Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

A Novel CNN Model with Entropy-Coded Genetic Algorithm for Blood Cell Classification

Abdul Muiz Fayyaz[1,*], Said Jadid Abdulkadir[1,2], Safwan Mahmood Al-Selwi[1,2], Ebrahim Hamid Sumiea[1], Saba Iqbal[3], Shahab Ul Hassan[1,4]

[1] Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia
[2] Center for Research in Data Science (CeRDaS), Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia
[3] Department of Computer Science, University of Wah, Wah Cantt 47040, Pakistan
[4] Centre for Intelligent Signal & Imaging Research (CISIR), Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia

ARTICLE INFO

Keywords:
White blood cells; CNN; entropy; GA; SVM

ABSTRACT

White blood cells are the immune system components that combat infections. White blood cells or WBCs, are blood cells that are present in the bone marrow are responsible for protecting against pathogens that kill healthy cells. Finding the immature cell formations early on will help to lessen the severity of this problem and eventually reduce the patients' rate of modality. In this study, an entirely new deep Convolutional Neural Network (CNN) model is presented inside this paradigm. The deep CNN model is pre-trained on medical imaging datasets to improve its performance. In preprocessing, K-Mean Clustering is utilized to highlight the region of interest (ROI). Post clustering, the updated dataset is used for feature extraction using a novel proposed CNN model with 33 layers. Moreover, the Entropy-Coded Genetic Algorithm is another significant contribution which is used for feature selection to choose the most optimal features. These selected features are subsequently classified using a Support Vector Machine (SVM). The results show that with 2025 features, the proposed model achieved 97.59% accuracy, 96.0% sensitivity and 97.91% specificity using the cubic-SVM classifier.

1. Introduction

Blood is the most important element of the human body, consisting primarily of liquid named plasma (55% of total volume) to allow blood to easily flow through a blood vessel [1]. The biological components of blood are classified into three main cells based on their colour, shape, size, texture and composition. RBCs are the red blood cells (Erythrocytes) [2], WBCs (Leukocytes) [3] and platelets [4]. Normal RBC counts in the blood vary from 4 to 6 million per microliter, with RBCs accounting for 40 to 45 percent of total blood volume [5,6], WBCs combat germs and protect humans from harmful illnesses, which is why they are sometimes referred to as "Defender cells". The number of WBCs per microliter of blood has fluctuated from 4,500 to 11,000, depending on the individual [7,8], since white

* *Corresponding author.*
E-mail address: abdulmuizfayyaz@gmail.com

https://doi.org/10.37934/araset.63.3.134148

blood cells (WBCs) account for just 1% of total blood volume, even a small alteration can make a significant effect since the human immune system is dependent on WBCs [9,10]. WBCs are significantly larger than other blood cells due to the presence of nuclei and cytoplasm and because of this distinguishing characteristic, WBCs are divided into two categories: granulocytes and agranulocytes. Compared to agranulocytes, which include bulky cytoplasm in their nucleus due to the lack of distinct granules, granulocytes have a nucleus with a 2-5 lobed shape and granules in the cytoplasm. In addition to these two groups, there are five major subcategories of immune cells: basophils, eosinophils, neutrophils, monocytes and lymphocytes. Basophils are the most common kind of immune cell [11].

Neutrophils are the most prevalent type of blood leukocyte; they account for 55-70 percent of all white blood cells and are capable of fighting fungal and bacterial infections; their nucleus is located close to the C, S and U, resulting in a multilobed appearance [12]. Eosinophils are produced in the bone marrow and are therefore primarily used in the immune system; they account for 2-4 percent of total white blood cell volume and fight in response to allergies, collagen disease, parasitic infections and infections of the central nervous system; their nuclei are shaped like two or three lobes [13]. Basophils account for 0-1 percent of total white blood cell volume and are primarily responsible for allergy and antigen responses as well as improving blood vessel dilatation. Their nuclei are bi or trilobed in form and they are responsible for allergic and antigen responses [14]. Lymphocytes are also the most important component of the immune system, accounting for 30% of its total volume. They are located in the spleen, lymph ducts and lymph nodes, among other places. Its nucleus is spherical and appears to be made of a single piece [15]. Approximately 5-8 percent of white blood cell volume is made up of monocytes, which generate macrophage tissues that travel through the circulation to eliminate dead cells, combat pathogens and release inflammatory mediators. The nucleus of the monocyte has an uneven structure inside the cytoplasm [16]. In this paper, the BCCD (Blood Cell Count and Detection) Dataset [17] is used for blood cell classification that is publicly available as Shown in Figure 1.

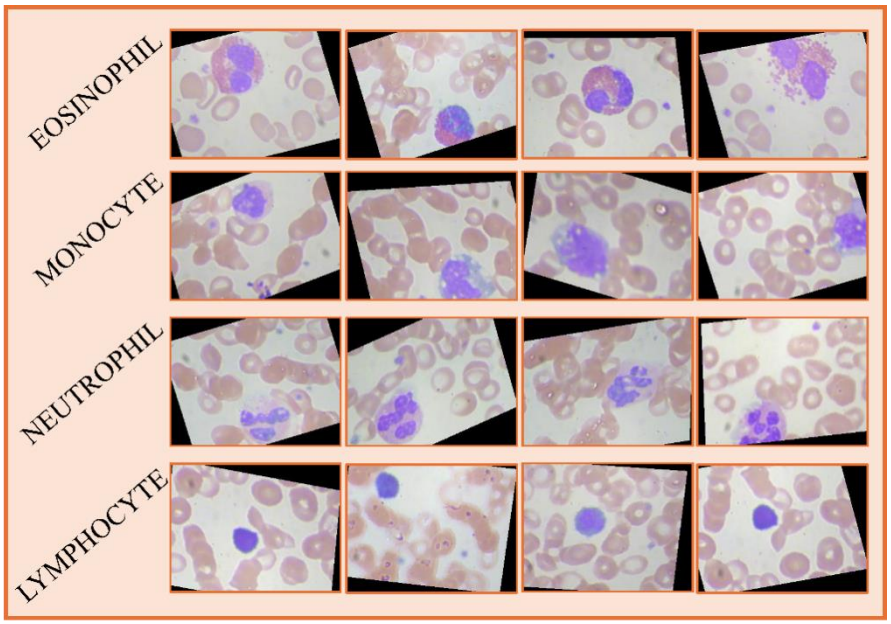

Fig. 1. The blood cell classification dataset

The significant contributions are as follows:

i. In this framework, a new deep CNN model is proposed and pre-trained using a medical imaging dataset.
ii. K-Mean Clustering is performed as pre-processing for Blood Cell Classification. The newly proposed deep CNN is used for feature extraction.
iii. The Entropy-Coded GA is used for features selection for selecting the core features. These selected features are given to the SVM Classifier for classification.

