifying white blood cells. They addressed the BCCD dataset, which initial comprised only 349 images of low quality, spread across four white blood cell categories: monocyte, lymphocyte, neutrophil, and eosinophil. Utilizing various techniques for data augmentation, they considerably expanded the dataset to encompass more than 3000 photos for every category. After that, groups for testing and training were created using this improved dataset. Across all four types of white blood cells, the average classification accuracy was a noteworthy 87%.

A methodology for the identification, localization, and classification of leukocytes has been proposed by Zhao et al. [18]. They utilized databases from Cella vision and Jiashan for the detection task, while the ALL-IDB database was employed for classification purposes. The workflow commenced with the initial identification of white blood cells (WBCs)

using morphological techniques, followed by the utilization of color and granularity features for classification. To be more precise, an SVM Classifier was used to differentiate between the eosinophil and basophil classes. In contrast, the remaining classes—neutrophils, lymphocytes, and monocytes—were classified using a hybrid model that combined a CNN (Convolutional Neural Network) and random forest. This all-encompassing method produced an astounding accuracy rate of 92.8%. Ma, Li, et al. [19] paper introduce an innovative leukocyte classification architecture combining a Deep Convolutional Generative Adversarial Network (DC-GAN) with ResNet that overcomes the drawbacks of conventinal WBC categorization techniques that rely on cell segmentation and feature extraction. Due to problems like insufficient data or class imbalances in deep learning applications, traditional approaches frequently have poor accuracy because of inadequate segmentation. The proposed DC-GAN improves uncertainty estimation by producing synthetic images, which strengthens the model's capacity to handle out-of-distribution inputs. To improve robustness and accelerate model convergence, transfer learning is used. The ResNet's modified loss function, an advancement over the standard softmax loss, enables more effective learning of WBC image characteristics, resulting in a superior classification model with accuracy 91.7%.

Yang et al. [20] utilized an identical dataset partitioning for both training and testing phases. Initially, they employed ResNet-18 [21] on two distinct image resolutions: 28 x 28 and 224 x 224. Subsequently, ResNet-50 was applied to the same resolutions [21]. In a parallel vein, Feurer et al. [22] employed automated machine learning (Auto-sklearn) for training and validation. Following this, Auto-keras by Haifeng et al. [23] was employed on the identical dataset. Finally, the authors introduced their proposed methodology, Google AutoML, which yielded the highest accuracy.

III. Methodology

A. The architecture of the hybrid model with EfficinetNet and Swin Transformer

In this study, we used PyTorch, torchvision, TIMM (PyTorch Image Models), along with other python libraries. To ensure the reproducibility of our results, we implemented deterministic behaviors by fixing the random seed and disabling certain CUDA optimizations, which are necessary for the consistent and faster outcomes across runs.

We initiated our process by defining the dataset and extracting label information, followed by preparing data transformations. This involved resizing images to 224x224 pixels, converting them into tensor format, and normalizing them for optimal processing. Subsequently, we downloaded and loaded the dataset's training, validation, and test splits. For each dataset split, we initialized DataLoaders with a batch size of 32, incorporating shuffling for the training set to ensure the randomness of input data order.

Our approach included the definition of a custom hybrid deep learning model that synergizes the capabilities of Swin Transformer and EfficientNet models. Leveraging pre-trained models on ImageNet, we omitted their final classification layers and introduced a linear classifier to make predictions based on the concatenated features from both models. The architecture of our proposed model is illustrated in Fig. 2.

Training was conducted on a single V100 GPU, utilizing CrossEntropyLoss for multi-class classification and Adam optimizer with a learning rate of 0.001. To enhance training efficiency, we implemented an early stopping mechanism, set to trigger if the validation accuracy did not improve after five epochs.

We commenced the training for a predetermined set of 35 epochs, during which we performed forward and backward passes and updated the model weights accordingly. After each epoch, we evaluated the model on the validation set to monitor its performance, making decisions regarding early stopping based on these evaluations. To visually track the learning progress, we plotted the training loss and accuracies (both training and validation) as functions of epoch number. Upon completion of the training or upon triggering the early stopping, we evaluated the best model on both the training and test sets to derive the final performance metrics. The training code of hybrid model available on GitHub link - https://github.com/pateltanvi2992/EfficientSwin-A-Hybrid-Model-for-Blood-Cell-Classification-with-saliency-maps-visualization/tree/main.

