An Automated Detection and Classification of Retinopathy of Prematurity Stages Using SWIN Transformer

Nazar Salih Absulhussein[1,*], Bashar I. Hameed[1], Humam K. Yaseen[1], Nebras H. Ghaeb[2], Mohamed Ksantini[3]

[1]Department of Computer Science, Al-Imam Al-Adham University College, Baghdad, Iraq

[2]Biomedical Engineering Department, AL-Khwarizmi College of Engineering, Baghdad University, Baghdad, Iraq

[3]ATISP LAB, ENET'Com, University of Sfax, Sfax, Tunisia

Emails: nazarsalih@imamaladham.edu.iq; bashar_ibrahim@imamaladham.edu.iq; humam.khalid@imamaladham.edu.iq; nebras@kecbu.uobaghdad.edu.iq; mohamed.ksantini@ipeis.usf.tn

Abstract

Retinopathy of prematurity (ROP) remains the leading cause of blindness in children. The detection and treatment of this disease mainly depend on subjective evaluation of the features of retinal blood vessels. This method is not only time-consuming but also prone to errors. The increasing number of such cases demands an urgent need for automated models to improve the accuracy and efficiency of diagnosis and treatment. This paper presents a method for early detection of ROP using the Swin Transformer, a hierarchical vision transformer architecture. This work focuses solely on the screening stages for ROP, as documented between 2015 and 2020, based on a dataset composed of 3720 retinal images from preterm infants, kindly made available by the Al-Amal Eye Center located in Baghdad, Iraq. The proposed model achieved a classification accuracy of 98.67% on a clinical ROP dataset. The results highlight the importance of the most recent in-depth learning methods in enhancing early detection techniques, ultimately leading to improved clinical outcomes for at-risk infants.

Keywords: Health Care; Deep Learning; Retinopathy of Prematurity; Fundus Images; Swin Transformer

1. Introduction

Deep learning has been widely adopted in various domains, including medical imaging. However, Challenges remain in developing adaptive models under limited data conditions [1].

ROP is a notable contributor to childhood blindness, as it develops malformed blood vessels in the eye, leading to subsequent vision impairment [2]. ROP is a highly severe illness that impacts the eyes of premature infants [3]. Infants with a low birth weight and gestational age at birth of less than 32 weeks are susceptible to developing premature retinopathy, defined as weighing less than 1.5 kg at birth [4]. In Figure 1, the severity of ROP is determined according to the International Classification of Retinopathy of Prematurity (ICROP) guidelines published in 1984 [5],1987 [6],2005 [7], and 2021 [8]. ROP can be categorized into stages (1-5) and zones (1-3) based on the specific location of the illness within the body. This research will specifically evaluate the identification of stages "2-4" of ROP for various causes. 1- As previously stated, stage 1 demonstrates no major deviations from the normal state. Stage 5 is a fully detached retina, which is permanent and results in complete loss of vision. 3- The remaining stages need to be categorized, as they can be treated with laser, injectable, or surgical methods.

ROP remains the leading cause of juvenile blindness throughout the world today. The diagnosis and management of ROP present complex challenges because the clinical diagnosis of plus disease in ROP shows wide variations among specialists working in this field regarding the appropriate diagnostic methods. 2- The shortage of

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

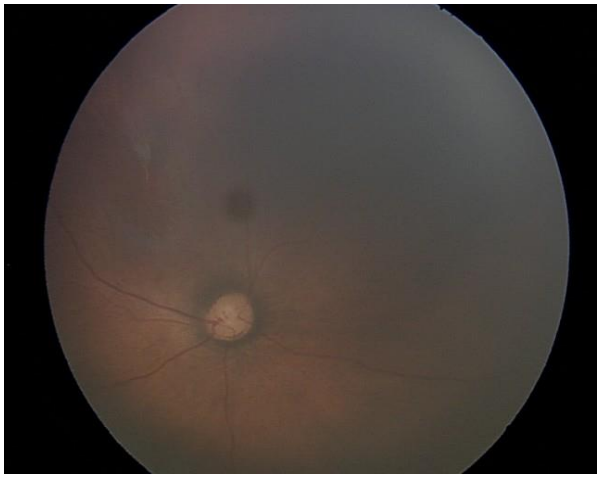
ophthalmologists and neonatologists who can treat ROP effectively stems from logistical issues, combined with extensive training duration, prolonged examination time, and high potential for mistakes. 3- The number of ROP cases in newborns continues to rise. Scientists have attempted to use quantitative and objective methods, such as computerized image analysis, to diagnose ROP due to these difficulties [9]. The development of Computer-Based Instructional Aid (CBIA) systems by various universities for ROP disorder detection has not produced an automated system that matches human medical diagnostic abilities. Professionals will be assisted in diagnosing by utilizing a fully automated and certified CBIA system, ensuring meticulous care. Additionally, it has the potential to enhance care accessibility by implementing widespread automated screening systems [10]. Artificial intelligence (AI) based learning models have recently gained significant popularity in medical picture analysis. The utilization of intelligent diagnostic technologies has become prevalent in the diagnosis of numerous disorders [11]. Deep learning (DL) is a state-of-the-art solution for many computational biology and artificial intelligence (CBIA) difficulties [12]. Furthermore, a recent development in deep learning marks a major improvement in the field of computer vision, as the Swin Transformer effectively manages high-resolution images by combining the features of transformers with innovative approaches. Its application in other fields indicates its adaptability and possible strength as a robust AI tool. When applied to medical imaging, particularly fundus imaging for ROP diagnosis, it can result in significant accuracy and efficiency gains.

In this study, we developed a new approach to detect and classify ROP stages using a built dataset of fundus images. To the best of our knowledge, the Swin transformer was used for the first time in ROP diagnosis. The proposed approach utilized a Swin transformer that leveraged a hierarchical vision transformer architecture with a shifted-window mechanism to efficiently and accurately process and analyze images. This approach was compared to that used in a previous study using the same dataset. Deep learning algorithms and belief function theory were developed to classify the same dataset for the ROP stages. The fusion of these models yielded an overall accuracy of 95.57% [13].