The rest of this paper is organized as follows: Section 2 highlights the related works. Section 3 discusses the material and methods used to conduct the experiments. Section 4 shows the achieved results. Section 5 provides a discussion of the results. Finally, this paper is concluded in Section 5.

2. Related Works

In the following section, we provide a comprehensive review of literature pertinent to our study. Through this exploration, we aim to elucidate the current landscape of research surrounding blood cell classification methodologies, particularly focusing on approaches involving convolutional neural networks (CNNs) augmented with Entropy and Genetic Algorithm (GA). By synthesizing existing works, we aim to underscore the innovative contributions and distinctiveness of our proposed model within this domain.

In the realm of white blood cell subtype classification, traditional methods rely heavily on feature design, while machine learning (ML) approaches demand extensive labelled data. The study by Song *et al.,* [18] presents a semi-supervised CNN, offering high accuracy with minimal labelling. This pioneering work lays the foundation for advancing diagnostic solutions in blood disease detection. On the other hand, Rahat *et al.,* [19] evaluate five deep learning models for automating blood cell classification in haematology, finding VGG19 as the top performer with positive accuracy. This highlights the potential of deep learning in enhancing medical diagnostics, particularly in haematology, with the practical implementation of VGG19 recommended. However, further investigation is needed for performance variations and generalization to unseen data. Moreover, the study by Liang *et al.,* [20] introduces an image-based RBC deformability assessment (IRIS), which sensitively detects differences in 0.001% glutaraldehyde-treated (Red Blood Cells) RBCs from controls.

Eshel *et al.,* [21] present a novel approach using infrared spectroscopy and ML to diagnose bacterial infections and monitor antibiotic therapy in febrile paediatric oncology patients with bacteraemia. The logistic regression classifier achieved over 95% success in distinguishing bacterial samples within an hour, while infrared spectroscopy coupled with ML demonstrated an 87.5% accuracy in assessing antibiotic treatment effectiveness. Also, ML models were applied by Pullakhandam *et al.,* [22] to classify Iron Deficiency Anemia (IDA) from CBC data, achieving a PR AUC of 0.87 and high recall/sensitivity across datasets. Feature importance analysis [23,24] highlighted critical factors such as low haemoglobin levels and higher age. The approach demonstrates the potential for enhancing current automated CBC analysers, supported by its high performance and consistency. In contrast, Gu *et al.,* [25] propose an improved YOLOv5 model, termed AYOLOv5, utilizing an attention mechanism to enhance cell detection in dense and complex distributions. Integrating convolutional block attention modules (CBAM) and transformer encoder blocks, AYOLOv5 enhances feature extraction in dense regions and improves network capacity for cell property recognition. Evaluation on the BCCD dataset yields a mean Average Precision (mAP) of

93.3%, outperforming prior methods, with validation set recognition accuracy rising from 89% to 98%. AYOLOv5 effectively extracts cell feature information, significantly enhancing cell picture quality and recognition performance.

Moreover, Zeng *et al.,* [26] propose a portable AI-based leukocyte detection system for rapid and cost-effective WBC analysis. It simplifies operations, integrates functions and achieves reliable results with less than 1% error rate in WBC counting and about 2% in classification, with a detection time of under 20 seconds. This system promises to standardize disease diagnosis in point-of-care testing (POCT) with its portability and scalability. To improve malaria parasite detection in microscopic images, a deep-CNN model combined with Random Forest is proposed by Murmu *et al.,* [27] which integrates domain-specific expertise, incorporates Global Average-Pooling layer (GAP) for enhanced parasite area visualization and employs Canny edge detection for precise boundary detection. Evaluation of diverse datasets showcases superior performance compared to existing models. Uzen *et al.,* [28] propose a novel approach for WBC classification, leveraging ConvMixer and Swin transformer models in a hybrid network termed SC-MP-Mixer. This model effectively extracts spatial details using ConvMixer and applies self-attention mechanisms via a Swin transformer, with a multipath structure enhancing patch representations.

Our literature review urges the study of methodologies of classifying human blood cells using deep CNN with different techniques such as entropy, GAs and attention mechanisms. These diverse methods, established to overcome the problem of blood cell classification and identification of the disease, have subsequently developed a spectrum of innovative approaches targeting different elements of blood cell analysis. Our proposed model combines various methodologies to improve the efficiency and applicability of blood cell classification systems.

3. Material and Method

In this proposed method, the K-Mean Clustering is used as preprocessing of the blood cells dataset. The new 33-layer deep CNN is proposed to extract the deep features. Entropy-Coded GA is used for feature selection. These selected features are given to SVM for the classification of Blood Cells. The detailed flow diagram of the proposed method is shown in Figure 2.

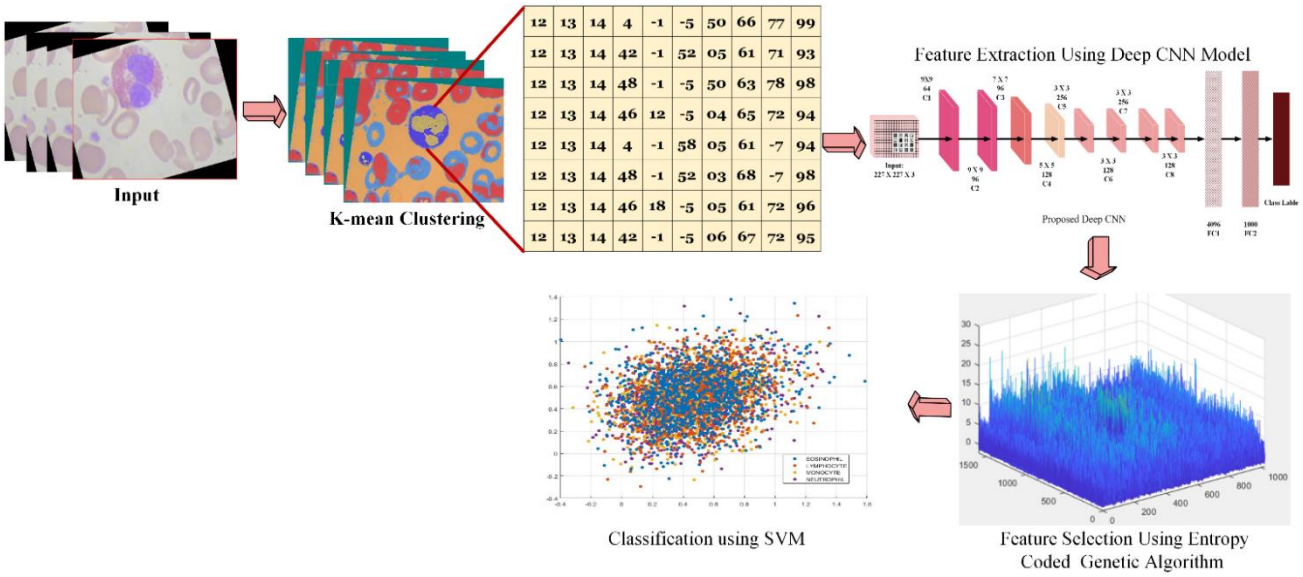

Fig. 2. The detail about proposed methodology

3.1 Preprocessing

In this preprocessing, K-Mean Clustering [29] is used to enhance the image as shown in Figure 3. Since there are fewer data points in K-Mean than there are clusters, we place each data point on the graph at the location of the cluster's centroid. Every centroid will be given a cluster number. If the number of data points exceeds the number of clusters, we find the smallest distance feasible by calculating the separation between each data point and each centroid. The cluster that this data sample belongs to is said to be the one that is closest to it.