B. Experiments and design

We conducted a set of examinations in our study to forge a novel hybrid model that merges the strengths of transformer models with the efficiency of CNN EfficientNet models. Our initial step involved fine-tuning base models from both domains to assess their standalone performance. We specifically focused on the EfficientNet family for our CNNs, drawing inspiration from recent studies [24]–[26] that highlight their superior performance over other CNN architectures.

The EfficientNet-B0 model, in particular, has garnered attention for its remarkable efficiency in learning from small datasets. This efficiency stems from its ability to represent learning effectively from limited data. Nonetheless, its performance is somewhat contingent on the specific characteristics of the dataset.

EfficientNet models are generally recognized for achieving high accuracy with a comparatively low count of parameters, which makes them exceptionally well-suited for small datasets where the risk of overfitting is high. The reduced parameter count in EfficientNet-B0 minimizes its propensity to overfit, thereby enhancing its ability to generalize from a constrained volume of training data. Despite its simplicity, the EfficientNet-B0 may not as effectively capture the complexities inherent in larger datasets characterized by a broader and more intricate distribution of data.

To make comparisons across different versions of EfficientNet, namely B0, B1, B2, and B3, we also conducted classification tests using models pre-trained on ImageNet. The outcomes of these tests, detailed in the results section, shed

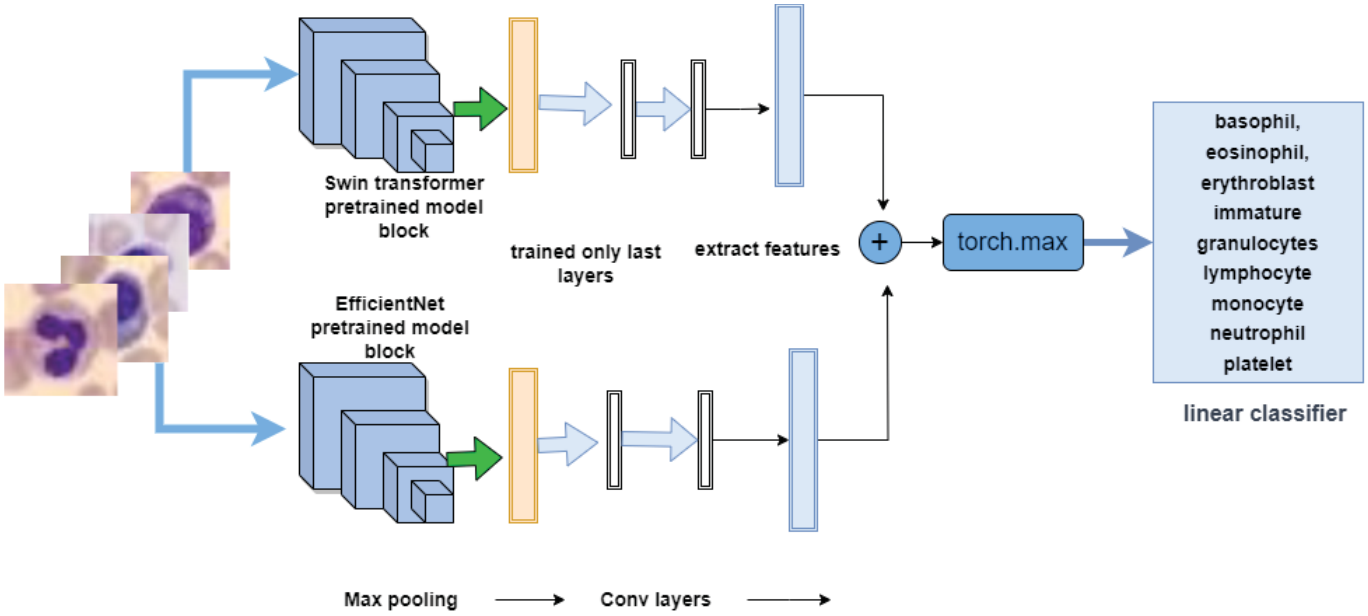


Fig. 2. Framework structure of the EfficientSwin hybrid model. Utilize the features of Swin Transformer and EfficientNet models to accurately capture the patterns, and combine features to connect the classification layer to classify blood cell types.

light on the performance variances among these EfficientNet variants.