(a)

(b)

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025



(c)

Figure 1. System model: (a) stage 2, (b) stage 3, and (c) stage 4.

2. Related Work

This section critically reviews relevant literature and provides the necessary information upon which the proposed methodology is developed. The Swin Transformer, introduced by Liu et al. [14], represents a radical evolution in vision transformer architectures with hierarchical structures and shifted window-based attention, enabling an efficient way to learn long-range dependencies for visual data. This makes it particularly powerful when used on medical image tasks. Building on this, several others have taken the Swin Transformer and extended it for image analysis applications. For instance, Jingyun et al. [15] proposed SwinIR, an image restoration model that achieves excellent results in various degradation tasks. For example, Liao et al. [16] proposed their model-Swin-PANet-For segmentation of melanoma, the Swin Transformer would be placed within a computer-aided diagnosis system, but this would be under constraints of transfer learning. Li et al. [17] employed DnSwin combined with a wavelet-based denoising method to corrupt real images. This demonstrated the capacity of the Swin transformer to deal with multiple varieties of information. In the field of remote sensing, Hao et al. [18] developed TSTNet, a system that streams two data streams in different directions: one based on Sobel's method of edge detection and the other on gradient descent. They both produced impressive results on satellite imagery. For ocular diseases of the retina, Dihin et al. [19] employed a combination of multiple wavelets along with the Swin transform to identify diabetic retinopathy. The binary classification was highly accurate, but the multi-classification performance was poor, suggesting that it has limited scalability to complex class distributions. Other scientists. [20] Employed the Swin transformer to differentiate fundus images associated with five different types of diabetic eye disease, resulting in an increase in the AUC of 56.8% compared to conventional methods.

Transformer-based image processing models, particularly the Swin Transformer, have achieved increasing success in various medical and non-medical imaging tasks, demonstrating their value in ROP stage classification. Building on this success, this paper applies the Swin Transformer to a real-world ROP dataset, achieving improved performance and explaining the architecture's suitability for demanding clinical imaging applications.

While the Vision Transformer (ViT) [21] pioneered the direct application of transformer encoders to image classification, it requires very large datasets and struggles with limited medical data. The Data-efficient Image Transformer (DeiT) [22] improved sample efficiency through knowledge distillation, making it more practical in low-data regimes. However, both ViT and DeiT treat images as flat sequences of patches, which may limit their ability to capture local context. In contrast, the Swin Transformer introduces hierarchical feature maps and shifted window attention, allowing it to model both local vascular patterns and global retinal structures more effectively for ROP classification.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

3. Materials and Methods

Our methodology harnesses the Swin Transformer architecture to address the challenge of early ROP detection in fundus images. The Swin Transformer does not use convolutional layers but captures relationships among rich visual data at different scales through a hierarchical implementation over multiple stages. Different transformer blocks are used in each stage to ensure effective illustration of patch interactions, where these blocks employ distinct window attention mechanisms and shifted windows for non-linear feature representation enhancement, including Multi-Layer Perceptron (MLP) and Layer Normalization components. Information is analyzed using down-sampling and up-sampling between stages, which helps address different scale problems. A classification head is added for ROP prediction. The architecture is adaptive because it uses parameters efficiently, making it suitable for tasks related to medical image analysis. However, we take it further by pre-training the model on large datasets and then fine-tuning it specifically for early ROP identification tasks from fundus images. This approach significantly enhances the accuracy and efficiency of early retinopathy of prematurity detection. It has enabled us to utilize a Swin Transformer architecture to integrate the fine details of fundus images into ROP diagnosis and classification, thereby ensuring high accuracy by capturing all relevant information. The hierarchical structure, comprising multiple stages and transformer blocks, allows long-range dependencies to be modeled easily alongside patch interactions. In addition, model performance is enhanced through unique window attention techniques that optimize local context modelling; this is achieved by shifting windows, thus enabling the feature quality to be further improved. Layer Normalization with MLP components helps improve the total quality of features and their discriminative power; hence, a better non-linear representation of features within image data leads to better output.

Down-sampling and up-sampling can be used at every stage so that data from various scales are managed properly. Perspectives on the features of ROP are thereby made possible through views at different resolutions of fundus images. A classification head in the last stage makes it possible to accurately predict the emergence of ROP, which would be desirable since it is known well ahead of time, such that timely intervention can be implemented. It would align well with Swin Transformer for medical image analysis, as this model's performance ranks among the most efficient within one of the most economical parameter settings. The reason why Swin Transformer has proved to be so efficient in detecting ROP is possibly based on the fact that it enables modeling local interactions between image patches but keeps long-range dependencies between distant patches, which is important for contextual information over a large area, as indicated by vessel behavior. It's highly versatile and can be easily combined with other imaging technologies to create a complete solution for the early detection of ROP. One method of maximizing the potential of the Swin Transformer is to pre-develop and optimize it. The training of a large dataset benefits from the benefits of Transformer-based representation learning. During this initial phase, the model learns the details of fundus images, laying a solid foundation for further optimization for specific tasks related to early ROP detection from fundus images. The entire process focuses on developing a model that can detect ROP early from fundus images. Furthermore, fine-tuning on task-specific datasets can improve the representations learned by this developed model, thereby improving detection accuracy and robustness to internal variations in ROP fundus images.

Our method for recognizing ROP in fundus images is based on the power of the Swin Transformer design and utilizes a training and fine-tuning procedure. This approach has a significant impact on effectiveness and accuracy, which enables the early detection of ROP as a critical indicator of high-risk children. Our research paper highlights the contrasts between the stages of ROP upon examination, using explanatory examples from the data set in Figure 1. This highly effective and large dataset, taken from a renowned clinical setting, significantly enhances the model's training effectiveness and expands its application scope to clinical situations, including the early detection of ROP.