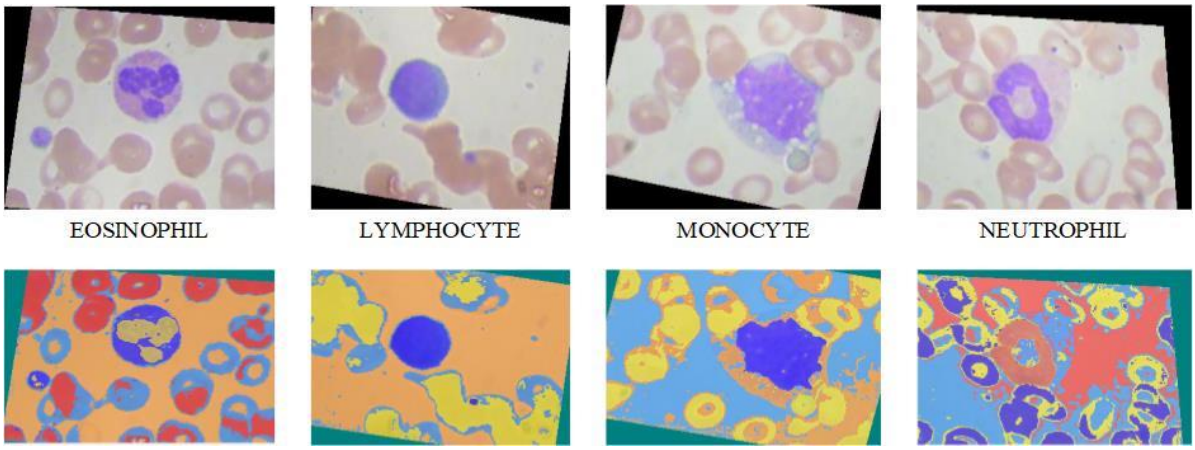

Fig. 3. Preprocessing using K-Mean clustering

.

3.2 Proposed Deep CNN

The proposed Deep CNN has thirty-three total layers, eight of which are convolutional layers and three of which are fully connected. In this architecture, Rectified Linear Units (ReLU) are used as an alternative to the tanh function. It was standard at this time [30]. The Proposed Deep CNN used the dropout layer for reducing the overfitting in fully connected layers. Two-dimensional convolution layers utilized trainable filters, frequently through optional trainable bias in each kernel [31]. Another major component of the CNN is a pooling layer. Maximum pooling is a pooling method wherein each of the patches on the map determines the maximum or the largest value. The Proposed Deep CNN model contains the following architecture shown in Figure 4 and Table 1.

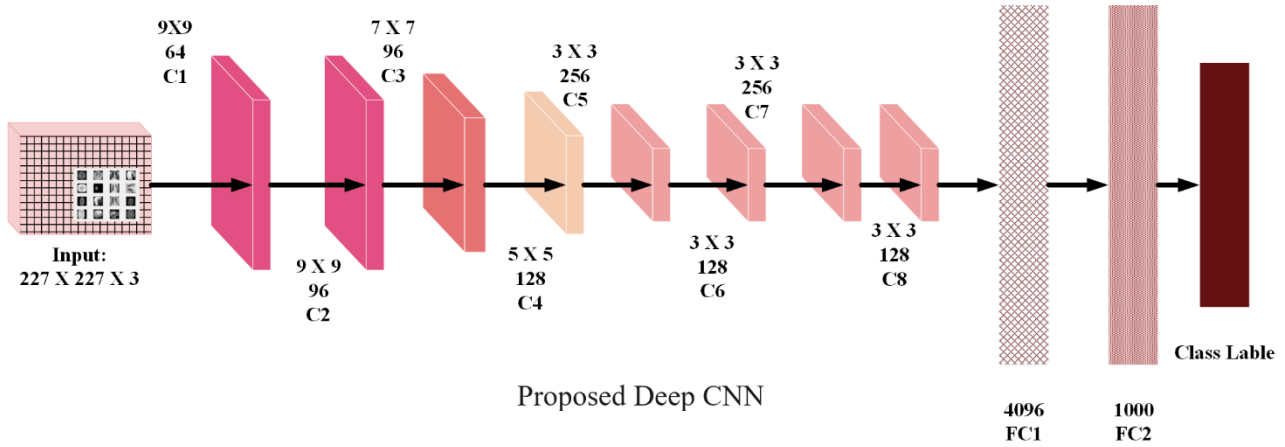

Fig. 4. The architectures of deep CNN

In Deep CNN the model architecture is based on eight convolutions layers and 3 fully connected layers. The input size of Deep CNN is “227 X 227 X 3” and the layer name is Input. Convolution layer 1 filters the image with 64 kernels-based size “9 X 9 X 3” and in convolution layer 2 filters the input of convolution layer 1 with 96 kernels-based size “9 X 9 X 3”. The input of convolution layer 2 is filtered by the convolution layer 3 with the 96 kernels-based size of “7 X 7 X 3”. Moreover, the outcomes from convolution layer 3 with 128 kernels-based size “5 X 5 X 3”, in the same way, the last convolution layer based on 128 kernel sizes based on 3 X 3 X 3. Deep CNN also has 3 features layer known as “Fully Connected”, where FC1 has “4096” features, FC2 consists of 1000 features and the FC layer contains the class labels. The detailed architecture is described in Table 1.

Table 1
Detail architecture of deep CNN model

Deep CNN		
Layers	Name	Properties
1	Input	227 X 227 X 3
2	C1	9 X 9, 64, BiasLearnRateFactor = 2, Stride = 4
3	R1	Relu
4	N1	5, K = 1
5	P1	3 X 3, Stride = 2
6	C2	9 X 9,96, BiasLearnRateFactor = 2, Padding = 4
7	R2	Relu
8	N2	5, K = 1
9	C3	7 X 7, 96, BiasLearnRateFactor = 2, Padding = 3
10	R3	Relu
11	N3	5, K = 1
12	P2	3 X 3, Stride = 2
13	C4	5 X 5, 128, BiasLearnRateFactor = 2, Padding = 2
14	R4	Relu
15	C5	3 X 3, 256, BiasLearnRateFactor = 2, Stride = 4, Padding = 1
16	R5	Relu
17	N4	5, K = 1
18	C6	3 X 3, 128,2, BiasLearnRateFactor = 2, Padding = 1
19	R6	Relu
20	C7	3 X 3, 256, BiasLearnRateFactor = 2, Stride = 4
21	R7	Relu
22	N5	5, K = 1
23	C8	3 X 3, 128, 2, BiasLearnRateFactor = 2, Padding = 1
24	R8	Relu
25	FC1	Features = 4096, BiasLearnRateFactor = 2
26	R9	Relu
27	D1	D = 0.5
28	FC2	Features = 1000, BiasLearnRateFactor = 2
29	R10	Relu
30	D2	D = 0.5
31	FC	Class label = 18, BiasLearnRateFactor = 2
32	PROB	Softmax
33	OUTPUT	Classification

Initially, the proposed CNN model is pre-trained using the Medical Imaging Dataset. The Medical Imaging Dataset has 18 categories. Each category has 1400 training and 100 testing images. This research uses 1500 images per category for training and assessment. The Fully Connected layer offers dataset image properties (a total of 27000 images). After pretraining, the proposed deep CNN is used for feature extraction.