Moreover, referencing the study by Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows" [27], which demonstrates the Swin Transformer's superiority over the Vision Transformer (ViT) across various image categorization benchmarks, we sought to delve deeper into the Swin Transformer's potential. Utilizing transfer learning, we trained the Swin Transformer on the same dataset to further ascertain its capabilities. This comprehensive approach allowed us to explore the effectiveness of combining transformer models with CNNs, particularly the EfficientNet models, in achieving enhanced model performance.

C. Methodology to compare the performance of the base model classifier with proposed model classifier

Following the training of the base models, we proceeded to evaluate their performance by comparing the confusion matrices. To achieve this, we employed McNemar's Test, which allowed us to construct a contingency table based on the outcomes represented in both matrices. The test is defined by the formula,

$$X^2 = \frac{(b-c)^2}{(b+c)}$$

Where b denotes the count of the instances classified as false positives in the first matrix and as true negatives in the second matrix, while c represents the count of false negatives in the first matrix and true positives in the second matrix. The resulting X^2 statistic follows a chi-square distribution with one degree of freedom, which can then be used to assess the statistical significance of the difference in performance between the two classifiers.

The primary goal of analyzing confusion matrices through this statistical method was to assess the performance of the proposed model. This was done by leveraging the chi-square distribution to calculate the p-value, which serves as a measure of the difference between the models' performances. Should the p-value fall below a predetermined significance level (for instance, 0.05), we would reject the null hypothesis, thereby concluding that there exists a statistically significant difference between the performance of the two models as evidenced by their respective confusion matrices. This approach enables a rigorous statistical comparison to determine the efficacy of the proposed model relative to the base models.

IV. Results

We conducted experiments across three distinct tasks. The initial task focused on the classification of various blood cell types—basophil, eosinophil, erythroblast, immune granulocytes, lymphocyte, monocyte, neutrophil, and platelet representing a multi-category classification challenge. During this task, we compared the performance of our hybrid model against that of the baseline model. The second task involved generating saliency maps with the classifier to visualize different classes of blood cells in microscopic images.

Furthermore, we undertook two comparative analyses. The first analysis compared the accuracy of our proposed hybrid model against findings reported in literature review articles, which included evaluations of classification performance across various blood cell datasets. The second analysis assessed the performance of our proposed model using the same dataset and identical data splitting criteria.

A. Hybrid model and base model evaluation

We initially fine-tuned the Swin Transformer model and evaluated its performance by calculating accuracy, precision, recall, and F1-score for each blood cell class, followed by computing the average for each metric. Similarly, we fine-tuned EfficientNet-B0 and calculated its average metrics, considering both as foundational models for our hybrid approach. Additionally, we have also apply fine-tunning process to EfficientNet-B1, B2, B3 to compare their performance outcomes. In Table I, we juxtaposed the average evaluation metrics of these models against those of our hybrid model. The comparison revealed that EfficientNet-B0 and B1 exhibited strong performance, with EfficientNet-B0 achieving an accuracy of 87.85% and EfficientNet-B1 slightly higher at 88.03%. The marginal difference of 0.18% between them, coupled with EfficientNet-B0's advantage of having fewer parameters—which translates into reduced training time—led us to select EfficientNet-B0 as the preferable base model for our hybrid configuration.

TABLE I. COMPARISON OF THE AVERAGE ACCURACY, PRECISION, RECALL AND F1-SCORE WITH THE BASE MODELS WITH THE PROPOSED EFFICIENTSWIN MODEL. ALL THE MODEL USED SAME TRAINING DATASET AND VALIDATION DATASET. THE HIGHEST RESULTS HIGHLIGHT WITH BOLD TEXT.

Model	Accuracy	Precision	Recall	F1-score
Swin Transformer	87.97%	86%	87%	86%
EfficientNet-B0	87.85%	87%	86%	86%
EfficientNet-B1	88.03%	87%	86%	86%
EfficientNet-B2	85.86%	84%	83%	84%
EfficientNet-B3	80.26%	78%	77%	77%
Hybrid approach	98.13%	98.07%	97.99%	98.02%

To further explore the model's performance in different class outcomes and its efficacy in categorization, we also created a confusion matrix on our proposed model performance (Fig. 3). In Fig. 4, shown the training accuracy, loss and validation accuracy over each epoch till epoch 30. The highest validation accuracy we received is 98.13%.