3.1 Preprocessing Steps for ROP Stages

3.1.1 Image Resizing

We downsized the original fundus images from a resolution of 640 x 480 pixels to 224 x 224 pixels. This step is critical for efficient model training and resource utilization.

3.1.2 Dataset Partitioning

Distinct suites have been created for training, testing, and validation. 70% of the dataset (3720 images) was used for training, 20% (744 images) for testing, and 10% (372 images) for validation, as shown in Table 1. This division ensures separate subgroups for typical training, assessment, and performance evaluation.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

3.1.3 Exclusion of Low-Quality Images

Removing hazy, fuzzy, or dark images improves the dataset's quality. They have conducted a thorough quality check and excluded images with low clarity or visibility. This step ensures the model is trained on high-quality data, improving its generalization ability to real-world scenarios.

3.1.4 Avoiding Dataset Overlap

Special training, testing, and verification groups have been established. 70% of the data set (3720 images) was used for training, 20% (744 images) for testing, and 10% (372 images) for validation. This division includes separate sub-clusters for pilot training, evaluation, and performance evaluation.

3.1.5 Dataset Class Distribution

We maintained a balanced distribution of ROP stages across datasets. They examined the distribution of ROP stages (Stage 2, Stage 3, Stage 4) in each dataset subset (Train set, Test set, Validation set). This documentation facilitates understanding of the dataset's composition and guides the model training process. These preprocessing steps, specifically tailored for ROP stages, contribute to preparing a focused and well-structured dataset for training and evaluating the Swin Transformer model in the context of early ROP detection in fundus images.

3.2 Training Procedure and Hyperparameters

In the training phase of the Swin Transformer model for early detection of ROP stages in fundus images, specific hyperparameters were meticulously chosen to optimize the learning process. The learning rate of 0.001 facilitated a balanced convergence rate during the improvement process. Using the 64-image batch size (BS) for each frequency, it achieved a trade-off between computational efficiency. The module was trained 200 times, ensuring comprehensive exposure to the dataset for optimal learning of its advantages. To assess the model performance and prevent overprocessing, a 15% cross-validation split was applied, with a distinct sub-cluster dedicated to performance evaluation during training. These super-markers, including the learning rate, batch size, number of epochs, and validation partitioning, are adjusted through experience to balance effective convergence with robust dissemination, thereby enhancing the effectiveness of the Swin Transformer in the early detection of ROP.

Table 1: Dataset for Retinopathy of Prematurity (ROP)

Page	Stage2	Stage3	Stage4
Train set (70%)	968	942	955
Validation set (10%)	276	270	273
Test set (20%)	138	135	136
Total	1382	1347	1346

3.3 Proposed Methodology

The Swin Transformer architecture is intricately designed, as shown in Figures 2 and 3, commencing with the 'Patch Partition' block responsible for segmenting input images into smaller patches. Subsequently, four stages, each housing one or more Swin Transformer blocks, are employed to correct and transform features iteratively. At the apex of each stage, patch merging or linear embedding is applied, termed 'Patch Merging' or 'Linear Embedding' exclusively in the initial layer. This process involves reducing the number of distinctive tokens by a factor of 4, resulting in an effective down-sampling of resolution by a factor of 2. Consequently, a pyramidal-shaped feature map emerges due to varying resolutions at each stage.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