3.3 Entropy Coded Genetic Algorithm

Entropy [32], when applied to feature selection, assesses the unpredictability of data and the effectiveness of a feature in grouping data. High entropy implies that data is amalgamated and harder to separate, while low entropy suggests that a feature provides informative characteristics that lead to more precise classification [33]. Consider $q_1, q_2, \dots, q_N$ denote the number of features gathered by the deep network. The subsequent Eq. (4) computes the entropy E_P value for each feature [34]:

$$E_P(q_1, q_2, \dots, q) = -\sum_{f_1} \dots \sum_{f_N} \theta(y_1, \dots, y_N) \log \theta(y_1, \dots, y_N) \quad (1)$$

Here, $y_1, \dots, y_N$ represents the values of the core random variable and $\theta(y, \dots, y_N)$ indicates their probabilities.

GA is a stochastic optimization technique inspired by the process of natural selection and genetics [35]. It operates by evolving a population of potential solutions over multiple generations to find an optimal solution [36]. In feature selection, GA mimics the process of natural selection by iteratively selecting the most promising features from one generation to the next, combining and mutating them to explore the solution space efficiently. GA has demonstrated effectiveness in various optimization tasks and problem domains due to its ability to handle complex search spaces and find near-optimal solutions [37].

In this paper, we propose an approach that combines Entropy ranking with a GA for feature selection. First, features are ranked based on their Entropy values, indicating their informativeness. Then, the GA is employed to iteratively select and refine a subset of features that maximizes classification accuracy or minimizes error. The feature selection process based on the GA is outlined in the following Algorithm 1 [37]:

Algorithm 1
Entropy-Coded Genetic Algorithm for feature selection

Given	$feat$	feature vector (instances x features)
	$label$	label vector (instance x 1)
	N	Number of chromosomes
	max_{Iter}	Maximum number of generations
	CR	Crossover rate
	MR	Mutation rates
	HO	Hold-Out partition object for cross-validation
Output	s_{feat}	Selected features (instances x features)
	SF	Selected feature index
	NF	Number of selected features
	F_r	Ranked features

Process # Initialization
1. Load the benchmark dataset and set aside a validation set.
2. Set algorithm parameters N, max_{Iter}, CR and MR.
3. Initialize solution matrix *X* with binary values (0 or 1) randomly.
4. Initialize global best fitness $f_{it}G = \infty$ and fitness vector f_{it} with zeros.

Iterations

for $feat = 1$ **to** N **do**
  Calculate Entropy $F_1..F_N$
  Calculate Information gain $F_1..F_N$
  Feature Ranking (Image, number of features)
  F_r = Key Sort(Λ, $X_1 \cdot X_1[1]$, reverse = True); *Ranked Features* = F_r
  $feat = F_r$
  for $t = 1$ **to** max_{Iter} **do**
    Selection based on fitness proportionate
    Calculate inverted fitness $Ifit = 1 - fit$ for selection
    Calculate selection probability $prob = Ifit/sum(Ifit)$
    Crossover
    foreach *pair i in X* **do**
      Perform crossover with probability CR
      Update child solutions $X1$, $X2$
    end
    Mutation
    foreach *solution i in X* **do**
      Perform mutation with probability MR
    end
    Fitness Evaluation and Update
    foreach *solution i in X* **do**
      Calculate fitness $fit(i)$ using $fun(feat, label, X(i, :) > thres, HO)$
      Update global best solution Xgb and fitness $fitG$ if current solution is better
    end
  end
end

Return
Selected features SF where X_{gb} > *thres*
Extracted selected features matrix s_{feat} from $feat$
Number of selected features NF

3.4 Images Classification

These selected features are passed to SVM for classification. In this study, the different variants of Support Vector Machine (SVM) are used for classification, which included the linear SVM (L-SVM), the quadratic SVM (Q-SVM), the fine Gaussian (FG-SVM), the medium Gaussian (MG-SVM) and the coarse Gaussian (CG-SVM), as well as the cubic SVM (C-SVM) [38,39].

3.5 Performance Evaluation

Five metrics were used to evaluate the generated model's performance: accuracy, sensitivity, specificity, training time and prediction speed. The following is the explanation of certain performance measures [40,41]:

i. Accuracy: is related to the proximity of a measure to its recognized values and is defined as "the amount to which the outcome of a measurement correlates to an appropriate value or norm."

$$\text{Acc} = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

ii. Sensitivity: also referred to as recall, is a test's capacity to accurately identify patients who have an illness. Sensitivity is the measure of how well a test yields a positive result for

those who have tested positive for the illness (sometimes referred to as the "true positive" rate).

$$\text{Sen} = \frac{TP}{TP+ FN} \tag{3}$$

iii. Specificity: also referred to as the "true negative" rate, is the ability of a test to consistently generate a negative result for individuals who do not have the health problem under assessment.

$$\text{Spe} = \frac{TN}{TN+ FP} \tag{4}$$

iv. Training time: it refers to the time duration required to train the model on a dataset until it achieves the desired performance.

$$T_{time} = E \times \frac{N}{B} \times Ti \tag{5}$$

Where E is the number of epochs, N is the number of training samples, B is the batch size and Ti is the average time per iteration.

v. Prediction Speed: it is the amount of time it takes a trained model to make predictions based on a new set of data. This performance parameter is crucial, especially for high-throughput or real-time applications where quick predictions are necessary.

$$T_{prediction} = \frac{T_{total\_inference}}{N_{samples}} \tag{6}$$

Whereas $T_{total\_inference}$ is the total time taken to make predictions on a batch of $N_{samples}$.

4. Results

The results of the tests and trials that were conducted to demonstrate a rigorous approach are detailed in this section. Both of these tests were performed on a Lenovo T440s personal computer equipped with an i7 processor, 8 GB of RAM and a 500 GB SSD. For modifying performance, the number of trials with a distinct collection of characteristics is carried out in a controlled environment. This paper demonstrates the efficacy of the proposed contribution/technique through the results presented. In Computer Vision and Machine Learning, there are many performance measures which are used for checking how the proposed work is robust and efficient. Some of these measures are Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), training time and prediction speed.