B. Comparative analysis of hybrid model vs. base models using McNemr's test

The evaluation of our proposed model against various baseline models using McNemar's test revealed significant findings in classification accuracy. The test showed a statistically significant improvement of our model over EfficientNet-B0, with a p-value of 0.0063. However, comparisons with EfficientNet-B1 and B2 did not yield statistically significant differences, with p-values of 0.125 and 0.625, respectively, indicating comparable performance levels. A notable exception was the comparison with EfficientNet-B3, where our model demonstrated a highly significant advantage, evidenced by a p-value of 0.00117. Furthermore, our model also showed a statistically significant improvement over the Swin Transformer model, with a p-value of 0.0390 showed in Table II. These outcomes underscore the superior performance and effectiveness of our proposed model against selected versions of EfficientNet and the Swin Transformer, affirming its potential in relevant applications.

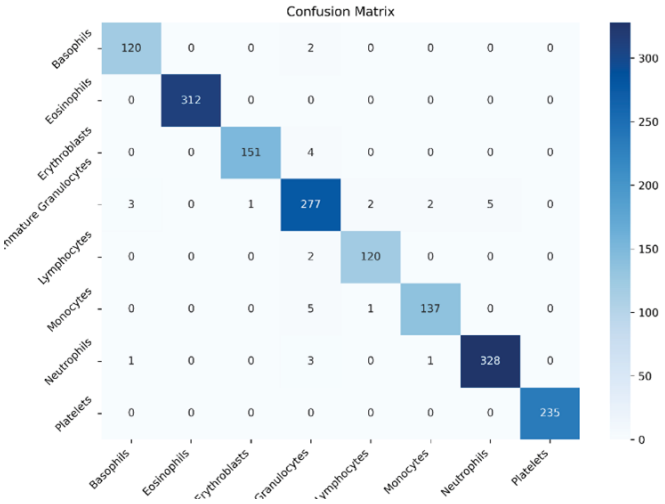


Fig. 3. Confusion matrix is showing the comprehensive assessment of classification performance of each class. 'platelet' and 'eosinophil' outperforms on other classes.

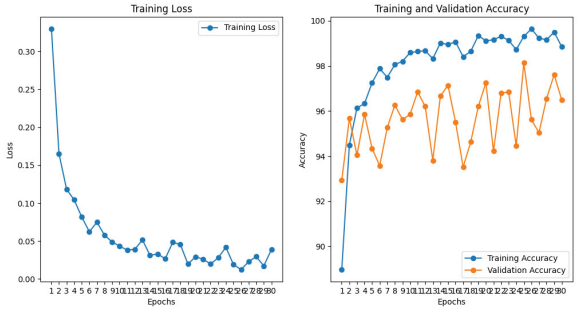


Fig. 4. Training Loss, Training Accuracy, and Validation Accuracy of the Model Over 30 Epochs. This figure illustrates the progression of training loss, alongside the improvements in training and validation accuracy with each successive epoch, highlighting the model's learning dynamics and convergence behavior up to epoch 30.

TABLE II. THE EVALUATION OF OUR PROPOSED MODEL AGAINST INDIVIDUAL BASELINE MODELS WAS CONDUCTED USING MCNEMAR'S TEST TO IDENTIFY STATISTICALLY SIGNIFICANT DIFFERENCES IN THE VALUES OF THE CONFUSION MATRICES ON A CLASS-BY-CLASS BASIS.

Comparison	McNemar's test P-value
Proposed model vs. EfficientNet-B0	0.0063
Proposed model vs. EfficientNet-B1	0.125
Proposed model vs. EfficientNet-B2	0.625
Proposed model vs. EfficientNet-B3	0.00117
Proposed model vs. Swin Transformer	0.0390

C. Saliency maps

Saliency maps visually highlight the areas within an image that are most influential in the model's decision-making process, which also known as Region of Interest (ROI). Which is

showing in Fig. 5 and 6. Through the highlight ROI in various labels.