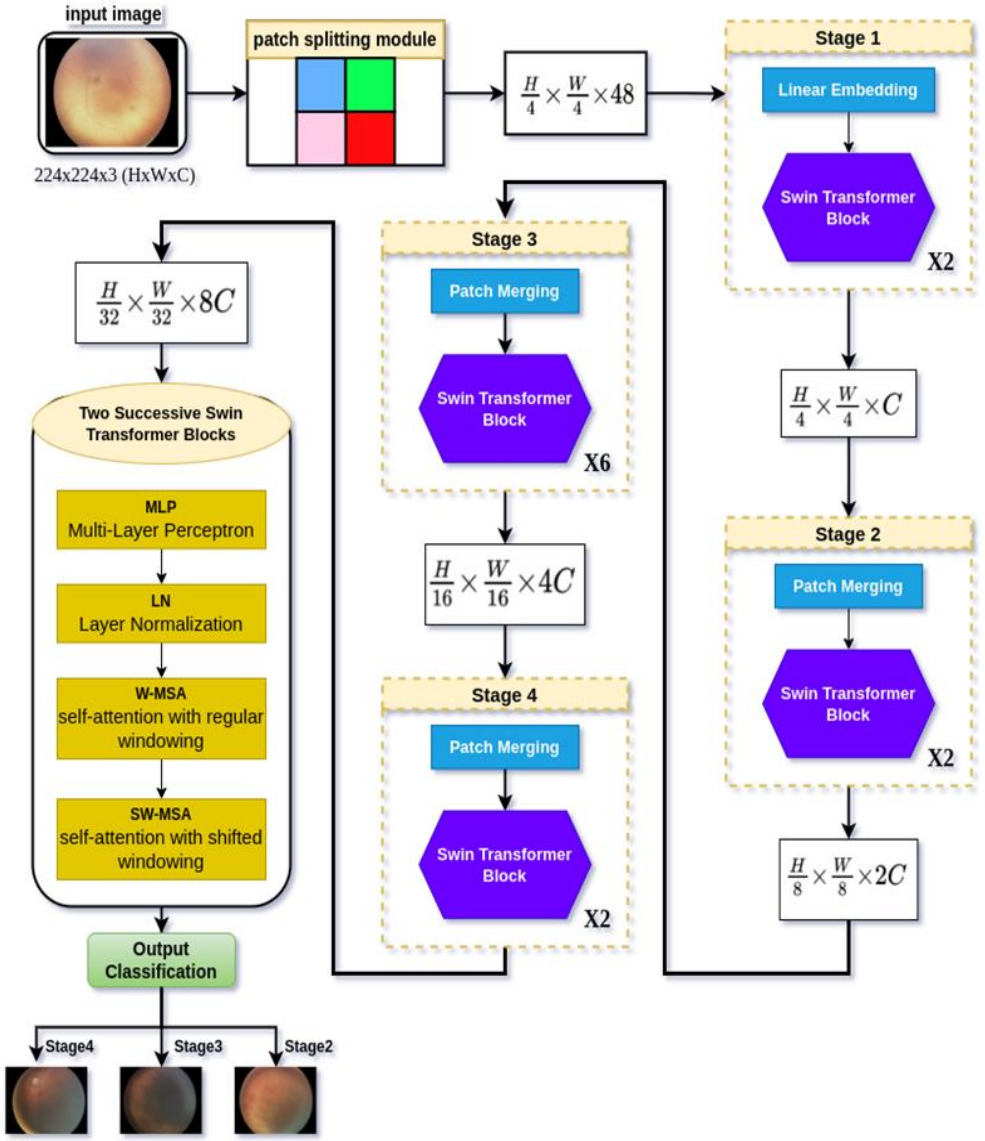


Figure 2. The suggested method

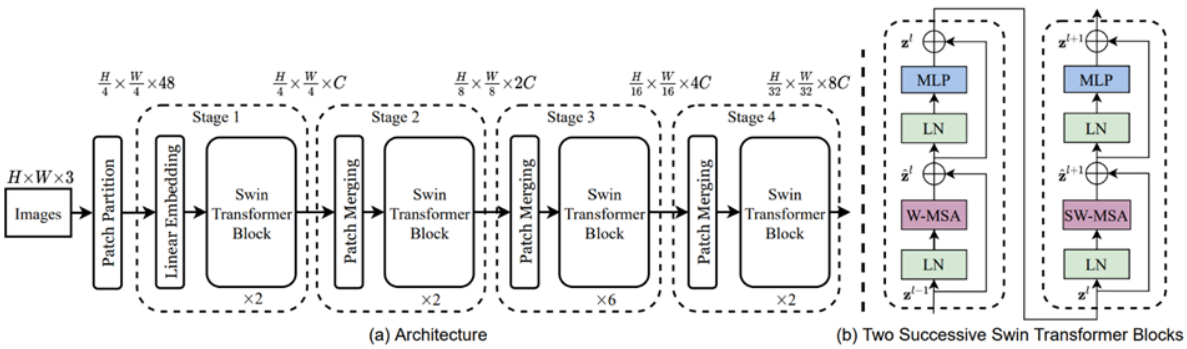


Figure 3. SWIN Transformer architecture

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

The final block, denoted 'Two Successive Swin Transformer Blocks', consists of Multi-Layer Perceptron (MLP), Layer Normalization (LN), Window-based Multi-Head Self-Attention (W-MSA), and Multi-Head Self-Attention Module with Regular Windowing (SW-MSA).

For our ROP classification task, we employed the Swin-Tiny variant with four hierarchical stages. Each fundus image (224×224) was partitioned into non-overlapping 4×4 patches, producing 56×56 tokens. We used an embedding dimension of 96 with depths [2, 2, 6, 2] and a window size of 7, balancing accuracy with computational feasibility. To adapt the architecture to multiclass classification (stages 2–4), we replaced the default ImageNet head with a three-class fully connected layer. Cross-entropy loss was used, with class weights adjusted to handle a slight imbalance between stages. Fine-tuning was performed after initializing with ImageNet pre-trained weights, while freezing the first two stages during the initial 20 epochs to stabilize training on medical images.

3.4 Evaluation Metrics

The trained models were evaluated using Precision, Recall, F1-measure, and area under the curve (AUC). The accuracy and relevance of the model trained in classifying ROP stages in this study have been assessed. Regardless of their absolute values, accuracy analyzes the relationship between many quantities and measures the trend toward health. With the help of Eq. 1, we can calculate accuracy by dividing the total number of real positives (TP) by the sum of both TP and FP.

$$\text{Precision}=TP/(TP+FP) \quad (1)$$

One method of measuring the number of comparable duplicate data points in a dataset is recall. It is, therefore, of the utmost importance to distinguish them from precision. As explained in the equation, recall is calculated by dividing the number of true positives (TPs) by the number of positive repetitions, where the positive category includes both the number of false negatives (FNs) and the number of TPs.

$$\text{Recall}=TP/(FN+TP) \quad (2)$$

F1-Score comprehensively assesses the model's performance in test data sets by including both recall and precision. It varies from 0 to 1 and is determined using the compromise medium for summoning and accuracy. Equation 3 provides the F1 degree formula.

$$\text{F1 Score}=2 \times (\text{Precision}.\text{Recall})/ (\text{Precision}+ \text{Recall}) \quad (3)$$

The area under the curve (AUC) in the receiver operating characteristic (ROC) is a vital performance measure that illustrates the model's ability to distinguish between distinct categories. The area under the more significant curve indicates a better detection model, so consider that. Equation 4 calculates the true positive rate (TPR) and the false positive rate (FPR), resulting in the ROC curve.