The proposed Deep CNN is pre-trained using a medical imaging dataset. This medical imaging dataset contains multiple different medical datasets known as the Brain CT-Scan Dataset [42], kvasir Dataset [43], Tumorslices Dataset, Tuberculosis Dataset [44], Pneumonia [45] and COVID-19 radiology dataset. This medical imaging dataset contains 27000 images having 3 channels. The proposed WebCNNNet deep model was first pretrained using the above medical imaging dataset. The dataset for training and validation in medical imaging is shown below in Figure 5.

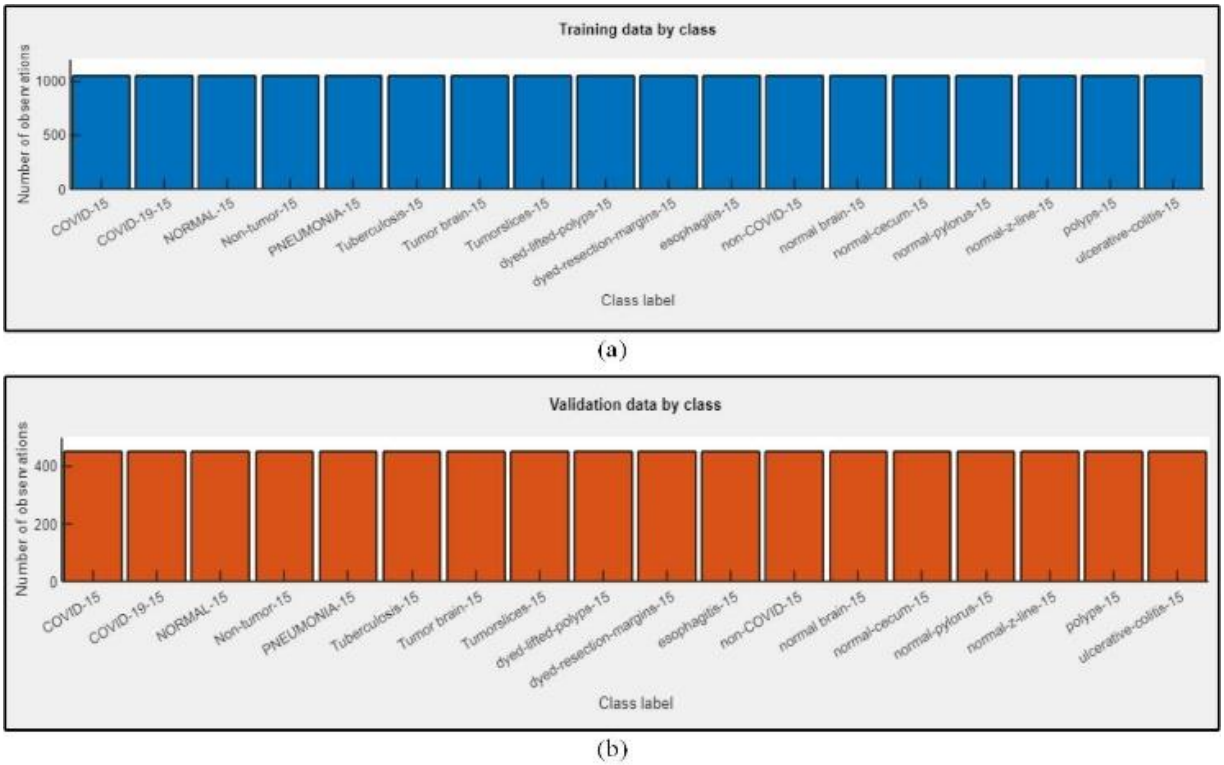

Fig. 5. The medical imaging dataset for training and validation

In the experiment setup, 250, 550, 750 and 2025 features are selected using Entropy-coded GA. The 5-fold cross-validation is applied which splits the data into 50 – 50 random training and testing sets. Then the labels are generated from this feature matrix by using multiple variants of SVM classifiers. The results of the experiments are shown in Table 2, Table 3, Table 4 and Table 5.

Table 2
The results of experiment setup 250 Features

Classifier	ACC	SEN	SPE
L-SVM	79.0%	77.28%	79.99%
Q-SVM	89.0%	89.26%	88.26%
C-SVM	91.1%	90.99%	90.0 %
MG-SVM	89.6%	81.05%	82.56%
CG-SVM	74.2%	78.53%	79.65%

Table 3
The results of experiment setup 550 features

Classifier	ACC	SEN	SPE
L-SVM	83.0%	83.28%	82.99%
Q-SVM	91.7%	90.26%	91.76%
C-SVM	93.2%	93.96%	92.99 %
MG-SVM	91.6%	90.01%	90.26%
CG-SVM	75.7%	75.03%	74.75%

Table 4
The results of experiment setup 750 features

Classifier	ACC	SEN	SPE
L-SVM	83.3%	83.18%	83.91%
Q-SVM	92.3%	91.99%	92.0%
C-SVM	95.7%	94.0%	94.99 %
MG-SVM	92.9%	92.01%	91.26%
CG-SVM	93.9%	93.03%	93.75%

Table 5
The results of experiment setup 2025 features

Classifier	ACC	SEN	SPE
L-SVM	91.11%	91.0%	91.93 %
Q-SVM	92.66%	91.99%	92.0%
C-SVM	97.59%	96.0%	97.91 %
MG-SVM	93.59%	92.99%	93.16%
CG-SVM	94.79%	93.99%	93.95%

The training time of the best outcomes is shown in Figure 6 and the prediction speed of the best outcomes is shown in Figure 7.

