Diabetic Retinopathy Classification Using Swin Transformer with Multi Wavelet

Rasha Ali Dihin
Faculty of Computer Science and Mathematics
Department of Mathematics
University of Kufa
Najaf, Iraq
rashaa.aljabry@uokufa.edu.iq
Orcid.org/0000-0002-3857-4013

Ebtesam N. AlShemmary
IT Research and Development Center
University of Kufa
Najaf, Iraq
dr.alshemmary@uokufa.edu.iq
Orcid.org/0000-0001-7500-9702

Waleed A. Mahmoud Al-Jawher
Uruk University
Bagdad, Iraq
profwaleed54@gmail.com

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100225
Received Jan. 3, 2023. Accepted for publication Aug.17, 2023

Abstract— Diabetic retinopathy (DR) impacts over a third of individuals diagnosed with diabetes and stands as the leading cause of vision loss in working-age adults worldwide. Therefore, the early detection and treatment of DR can play a crucial role in minimizing vision loss. This research paper proposes a novel technique that combines Wavelet and multi-Wavelet transforms with Swin Transformer to automatically identify the progression level of diabetic retinopathy. A notable innovation of this study lies in the implementation of the multi-Wavelet transform for extracting relevant features. By incorporating the resulting images into the Swin Transformer model, a unique approach is introduced during the feature extraction phase. The researchers conducted experiments using the publicly available Kaggle APTOS 2019 dataset, which comprises 3662 images. The achieved training accuracy in the experiments was an impressive 97.78%, with a test accuracy of 97.54%. The highest accuracy observed during training reached 98.09%. In comparison, when applying the multi-Wavelet approach to multiclass classification, the training and validation accuracies were 91.60% and 82.42%, respectively, with a testing accuracy of 82%. These results indicate that the multi-Wavelet approach outperforms alternative methods in the study. The model demonstrated exceptional performance in binary classification tasks, exhibiting high accuracies on both the training and test sets. However, it is important to note that the model's accuracy decreased when employed in multiclass classification, emphasizing the need for further investigation and refinement to handle more diverse classification scenarios.

Keywords— Diabetic retinopathy, Swin transformer, muti- Wavelet, APTOS 2019, Vision transformers.

I. INTRODUCTION

Diabetic retinopathy is one of the diseases that affect the eye and the cause is due to diabetes, and thus it is a major cause of vision loss, as about 34.6%) of patients with diabetes suffer from (DR) and this is found in Asia, Europe and the United States [1]. With the increase in the number of people with diabetes all over the world, the number of people with diabetes was estimated from 108 million in 1980 to an estimated 425 million in 2017, and to approximately 629 million in 2045. These statistics have become a global epidemic [2]. People with diabetes usually remain asymptomatic until the advanced stages of DR are uncontrollable and therefore early examination is necessary to start treatment in time [3].

Numerous obstacles are associated with eye care, encompassing the exorbitant expenses of healthcare and the limited availability of eye specialists in low- and middle-income countries. These regions lack adequate healthcare infrastructure to effectively detect and address eye-related ailments. Early diagnosis is imperative to effectively manage the disease and avert vision impairment [4].Vision transformers (ViTs) were first proposed for the machine translation task in the Natural Language Processing NLP domain. The transformer-based

methods have achieved state-of-the-art performance in various tasks [5]. The drawback of ViT is that it requires pre-training on its large dataset [6]. Transformers have been widely used in numerous vision problems, especially for visual recognition and detection [7].

The pioneering work of Swin Transformer [8] has two outstanding contributions that distinguish it from other works:

(i) Presents a hierarchical feature representation scheme that demonstrates impressive performances with linear computational complexity. These hierarchical features can make Swin Transformer suitable as a general backbone for kinds of computer vision tasks.

(ii) Swin Transformer proposes a key design where shifted windows are equipped between consecutive attention layers, which can enhance modeling power while performing a computation-efficient strategy [8].

This paper is structured into different sections. Section II provides an overview of related works. In Section III, the proposed framework is explained in detail, covering the Wavelet and multi-Wavelet transform with Swin transformer. The experimental results are presented and discussed in Section IV. Finally, the conclusions are summarized in Section V.

II. RELATED WORKS

This section reviews the relevant literature and provides fundamental knowledge for the proposed method.

Gu. Yeonghyeon et al [9] introduced a model called STHarDNet, which effectively combines Swin transformer blocks with a lightweight U-Net type architecture. The STHarDNet model utilizes an encoder-decoder structure based on HarDNet blocks and demonstrates remarkable performance in segmenting stroke MRI scan images. The incorporation of the first layer Swin adapter enables the extraction of hierarchical features. STHarDNet exhibits a CNN-like nature, efficiently completing the task while also addressing the limitations of specific regions. Due to its balanced trade-off between performance and speed, the proposed STHarDNet model has been recognized as the optimal choice.