$$FPR=FP/(FP+TN) \quad (4)$$

4. Results and Discussion

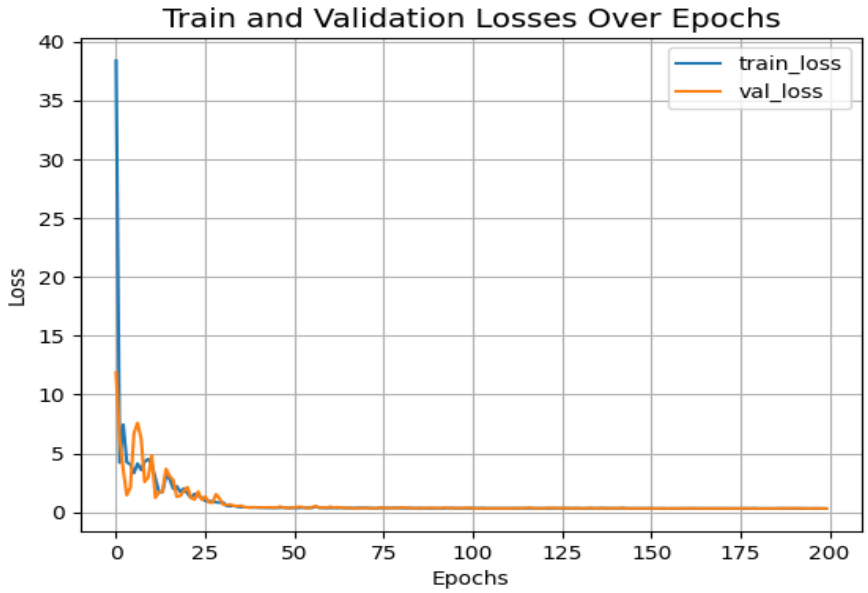
The research work discussed the evaluation of the relationship between various stages of ROP with complete details and analysis. The development process was from inception to completion. SWIN transformers were used to develop the models while a wide range of different-natured datasets were used in training; accuracy and recall, F1 score, and AUC metrics were mainly considered for effective prediction creation. By applying the capabilities of state-of-the-art transformers, the study aims to gain valuable insights into the complex relationship between the various stages of ROP and early birth. Such an understanding sought is about this intricate medical condition, which will lead to better diagnosis or treatment, thereby paving the way for a complete eradication.

4.1 Experimental setup

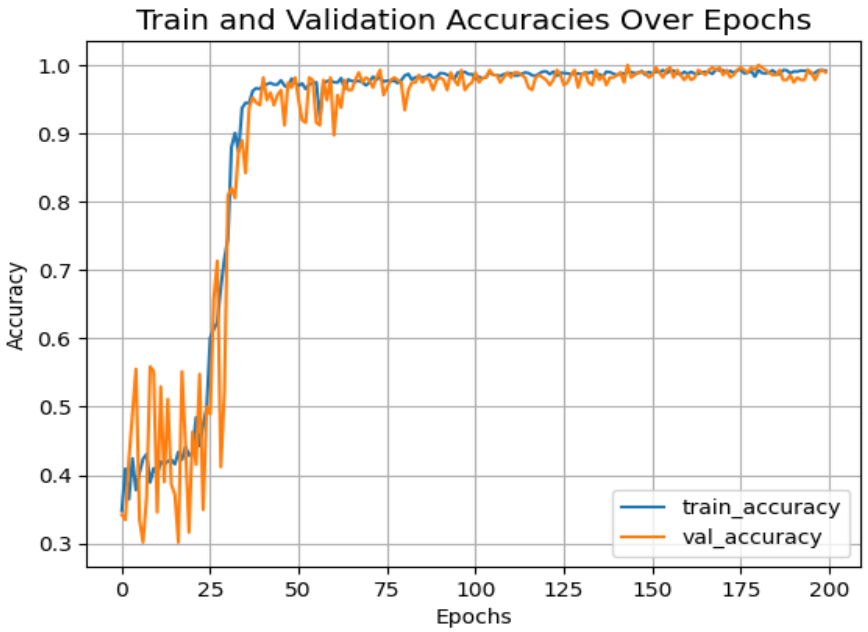
The modules were trained for 200 epochs at a rate of 0.001 per repeat and the size of the 64th batch. Adam was the optimizer, with cross-entrepreneur loss as the loss function. The training images are supplemented by random flipping and horizontal rotation to make the training dataset more diverse and less prone to overfitting. Model training has become more efficient using GPU acceleration. Swin Transformer underwent comprehensive training independently and was evaluated using the F1 score, recall, accuracy, and precision. Swin Transformer was the most effective in detecting precipitous retinal malformation early, with the best accuracy, precision, recall, and F1 score. The analysis was conducted on an Intel Core i7 computer with a random-access memory (RAM) of 8 gigabytes and a central processing unit at a speed of 2.7 gigahertz. Scikit-learn is an open-source machine learning program based on Python. To make the study analysis more efficient and accessible, we used Google Colab, a free web-based and open-source basic system, to create, share, and cooperate in real-time with reports, images, equations, and encrypted prose.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

Additionally, the models have undergone a comprehensive test and review using various datasets and rules to ensure they are robust and ready for sharing. The training included the careful setting of hyperparameters to achieve better results and minimize the risk of bias or excessive processing. Additionally, extensive tests have examined various design choices and setups to achieve the best possible outcome. Training has been distributed by utilizing more computer power, including multiple GPUs and parallel systems. Using this method ensured a fair build cost and work balance, providing good handling in real terms and dealing with large datasets, thereby reducing the time needed for training. New models have been developed based on improved steps and up-to-date group methods that ensure the best approach is taken and incorrect use is avoided.



(a)



(b)

Figure 4. Training and validation over epochs for the ROP stages dataset, (a) loss, (b) accuracy

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

Great computational power has led to distributed training methods, where users have more than one graphic processing unit, along with a parallel computing framework. This pathway not only leads to development and efficient treatment but also enables the management of big data very quickly, thereby reducing the time required for training. Better algorithms and modern methods of organization enable trained models to optimize convergence without misuse.