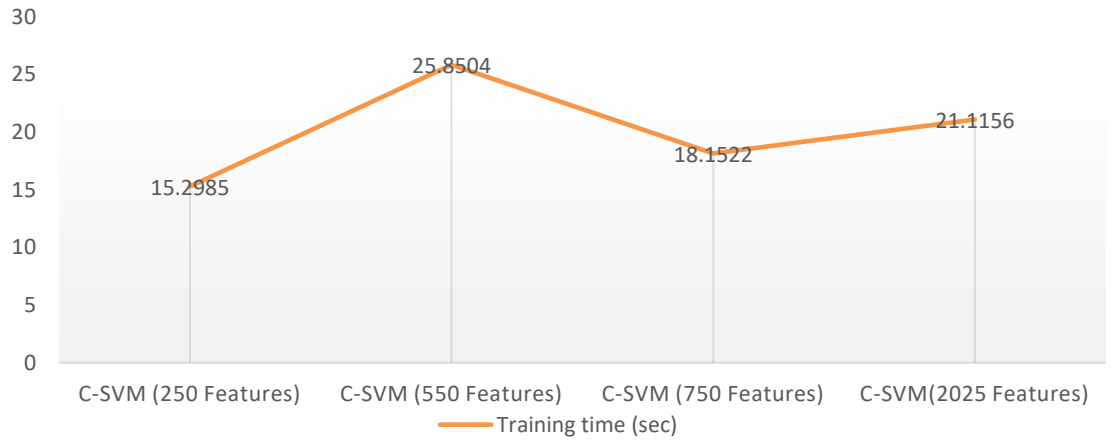

Fig. 6. Training time

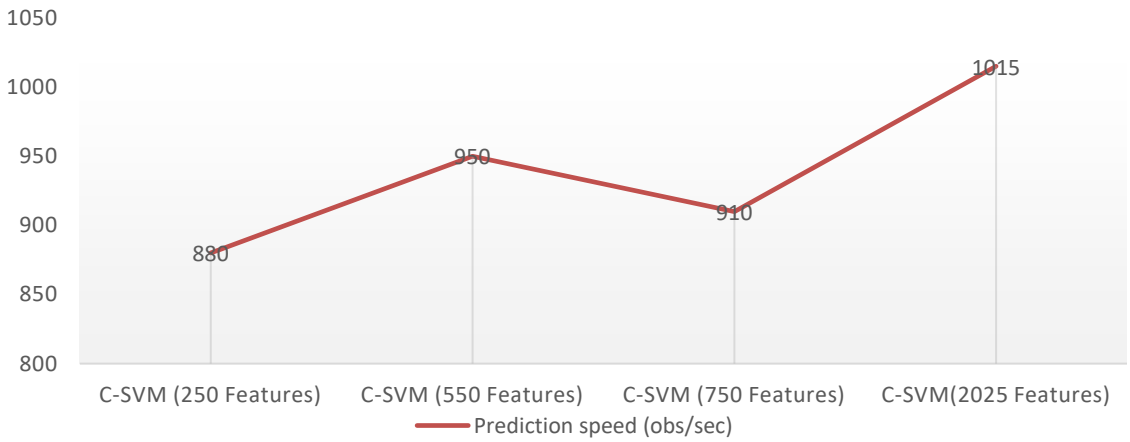

Fig. 7. Prediction speed

5. Discussion

This section discusses the detailed analysis of the experimental results obtained from various feature sets (250, 550, 750 and 2025 features) using different SVM classifiers (L-SVM, Q-SVM, FG-SVM, MG-SVM, CG-SVM and C-SVM). Each experiment was conducted to evaluate the performance of the proposed Deep CNN model combined with the Entropy-Coded Genetic Algorithm for feature selection in blood cell classification. The results highlight how the proposed model performs across different feature sets and classifiers, providing insights into the optimal configuration for achieving the best classification outcomes.

To start with the results of the experimental setup with 250 features (Table 2). The cubic SVM (C-SVM) achieved the highest accuracy of 91.1%, with a sensitivity of 90.99% and specificity of 90.0%. The quadratic SVM (Q-SVM) also performed well, with an accuracy of 89.0%, sensitivity of 89.26% and specificity of 88.26%. The linear SVM (L-SVM) had lower performance metrics, with an accuracy of 79.0%, sensitivity of 77.28% and specificity of 79.99%. The medium Gaussian SVM (MG-SVM) and coarse Gaussian SVM (CG-SVM) showed moderate performance, with accuracies of 89.6% and 74.2%, respectively. These results indicate that with 250 features, C-SVM and Q-SVM are more effective classifiers.

For the experimental setup with 550 features (Table 3), the C-SVM again outperformed other classifiers, achieving an accuracy of 93.2%, sensitivity of 93.96% and specificity of 92.99%. The Q-SVM followed closely with an accuracy of 91.7%, sensitivity of 90.26% and specificity of 91.76%. The L-SVM improved slightly compared to the 250 features setup, with an accuracy of 83.0%, sensitivity of 83.28% and specificity of 82.99%. The MG-SVM and CG-SVM showed better results than in the previous setup, with accuracies of 91.6% and 75.7%, respectively. These findings suggest that increasing the feature set to 550 enhances the performance of the classifiers, especially the C-SVM and Q-SVM.

In the setup with 750 features (Table 4), the C-SVM reached an impressive accuracy of 95.7%, sensitivity of 94.0% and specificity of 94.99%. The Q-SVM also showed strong performance, with an accuracy of 92.3%, sensitivity of 91.99% and specificity of 92.0%. The L-SVM maintained a steady performance with an accuracy of 83.3%, sensitivity of 83.18% and specificity of 83.91%. The MG-SVM and CG-SVM achieved accuracies of 92.9% and 93.9%, respectively. This setup highlights that further increasing the number of features continues to benefit classifier performance, particularly for the C-SVM.

Finally, in the results for the setup with 2025 features (Table 5), the C-SVM achieved the highest accuracy of 97.59%, with a sensitivity of 96.0% and specificity of 97.91%. The Q-SVM also performed well, with an accuracy of 92.66%, sensitivity of 91.99% and specificity of 92.0%. The L-SVM showed a marked improvement with an accuracy of 91.11%, sensitivity of 91.0% and specificity of 91.93%. The MG-SVM and CG-SVM had accuracies of 93.59% and 94.79%, respectively. These results demonstrate that using 2025 features significantly enhances the performance of the classifiers, with the C-SVM providing the most robust and efficient outcomes.

Moreover, from the training time (Figure 6) and prediction speed (Figure 7) of the C-SVM classifier across these different feature sets, the training time peaks at 25.85 seconds for 550 features, decreases to 18.15 seconds for 750 features and rises again to 21.12 seconds for 2025 features, indicating a non-linear relationship. Conversely, the prediction speed increases from 880 obs/sec with 250 features to 1015 obs/sec with 2025 features, with a slight dip at 750 features. This analysis demonstrates that while training time fluctuates with more features, prediction speed generally improves, highlighting the trade-off between training complexity and prediction efficiency.

In conclusion, the findings from our test cases indicate that the overall experimental arrangement produces the best results with 2025 features. Across all setups, the C-SVM consistently delivered superior and acceptable outcomes. The performance improvements with increasing feature sets validate the efficacy of the proposed Deep CNN model combined with the Entropy-Coded Genetic Algorithm for feature selection. Overall, the results demonstrated that the proposed model is highly effective for blood cell classification, particularly when using a large number of features.

6. Conclusion

In this paper, we proposed a novel deep CNN model with thirty-three layers, including eight convolutional layers and two feature map layers, specifically designed for blood cell classification. The model was pre-trained using a comprehensive medical imaging dataset containing 27,000 images from multiple sources, such as Brain CT-Scan, kvasir, Tumorslices, Tuberculosis, Pneumonia and COVID-19 radiology datasets. To enhance the focus on regions of interest (ROI), K-means clustering was employed during preprocessing.