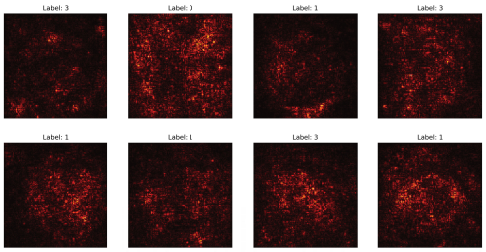


Fig. 5. Saliency maps overlaid with labels

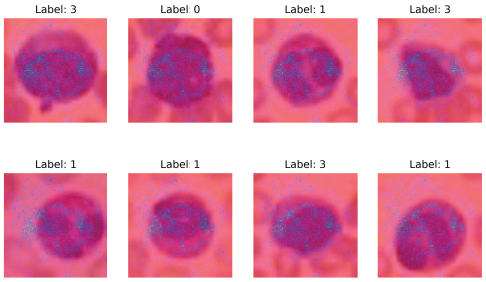


Fig. 6. The overlapping saliency maps on images provide insights into the regions of interest crucial for accurate classification

D. Comparative evaluation of proposed method accuracy vs. previous studies accuracy

V. Discussion

In the literature review, we examined previous research conducted on blood cell dataset classification using diverse methodologies. In Table IV, we compared the findings of these studies, including their respective methods and achieved accuracies. Subsequently, we contrasted these results with those obtained from our proposed hybrid model, which yielded the highest accuracy of 98.13%. However, it's worth noting that the studies in Table IV utilized varying train-validation splits and datasets. In Table III, we conducted a comparison using the same blood cell dataset and splitting methodology as previous research, achieving the highest accuracy even when compared to the proposed method by Yang, Jiancheng, et al. [20]. Moving backward to Table I, we compared the results of the base models with those of our proposed model. To ensure a robust comparison, we employed McNemar's test, revealing a significant difference between the confusion matrices of EfficientNet-B0 and EfficientSwin, with a p-value of 0.0063. Additionally, comparisons with other base models yielded p-values of 0.125, 0.625, and 0.00117 for EfficientNet-B1, EfficientNet-B12, and EfficientNet-B3, respectively, while the comparison with the Swin Transformer model resulted in a p-value of 0.0390. Furthermore, we calculated the specificity of each model, with values of 0.9823, 0.9827, 0.9795, 0.9713, 0.9828, and 0.9973 for EfficientNet-B0, EfficientNet-B1, EfficientNet-B2, EfficientNet-B3, Swin Transformer, and the proposed model, respectively. Notably, our proposed model achieved 100% accuracy for the 'platelet' and 'eosinophil' classes on the validation dataset.

TABLE III. Comparison of the accuracy of different deep learning CNN architecture which was listed in the literature review section and used the same BloodMNIST dataset. The best accuracy values are in bold.

Reference	Methods	Resolution	Accuracy
Yang, Jiancheng, et al. [20]	ResNet-18 [21]	28	95.8%
Yang, Jiancheng, et al. [20]	ResNet-18 [21]	224	96.3%
Yang, Jiancheng, et al. [20]	ResNet-50 [21]	28	95.6%
Yang, Jiancheng, et al. [20]	ResNet-50 [21]	224	95%
Yang, Jiancheng, et al. [20]	Auto-sklearn [22]	224	87.8%
Yang, Jiancheng, et al. [20]	AutoKeras [23]	224	96.1%
Yang, Jiancheng, et al. [20]	Google AutoML Vision [20]	224	96.6%
Proposed hybrid model	**Hybrid approach**	**224**	**98.13%**

The classification performance of the model across different cell types varied, with F1-scores ranging from 95% to 100%. Despite high F1-scores for most classes, the model exhibited challenges in accurately categorizing Immature Granulocytes, mislabeling a small number as Monocytes or Erythroblasts. Notably, the model demonstrated robust performance in distinguishing Platelets and Monocytes, achieving accurate classifications for 328 Monocytes and 235 Platelets. In Fig. 5, 6 we have used saliency maps as visualization tools to understand model decision by highlighting the most relevant regions of input data. We have found out that each class has a different pattern.