Additionally, high-level data processing is discussed in the training pipeline, where advantages and normalization measurements are calculated before dimensionality reduction. These preprocessing steps helped to firmly merge the input data and extract features that would be useful for the model to learn from and make predictions later on. Also, high-level data processing in the training pipeline includes benefit measurement, normalization, and dimensionality reduction. These preparatory steps enable the reliable combination of input data and the extraction of useful features, allowing for the effective learning of models designed to achieve accurate predictions. The combination of advanced model architectures, rigorous training procedures, state-of-the-art algorithms, and powerful computing resources often leads to superior performance in early network detection. This study is useful in medical image analysis because it proves the worth of transformer-based architectures in ROP stage classification. Other studies can easily be borrowed from it. Medical image analysis is also an emerging field that demonstrates the potential of automated learning to improve healthcare outcomes. The visual representation of the training process is shown in Figures 4 and 5

Table 2: Dataset for Retinopathy of Prematurity (ROP)

epoch	Train-Acc	Train-Loss	Val-Acc	Val-Loss
10	0.4101	4.5261	0.5515	2.9745
20	0.4285	2.0223	0.3162	1.8037
30	0.7157	0.8163	0.5221	1.1593
40	0.9648	0.4015	0.9412	0.4175
50	0.9742	0.3689	0.9816	0.3540
60	0.9775	0.3606	0.9816	0.3509
70	0.9701	0.3663	0.9816	0.3614
80	0.9775	0.3620	0.9743	0.3594
90	0.9820	0.3466	0.9779	0.3658
100	0.9865	0.3398	0.9743	0.3860
150	0.9898	0.3305	0.9816	0.3309
200	0.9918	0.3137	0.9890	0.3197

The presented table (Table 2) provides a comprehensive overview of the performance metrics of a machine learning model across various epochs, illustrating its development during the training process. In the initial epochs (10-30), the training accuracy steadily ascended from 41.01% to 71.57%, coinciding with a decline in training loss from 4.5261 to 0.8163. This indicates that the model is progressively learning and improving its ability to classify instances in the training dataset accurately. Nevertheless, the validation accuracy fluctuates during this period, suggesting potential difficulties in generalizing to new data. Following this, in epochs 40-100, training and validation accuracies significantly improve, reaching impressive values exceeding 95%. The training loss continues to decrease, demonstrating the model's effective adaptation to the training data, while the validation loss also diminishes, showcasing improved generalization. The model displays considerable stability in the later epochs (150-200), achieving a peak validation accuracy of 98.16% with minimal validation loss (0.3309). These results collectively indicate an efficiently performing model with strong learning capabilities and effective generalization to novel, unseen data, as evidenced by robust accuracy and low loss values across training and validation datasets.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

```
Epoch 196/200
39/39 [==============================] - 31s 802ms/step - loss: 0.3251 - accuracy: 0.9865 - top
-5-accuracy: 1.0000 - val_loss: 0.3175 - val_accuracy: 0.9890 - val_top-5-accuracy: 1.0000
Epoch 197/200
39/39 [==============================] - 31s 803ms/step - loss: 0.3203 - accuracy: 0.9885 - top
-5-accuracy: 1.0000 - val_loss: 0.3308 - val_accuracy: 0.9779 - val_top-5-accuracy: 1.0000
Epoch 198/200
39/39 [==============================] - 31s 803ms/step - loss: 0.3154 - accuracy: 0.9922 - top
-5-accuracy: 1.0000 - val_loss: 0.3233 - val_accuracy: 0.9890 - val_top-5-accuracy: 1.0000
Epoch 199/200
39/39 [==============================] - 31s 803ms/step - loss: 0.3151 - accuracy: 0.9922 - top
-5-accuracy: 1.0000 - val_loss: 0.3146 - val_accuracy: 0.9926 - val_top-5-accuracy: 1.0000
Epoch 200/200
39/39 [==============================] - 31s 803ms/step - loss: 0.3137 - accuracy: 0.9918 - top
-5-accuracy: 1.0000 - val_loss: 0.3197 - val_accuracy: 0.9890 - val_top-5-accuracy: 1.0000
```

Figure 5. Part of the training process implementation for classification using the Swin transformer

4.2 Confusion matrix

A confusion matrix is a tabular representation that provides a concise summary of the number of accurate positive predictions, inaccurate positive forecasts, accurate negative predictions, and inaccurate negative predictions for each class. Figure 6 displays the confusion matrix of the SWIN transformer. The table's rows correspond to the data's true labels, while the columns correspond to the anticipated labels. The figures in the table indicate the frequency with which the algorithm accurately or inaccurately identified each incident. The confusion matrix in this image displays the predicted labels as "stage2", "stage3", and "stage4", but the actual labels also correspond to "stage2", "stage3", and "stage4".

The Swin Transformer's architectural design is quite complex, beginning with the Patch Partition module, which is responsible for partitioning the image into smaller patches. This ensures that the input is optimally distributed for further processing. After initialization, the architecture consists of four stages, each containing one or more Swin Transformer modules to match features and transfer them to another domain. Absorbers play a key role in building the foundation. They ensure that information from different layers is effectively captured, allowing unique strengths to be linearly combined at the top level. This is where patch merging (or linear embedding) comes into play, performed only at the primary level. This merges patches to reduce the number of unique symbols, typically by a factor of four. This also does not incur any additional loss of accuracy and is again divided by two. It guarantees structural flexibility to accommodate images of different sizes, and makes it a lot more mathematical, especially after we reduce.

This produces a feature map of pyramidal varieties at different resolutions at each stage. The last block, "Two Successive Swin Transformer Blocks", stacks twice the Swin Transformer blocks one after another in a manner that allows for very fine granularity of feature transformation, thereby improving the expressivity of the entire architecture. These modules consist of two main components: an MLP (Multi-layer Perceptron) and a Layer Normalization (LN). After performing token-level self-attention, a simple feed-forward MLP layer updates the features of each token. As a multi-layer artificial neural network with linear and non-linear transformations, it can learn complex patterns and dependencies in visual data.