The processed blood cell datasets were then utilized for classification. Feature extraction was conducted using the proposed Deep CNN model, which effectively captured critical features from the medical images. Subsequently, the Entropy-Coded Genetic Algorithm (GA) was applied for feature selection, where entropy assessed the randomness and grouping efficiency of the data. The GA iteratively evolved a population of potential solutions over multiple generations to identify the optimal set of features.

These selected features were then passed to various SVM classifiers (L-SVM, Q-SVM, FG-SVM, MG-SVM, CG-SVM and C-SVM) for the final classification task. The experimental results demonstrated that the proposed model achieved the highest performance with 2025 features, where the C-SVM classifier achieved an accuracy of 97.59%, sensitivity of 96.0% and specificity of 97.91%. This study highlights the effectiveness of combining a deep CNN model with an Entropy-Coded Genetic Algorithm for feature selection, providing a robust and efficient approach to blood cell classification. This method has the potential to significantly aid in early diagnosis and treatment by improving the accuracy and efficiency of blood cell classification.

Acknowledgment

Universiti Teknologi PETRONAS fully supports this research.

References

[1] Kuan, Da-Han, Chia-Chien Wu, Wei-Yu Su and Nien-Tsu Huang. "A microfluidic device for simultaneous extraction of plasma, red blood cells and on-chip white blood cell trapping." *Scientific reports* 8, no. 1 (2018): 15345. https://doi.org/10.1038/s41598-018-33738-8

[2] Farag, Mayada Ragab and Mahmoud Alagawany. "Erythrocytes as a biological model for screening of xenobiotics toxicity." *Chemico-biological interactions* 279 (2018): 73-83. https://doi.org/10.1016/j.cbi.2017.11.007

[3] Rezatofighi, Seyed Hamid and Hamid Soltanian-Zadeh. "Automatic recognition of five types of white blood cells in peripheral blood." *Computerized Medical Imaging and Graphics* 35, no. 4 (2011): 333-343. https://doi.org/10.1016/j.compmedimag.2011.01.003

[4] Faggio, Caterina, Antoni Sureda, Silvia Morabito, Ana Sanches-Silva andrei Mocan, Seyed Fazel Nabavi and Seyed Mohammad Nabavi. "Flavonoids and platelet aggregation: A brief review." *European journal of pharmacology* 807 (2017): 91-101. https://doi.org/10.1016/j.ejphar.2017.04.009

[5] Pandit, Amruta, Shrikrishna Kolhar and Pragati Patil. "Survey on automatic rbc detection and counting." *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering* 4, no. 1 (2015): 128-131. https://doi.org/10.15662/ijareeie.2015.0401012

[6] Al-Hafiz, Fatimah, Shiroq Al-Megren and Heba Kurdi. "Red blood cell segmentation by thresholding and Canny detector." *Procedia Computer Science* 141 (2018): 327-334. https://doi.org/10.1016/j.procs.2018.10.193

[7] Al-Dulaimi, Khamael Abbas Khudhair, Jasmine Banks, Vinod Chandran, Inmaculada Tomeo-Reyes and Kien 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)* (2018): 17-25.
[8] Rahadi, Irwan, Meechoke Choodoung and Arunsri Choodoung. "Red blood cells and white blood cells detection by image processing." In *Journal of Physics: Conference Series*, vol. 1539, no. 1, p. 012025. IOP Publishing, 2020. https://doi.org/10.1088/1742-6596/1539/1/012025
[9] van der Meijden, Paola EJ and Johan WM Heemskerk. "Platelet biology and functions: new concepts and clinical perspectives." *Nature Reviews Cardiology* 16, no. 3 (2019): 166-179. https://doi.org/10.1038/s41569-018-0110-0
[10] Dhurat, Rachita and MS25722595 Sukesh. "Principles and methods of preparation of platelet-rich plasma: a review and author's perspective." *Journal of cutaneous and aesthetic surgery* 7, no. 4 (2014): 189-197. https://doi.org/10.4103/0974-2077.150734
[11] Hegde, Roopa B., Keerthana Prasad, Harishchandra Hebbar and Brij Mohan Kumar 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* 39, no. 2 (2019): 382-392. https://doi.org/10.1016/j.bbe.2019.01.005
[12] Burn, Garth Lawrence, Alessandro Foti, Gerben Marsman, Dhiren Ferise Patel and Arturo Zychlinsky. "The neutrophil." *Immunity* 54, no. 7 (2021): 1377-1391. https://doi.org/10.1016/j.immuni.2021.06.006
[13] Robinson, Douglas S., A. Barry Kay and Andrew J. Wardlaw. "Eosinophils." *Inflammatory Mechanisms in Allergic Diseases* (2023): 43-75. https://doi.org/10.1201/9780429134432-5
[14] RP Siraganian. "Basophils." *Encyclopedia of Immunology (Second Edition), P. J. Delves Ed. Oxford: Elsevier*, (1998): 332-334. https://doi.org/10.1006/rwei.1999.0086
[15] Al-Shura, Anika Niambi. *Advanced Hematology in Integrated Cardiovascular Chinese Medicine: Volume 3*. Academic Press, 2019. https://doi.org/10.1016/B978-0-12-817572-9.00001-X
[16] Daskalopoulos, Evangelos P., Kevin CM Hermans, Lieke van Delft, Raffaele Altara and W. Matthijs Blankesteijn. "The role of inflammation in myocardial infarction." In *Inflammation in Heart Failure*, pp. 39-65. Academic Press, 2015. https://doi.org/10.1016/B978-0-12-800039-7.00003-7
[17] Mooney, P. "Blood Cell Count and Detection (BCCD)." https://www.kaggle.com/datasets/paultimothymooney/blood-cells
[18] Song, Huihui and Zheng Wang. "Automatic Classification of White Blood Cells Using a Semi-Supervised Convolutional Neural Network." *IEEE Access* (2024). https://doi.org/10.1109/ACCESS.2024.3380896
[19] Rahat, Irfan Sadiq, Mohammed Altaf Ahmed, Donepudi Rohini, A. Manjula, Hritwik Ghosh and Abdus Sobur. "A Step Towards Automated Haematology: DL Models for Blood Cell Detection and Classification." *EAI Endorsed Transactions on Pervasive Health and Technology* 10 (2024). https://doi.org/10.4108/eetpht.10.5477
[20] Liang, Minhui, Jianwei Zhong, Choo Sheriel Shannon, Rupesh Agrawal and Ye Ai. "Intelligent image-based deformability assessment of red blood cells via dynamic shape classification." *Sensors and Actuators B: Chemical* 401 (2024): 135056. https://doi.org/10.1016/j.snb.2023.135056
[21] Eshel, Yotam D., Uraib Sharaha, Guy Beck, Gal Cohen-Logasi, Itshak Lapidot, Mahmoud Huleihel, Shaul Mordechai, Joseph Kapelushnik and Ahmad Salman. "Monitoring the efficacy of antibiotic therapy in febrile pediatric oncology patients with bacteremia using infrared spectroscopy of white blood cells-based machine learning." *Talanta* 270 (2024): 125619. https://doi.org/10.1016/j.talanta.2023.125619
[22] Pullakhandam, Siddartha and Susan McRoy. "Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning." *BioMedInformatics* 4, no. 1 (2024): 661-672. https://doi.org/10.3390/biomedinformatics4010036
[23] Mohamad, Muhammad Arif and Muhammad Aliif Ahmad. "Handwritten Character Recognition using Enhanced Artificial Neural Network." *Journal of Advanced Research in Computing and Applications* 36, no. 1 (2024): 1-9. https://doi.org/10.37934/arca.36.1.19
[24] Zhahir, Amirul Asyraf, Siti Munirah Mohd, Mohd Ilias M. Shuhud, Bahari Idrus, Hishamuddin Zainuddin, Nurhidaya Mohd Jan and Mohamed Ridza Wahiddin. "Enhancing Quantum Information Processing–SU (2) Operator Model Development for Three-Qubit Quantum Systems Entanglement Classification." *International Journal of Computational Thinking and Data Science* 3, no. 1 (2024): 1-19. https://doi.org/10.37934/ctds.3.1.119
[25] Gu, Wencheng and Kexue Sun. "AYOLOv5: Improved YOLOv5 based on attention mechanism for blood cell detection." *Biomedical Signal Processing and Control* 88 (2024): 105034. https://doi.org/10.1016/j.bspc.2023.105034
[26] Zeng, Lanqing, Yusheng Fu, Jiuchuan Guo, Honghua Hu, Hongyu Li, Ning Wang, Xiwei Huang and Jinhong Guo. "AI-Based Portable White Blood Cells Classification and Counting System in POCT." *IEEE Sensors Journal* (2024). https://doi.org/10.1109/JSEN.2023.3348979