Introduced by Vaswani et al. [29] in their seminal work "Attention is All You Need," transformers have since been applied across various fields, showcasing their adaptability and effectiveness in tackling complex classification tasks [30], [31]. Meanwhile, the merits of EfficientNet models on smaller datasets should not be overlooked, underscoring the distinct advantages of various architectures based on dataset size. In our initial experiments, we trained baseline models and observed that EfficientNet-B0 outperformed its successors (B1, B2, B3) in terms of accuracy. Subsequently, training the Swin Transformer on the same dataset and split yielded accuracy comparable to EfficientNet-B0. However, by integrating these

TABLE IV. EVALUATION COMPARISON RESULTS FOR BLOOD CELL CLASSIFICATION VIA THE PROPOSED METHOD AGAINST THE LATEST DEEP LEARNING WORKS IN THE LITERATURE REVIEW SECTION. THE HIGHEST RESULTS HIGHLIGHT THROUGH THE BOLD TEXT.

Reference	**Data**	**Methods**	**Accuracy**
Jung et. al. [14]	Data privacy and generated by GAN	W-Net and transfer learning	97%
Ma, Li, et al. [19]	BCCD	DCGAN and Transfer learning	91.7%
Acevedo et. al. [15]	Private dataset	CNN, VGG16 and Inceptionv3	96.2%
Qin et. al. [16]	Private dataset	ResNet, SVM	76.84%
Sharme et. al. [17]	BloodMNIST	"VGGs", "Inception V3", and "Xception"	87%
Zhao et al. [18]	Cell vision, ALL-IDB, Jiashan	CNN, SVM, and random forest	92.8%
Şengür, Abdulkadir, et al. [28]	White Blood Cells (WBCs)	Image processing (IP) and machine learning (ML)	85.7%
Proposed hybrid model	**BloodMNIST**	**Hybrid Swin transformer and EfficientNet**	**98.13%**

models into our proposed architecture, we achieved significantly superior results.

This study also has certain limitations. The comparison with baseline models primarily focuses on quantitative aspects, overlooking qualitative factors such as interpretability and computational efficiency. Furthermore, it does not delve into hyperparameters or alternative model architectures beyond the hybrid approach.

VI. CONCLUSION

The hybrid model, which combines features from the EfficientNet and Swin Transformer designs, performs better than the baseline models alone, according to the evaluation results. In particular, the hybrid model outperforms the baseline models in certain classes in terms of accuracy, precision, recall, and F1-score. This shows that a more robust and efficient model is produced for the job at hand by utilizing the qualities of both designs. Additionally, the comparison of previous study and their results also support the higher accuracy of this approach. Furthermore, statistical tests like McNemar's Test could be used to confirm the importance of these enhancements and offer more in-depth understanding of the models' comparative performances. Through the saliency maps we shows the different labels Region of Interest (ROI) as well.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my professors for their unwavering support and guidance.

REFERENCES

[1] A. Ojaghi, G. Carrazana, C. Caruso, A. Abbas, D. R. Myers, W. A. Lam, and F. E. Robles, "Label-free hematology analysis using deep-ultraviolet microscopy," *Proceedings of the National Academy of Sciences*, vol. 117, no. 26, pp. 14 779–14 789, 2020.

[2] R. B. Hegde, K. Prasad, H. Hebbar, and B. M. K. Singh, "Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 382–392, 2019.

[3] L.-T. Huang, C.-P. Zhang, Y.-B. Wang, and J.-H. Wang, "Association of peripheral blood cell profile with alzheimer's disease: a meta-analysis," *Frontiers in Aging Neuroscience*, vol. 14, p. 888946, 2022.

[4] K. A. K. Al-Dulaimi, J. Banks, V. Chandran, I. Tomeo-Reyes, and K. Nguyen Thanh, "Classification of white blood cell types from microscope images: Techniques and challenges," *Microscopy science: Last approaches on educational programs and applied research (Microscopy Book Series, 8)*, pp. 17–25, 2018.

[5] Z. F. Mohammed and A. A. Abdulla, "An efficient cad system for all cell identification from microscopic blood images," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 6355–6368, 2021.

[6] Y. Y. Baydilli and Ü. Atila, "Classification of white blood cells using capsule networks," *Computerized Medical Imaging and Graphics*, vol. 80, p. 101699, 2020.