However, the LN layer employs additional normalization factors beyond the feature dimension; this is beneficial because it prevents the distribution of input values from becoming unstable. This directly increases the stability of training, which in turn increases the overall performance of the model. Additionally, the Swin Transformer employs a two-stage mechanism for attention: W-MSA (window-based multiple heads of attention) and SW-MSA (multiple heads of attention with a regular window). W-MSA is different from traditional multiple-headed self-observation mechanisms, which are situated around every input location. This facilitates the focusing of attention on local dependencies that depend on features from other tokens, enabling the efficient transmission of information across small regions.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

SW-MSA is similar to W-MSA, but instead of a full width and height, it has a window that covers half of the width and height. The window is incremented in uniform steps within the original image. This facilitates communication between the W-MSA layers, labeling, long-range relationships, and essential features. This is particularly important when a comprehensive understanding of the context of images is required. This detailed design and the manner in which components' features are incorporated into the Swin Transformer design make it highly effective for complex image processing tasks, including the classification of images (where it correctly identifies the classes of interest) and the segmentation of images (where it identifies and describes the objects of interest). The Swin Transformer's capacity to harvest both near and long-range dependencies, along with effective downsampling and attribute refinement, makes it a great choice for many practical applications that require complex image analysis and comprehension.

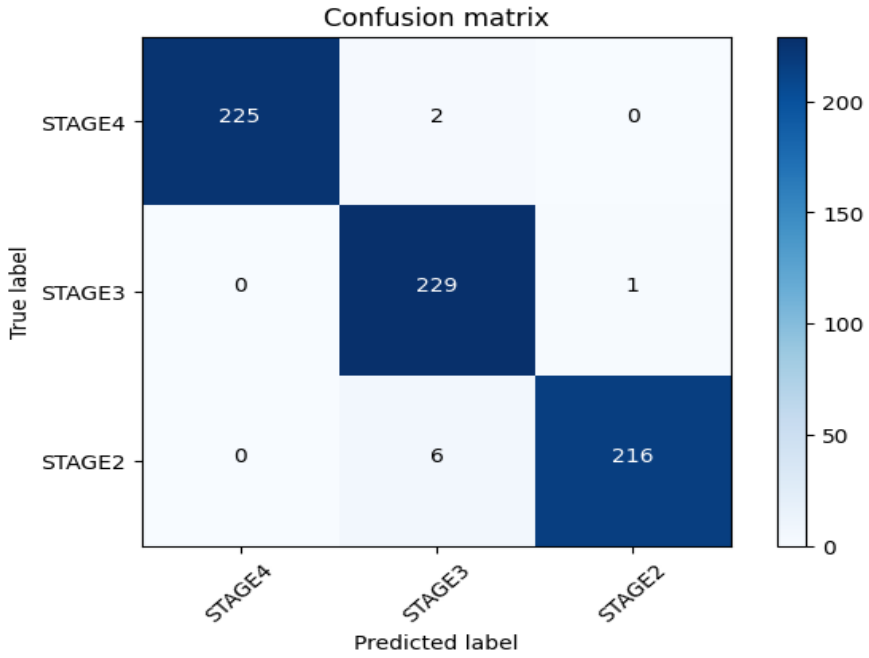


Figure 6. Confusion matrix of the Swin transformer

The Swin Transformer is intricate. Its first module, "Patch Partition," centers the image on a smaller, more focused scope. This facilitates the efficient distribution of the input for additional processing. Once built, the structure will undergo four steps. Each step consists of one or more blocks of the Swin Transformer type for further refinement and is then transferred to another field. These steps are important components of the architecture because they accurately gather information from different layers and represent it linearly, without necessitating the top layer to wait for the other layers to have finished their tasks before taking advantage of them. This is referred to as block merging/linear embedding—only performed at the top layer, the merge patch, and the removal of special symbols results in each block having four times fewer blocks than would actually exist, resulting in a twofold decrease in accuracy. This facilitates the structure's flexibility to adapt to different-sized images while still maintaining its mathematical complexity after downsampling.

This creates a feature map pyramid, considering different resolutions at each stage. The final block, "Two Consecutive Swin Transformer Blocks," consists of two consecutive occurrences of a Swin Transformer block. This enables a more refined transformation of features and enhances the generality of the architecture. There are two components in these blocks: Multi-Layer Perceptron (MLP) and Layer Normalization (LN). After the performance of a self-attention calculation, MLP layers update token features in separate paths. It can encode complex patterns and dependencies in visual data due to multimedia content being processed through a high-dimensional artificial neural network comprising many layers of linear and non-linear transformations.

SW-MSA is a sliding window version of W-MSA that shifts the window by half its size at all places in the original height and width dimensions of an image. SW-MSA layers have enough capacity for modeling long-range dependencies at the level of distinct tokens, thus opening rudimentary communication between multiple windows when both an image plus possibly all its context, must be understood for a certain task. Therefore, it enables the

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

Swin Transformer to support general and complex visual tasks ranging from image classification, i.e., inferring class labels of images, to image segmentation, i.e., predicting objects or regions of interest present in an image. We demonstrate that maintaining local and global links with the Swin Transformer is facilitated by its effective down-sampling and feature cleaning steps, providing a foundational component for various applications requiring high-level image analysis and understanding capabilities.

4.3 Comparison with Previous Studies

The Swin Transformer model is a new architecture for image classification that was applied in this paper. As a result, our model achieved an accuracy of 98.67%, surpassing the best accuracy reported in the previous study by a substantial margin, as the best result from earlier work was 95.57% [13]. This further improves the demonstration of how transformer-based models can review medical images to facilitate early analysis of retinopathy of prematurity, which is mostly found in preterm infants. The increased effectiveness of Transformer-based models suggests that these systems could be employed to augment automated diagnostic systems and introduce a revolution in early diagnosis and treatment, ultimately benefiting the health of preterm infants at risk of ROP.

Compared to other Transformer models such as ViT and DeiT, the Swin Transformer is particularly well-suited for ROP analysis because its hierarchical design reduces computational workload while preserving local structural information, which is critical for detecting subtle, stage-related retinal vascular changes.. This contextualizes our choice of Swin Transformer over other vision transformers for this task.

6. Conclusion

In conclusion, the early detection of ROP is essential for preventing irreversible vision loss in premature infants. ROP has a significant impact on childhood blindness. This study proposes a novel approach that utilizes the Swin Transformer architecture for early identification.