[27] Murmu, Anita and Piyush Kumar. "Dlrfnet: deep learning with random forest network for classification and detection of malaria parasite in blood smear." *Multimedia Tools and Applications* (2024): 1-23. https://doi.org/10.1007/s11042-023-17866-6
[28] Üzen, Hüseyin and Hüseyin Fırat. "A hybrid approach based on multipath Swin transformer and ConvMixer for white blood cells classification." *Health Information Science and Systems* 12, no. 1 (2024): 33. https://doi.org/10.1007/s13755-024-00291-w
[29] Arthur, David and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Stanford, 2006.
[30] Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).
[31] Muhammad, Taseer and Hamayoon Ghafory. "Sql injection attack detection using machine learning algorithm." *Mesopotamian journal of cybersecurity* 2022 (2022): 5-17. https://doi.org/10.58496/MJCS/2022/002
[32] Fayyaz, Abdul Muiz, Mudassar Raza, Muhammad Sharif, Jamal Hussain Shah, Seifedine Kadry and Oscar Sanjuán Martínez. "An integrated framework for COVID-19 classification based on ensembles of deep features and entropy coded GLEO feature selection." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 31, no. 01 (2023): 163-185. https://doi.org/10.1142/S0218488523500101
[33] Bein, Berthold. "Entropy." *Best Practice & Research Clinical Anaesthesiology* 20, no. 1 (2006): 101-109. https://doi.org/10.1016/j.bpa.2005.07.009
[34] Muiz Fayyaz, Abdul, Mahyar Kolivand, Jaber Alyami, Sudipta Roy and Amjad Rehman. "Computer Vision-Based Prognostic Modelling of COVID-19 from Medical Imaging." In *Prognostic Models in Healthcare: AI and Statistical Approaches*, pp. 25-45. Singapore: Springer Nature Singapore, 2022. https://doi.org/10.1007/978-981-19-2057-8_2
[35] Holland, John H. "Genetic algorithms." *Scientific american* 267, no. 1 (1992): 66-73. https://doi.org/10.1038/scientificamerican0792-66
[36] Mirjalili, Seyedali. "Evolutionary algorithms and neural networks." *Studies in computational intelligence* 780 (2019): 43-53. https://doi.org/10.1007/978-3-319-93025-1_4
[37] Halim, Zahid, Muhammad Nadeem Yousaf, Muhammad Waqas, Muhammad Sulaiman, Ghulam Abbas, Masroor Hussain, Iftekhar Ahmad and Muhammad Hanif. "An effective genetic algorithm-based feature selection method for intrusion detection systems." *Computers & Security* 110 (2021): 102448. https://doi.org/10.1016/j.cose.2021.102448
[38] Yue, Shihong, Ping Li and Peiyi Hao. "SVM classification: Its contents and challenges." *Applied Mathematics-A Journal of Chinese Universities* 18 (2003): 332-342. https://doi.org/10.1007/s11766-003-0059-5
[39] Fayyaz, Abdul Muiz, Muhammad Imran Sharif, Sami Azam, Asif Karim and Jamal El-Den. "Analysis of diabetic retinopathy (DR) based on the deep learning." *Information* 14, no. 1 (2023): 30. https://doi.org/10.3390/info14010030
[40] Suhaili, Shamsiah, Joyce Shing Yii Huong, Asrani Lit, Kuryati Kipli, Maimun Huja Husin, Mohamad Faizrizwan Mohd Sabri and Norhuzaimin Julai. "Development of digital image processing algorithms via fpga implementation." *Semarak International Journal of Electronic System Engineering* 3, no. 1 (2024): 28-45. https://doi.org/10.37934/sijese.3.1.2845
[41] Malik, Mehak Mushtaq, Abdul Muiz Fayyaz, Mussarat Yasmin, Said Jadid Abdulkadir, Safwan Mahmood Al-Selwi, Mudassar Raza and Sadia Waheed. "A novel deep CNN model with entropy coded sine cosine for corn disease classification." *Journal of King Saud University-Computer and Information Sciences* 36, no. 7 (2024): 102126. https://doi.org/10.1016/j.jksuci.2024.102126
[42] Singh, Amarjot, Shivesh Bajpai, Srikrishna Karanam, Akash Choubey and Thaluru Raviteja. "Malignant brain tumor detection." *International Journal of Computer Theory and Engineering* 4, no. 6 (2012): 1002-1006. https://doi.org/10.7763/IJCTE.2012.V4.626
[43] Pogorelov, Konstantin, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato *et al.,* "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection." In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164-169. 2017. https://doi.org/10.1145/3083187.3083212
[44] Er orhan, Feyzullah Temurtas and A. Çetin Tanrıkulu. "Tuberculosis disease diagnosis using artificial neural networks." *Journal of medical systems* 34 (2010): 299-302. https://doi.org/10.1007/s10916-008-9241-x
[45] Rahman, Tawsifur, Muhammad EH Chowdhury, Amith Khandakar, Khandaker R. Islam, Khandaker F. Islam, Zaid B. Mahbub, Muhammad A. Kadir and Saad Kashem. "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray." *Applied Sciences* 10, no. 9 (2020): 3233. https://doi.org/10.3390/app10093233