[7] F. Ajesh, F. Philip, P. Sajan, and R. Rahim, "Cad systems for automatic detection and classification of covid19 using image processing and machine learning," in *Proceedings of the 2nd Biennial International Conference on Safe Community, B-ICSC 2022, 20-21 September 2022, Bandar Lampung, Lampung, Indonesia*, 2023.

[8] J. L. Diaz Resendiz, V. Ponomaryov, R. Reyes Reyes, and S. Sadovnychiy, "Explainable cad system for classification of acute lymphoblastic leukemia based on a robust white blood cell segmentation," *Cancers*, vol. 15, no. 13, p. 3376, 2023.

[9] T. Tran, O.-H. Kwon, K.-R. Kwon, S.-H. Lee, and K.-W. Kang, "Blood cell images segmentation using deep learning semantic segmentation," in *2018 IEEE international conference on electronics and communication engineering (ICECE)*. IEEE, 2018, pp. 13–16.

[10] K. AL-DULAIMI and T. Makki, "Blood cell microscopic image classification in computer aided diagnosis using machine learning: A review," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 43–55, 2023.

[11] K. Yu, X. Yang, S. Jeon, and Q. Dou, "An end-to-end medical image fusion network based on swin-transformer," *Microprocessors and Microsystems*, vol. 98, p. 104781, 2023.

[12] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From cnn to transformer: A review of medical image segmentation models," *arXiv preprint arXiv:2308.05305*, 2023.

[13] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, "A dataset of microscopic peripheral blood cell images for development of automatic recognition systems," *Data in brief*, vol. 30, 2020.

[14] C. Jung, M. Abuhamad, D. Mohaisen, K. Han, and D. Nyang, "Wbc image classification and generative models based on convolutional neural network," *BMC Medical Imaging*, vol. 22, no. 1, pp. 1–16, 2022.

[15] A. Acevedo, S. Alférez, A. Merino, L. Puigví, and J. Rodellar, "Recognition of peripheral blood cell images using convolutional neural networks," *Computer methods and programs in biomedicine*, vol. 180, p. 105020, 2019.

[16] F. Qin, N. Gao, Y. Peng, Z. Wu, S. Shen, and A. Grudtsin, "Fine-grained leukocyte classification with deep residual learning for microscopic images," *Computer methods and programs in biomedicine*, vol. 162, pp. 243–252, 2018.

[17] M. Sharma, A. Bhave, and R. R. Janghel, "White blood cell classification using convolutional neural network," in *Soft Computing and Signal Processing: Proceedings of ICSCSP 2018, Volume 1*. Springer, 2019, pp. 135–143.

[18] J. Zhao, M. Zhang, Z. Zhou, J. Chu, and F. Cao, "Automatic detection and classification of leukocytes using convolutional neural networks," *Medical & biological engineering & computing*, vol. 55, pp. 1287–1301, 2017.

[19] L. Ma, R. Shuai, X. Ran, W. Liu, and C. Ye, "Combining dc-gan with resnet for blood cell image classification," *Medical & biological engineering & computing*, vol. 58, pp. 1251–1264, 2020.

[20] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: Hands-free automl via meta-learning," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 11 936–11 996, 2022.

[23] H. Jin, Q. Song, and X. Hu, "Auto-keras: An efficient neural architecture search system," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 1946–1956.

[24] V.-T. Hoang and K.-H. Jo, "Practical analysis on architecture of efficientnet," in *2021 14th International Conference on Human System Interaction (HSI)*. IEEE, 2021, pp. 1–4.

[25] A. Abdelrahman and S. Viriri, "Efficientnet family u-net models for deep learning semantic segmentation of kidney tumors on ct images," *Frontiers in Computer Science*, vol. 5, p. 1235622, 2023.

[26] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-label classification of fundus images with efficientnet," *IEEE access*, vol. 8, pp. 212 499–212 508, 2020.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[28] A. Şengür, Y. Akbulut, Ü. Budak, and Z. Cömert, "White blood cell classification based on shape and deep features," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*. Ieee, 2019, pp. 1–4.

[29] V. Ashish, "Attention is all you need," *arXiv preprint arXiv: 1706.03762*, 2017.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.