By leveraging a dataset of 3720 retinal images collected over six years by a private ophthalmology clinic, our model achieved an accuracy rate of 98.76% in distinguishing between stage 2, stage 3, and stage 4. This high accuracy underscored the potential of advanced deep learning methodologies in enhancing clinical outcomes in infants at risk of ROP.

In the future, we will focus on the continued development of algorithms, additional methodologies, and the creation of a larger training dataset, all of which will contribute to advancing medical reform in these current circumstances.

Funding: "This research received no external funding."

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] G. Skedsmo and S. G. Huber, "Assessing learning gaps and gains?," *Educ. Assess. Eval. Account.*, vol. 35, no. 4, pp. 471–473, Nov. 2023, doi: 10.1007/s11092-023-09423-4.

[2] E. H. Hong, Y. U. Shin, and H. Cho, "Retinopathy of prematurity: a review of epidemiology and current treatment strategies," *Clin. Exp. Pediatr.*, vol. 65, no. 3, pp. 115–126, Oct. 2021, doi: 10.3345/cep.2021.00773.

[3] J. M. Brown et al., "Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks," *JAMA Ophthalmol.*, vol. 136, no. 7, p. 803, Jul. 2018, doi: 10.1001/jamaophthalmol.2018.1934.

[4] Early Treatment for Retinopathy of Prematurity Cooperative Group, "The Incidence and Course of Retinopathy of Prematurity: Findings From the Early Treatment for Retinopathy of Prematurity Study," *Pediatrics*, vol. 116, no. 1, pp. 15–23, Jul. 2005, doi: 10.1542/peds.2004-1413.

[5] "An international classification of retinopathy of prematurity," *Pediatrics*, vol. 74, no. 1, pp. 127–133, Jul. 1984.

[6] "An international classification of retinopathy of prematurity. II. The classification of retinal detachment. The International Committee for the Classification of the Late Stages of Retinopathy of Prematurity," *Arch. Ophthalmol.*, vol. 105, no. 7, pp. 906–912, Jul. 1987.

[7] International Committee for the Classification of Retinopathy of Prematurity, "The International Classification of Retinopathy of Prematurity revisited," *Arch. Ophthalmol.*, vol. 123, no. 7, pp. 991–999, Jul. 2005, doi: 10.1001/archopht.123.7.991.

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025

[8] M. F. Chiang et al., "International Classification of Retinopathy of Prematurity, Third Edition," *Ophthalmology*, vol. 128, no. 10, pp. e51–e68, Oct. 2021, doi: 10.1016/j.ophtha.2021.05.031.

[9] J. Wang et al., "A Deep Learning System for Automated Diagnosis of Plus Disease in Retinopathy of Prematurity with Quantifiable Measurement of Vascularity," *JAMA Ophthalmol.*, vol. 140, no. 5, pp. 491–499, May 2022, doi: 10.1001/jamaophthalmol.2022.0783.

[10] G. M. Richter, S. L. Williams, J. Starren, J. T. Flynn, and M. F. Chiang, "Telemedicine for Retinopathy of Prematurity Diagnosis: Evaluation and Challenges," *Surv. Ophthalmol.*, vol. 54, no. 6, pp. 671–685, Nov. 2009, doi: 10.1016/j.survophthal.2009.02.020.

[11] Das et al., "Breast cancer detection using an ensemble deep learning method," *Biomed. Signal Process. Control*, vol. 70, p. 103009, Sep. 2021, doi: 10.1016/j.bspc.2021.103009.

[12] Bhandary et al., "Deep-learning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images," *Pattern Recognit. Lett.*, vol. 129, pp. 271–278, Jan. 2020, doi: 10.1016/j.patrec.2019.11.013.

[13] N. Salih et al., "An Advanced Approach for Predicting ROP Stages: Deep Learning Algorithms and Belief Function Technique," *Iraqi J. Sci.*, pp. 4047–4060, Jul. 2024, doi: 10.24996/ijs.2024.65.7.39.

[14] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv:2103.14030*, Aug. 2021.

[15] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," *arXiv:2108.10257*, Aug. 2021.

[16] Z. Liao, K. Xu, and N. Fan, "Swin Transformer Assisted Prior Attention Network for Medical Image Segmentation," in *Proc. 8th Int. Conf. Comput. Artif. Intell.*, 2022, pp. 491–497, doi: 10.1145/3532213.3532287.

[17] H. Li et al., "DnSwin: Toward real-world denoising via a continuous Wavelet Sliding Transformer," *Knowl.-Based Syst.*, vol. 255, p. 109815, Nov. 2022, doi: 10.1016/j.knosys.2022.109815.

[18] S. Hao, B. Wu, K. Zhao, Y. Ye, and W. Wang, "Two-Stream Swin Transformer with Differentiable Sobel Operator for Remote Sensing Image Classification," *Remote Sens.*, vol. 14, no. 6, p. 1507, Mar. 2022, doi: 10.3390/rs14061507.

[19] R. A. Dihin, E. AlShemmary, and W. Al-Jawher, "Diabetic Retinopathy Classification Using Swin Transformer with Multi Wavelet," *J. Kufa Math. Comput.*, vol. 10, no. 2, pp. 167–172, Aug. 2023, doi: 10.31642/JoKMC/2018/100225.

[20] Md. M. Haque, S. Akter, and A. F. Ashrafi, "SwinMedNet: Leveraging Swin Transformer for Robust Diabetic Retinopathy Classification from the RetinaMNIST2D Dataset," in *Proc. 6th Int. Conf. Elect. Eng. Inf. Commun. Technol. (ICEEICT)*, May 2024, pp. 1286–1291, doi: 10.1109/ICEEICT62016.2024.10534544.

[21] Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021. [Online]. Available: https://arxiv.org/abs/2012.12877

DOI: https://doi.org/10.54216/FPA.210215
Received: April 15, 2025 Revised: June 19, 2025 Accepted: August 30, 2025