Liao. Zhihao .el.at [10] introduced the Swin-PANet model, which incorporates a window-based self-attention mechanism utilizing the Swin switch in an intermediate supervision network. By leveraging the advantages of this switch, the proposed Swin PANet was applied in Computer Aided Diagnosis (CAD) for melanoma diagnosis, with the aim of enhancing segmentation accuracy. The model demonstrated superior performance compared to recent models; however, it still encounters certain limitations in the context of transfer learning.

L. Jingyun et al [11] were proposed a robust baseline model called SwinIR for image restoration, leveraging the power of the Swin Transformer. SwinIR is composed of three main components: shallow feature extraction, deep feature extraction, and human resource reconstruction units. The model, SwinIR, exhibited exceptional performance across various image recovery tasks and six different settings, establishing its competence in all aspects of image restoration.

In their study,H. Siyuan et al [12] introduced a novel two-stream Swin transformer network (TSTNet) designed specifically for Remote Sensing (RS) image classification. Each stream of the network utilized the Swin transformer as its backbone, which yielded impressive performance. Through experiments conducted on three challenging and widely-used datasets, it was demonstrated that TSTNet outperformed other state-of-the-art models in terms of classification accuracy. These findings highlight the effectiveness of TSTNet in RS image classification tasks.

In their research, A. Hatamizadeh et al [13] introduced the Swin UNITRansformers (Swin UNITR) model for the semantic stratification of 3D brain tumors. The input data was represented as a 1D sequence of embeddings, which served as the input to the Swin transformer. The model exhibited exceptional performance during the validation and testing phases, outperforming other approaches and achieving the best results. The Swin UNITR model demonstrates its effectiveness in accurately classifying and stratifying 3D brain tumors based on their semantic characteristics.

III. METHOD

In this section, a detailed introduction of the proposed method will be provided, focusing on the method's components depicted in Figure 1. Firstly, Section 3.1 will present an overview of the transform utilized in the proposed approach, encompassing the multi-Wavelet transform. Subsequently, Section 3.2 will delve into the specifics of the Swin Transformer block.

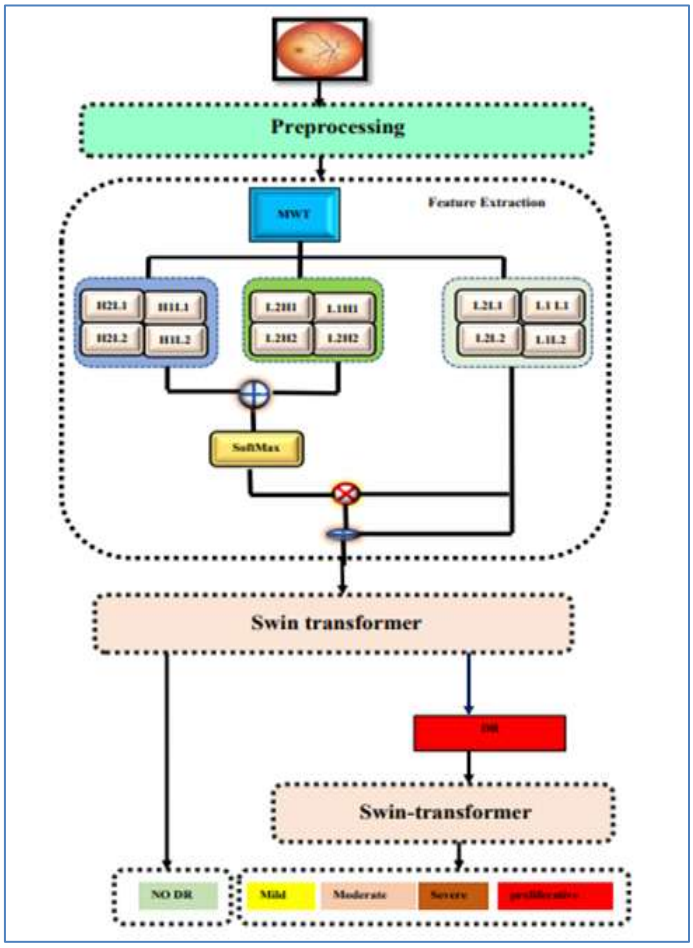


Fig. 1. The proposed method.

A. Preprocessing

In order to prepare the model for training, a series of preprocessing procedures are implemented on the input images.

These procedures encompass applying data augmentations, resizing the images, and auto cropping them.

B. Feature Extraction

In the feature extraction stage, the incorporation of multiwavelet transformation was implemented. Waves prove to be highly valuable tools in signal processing applications, as they fulfill various functions such as noise reduction and image compression [14],[15],[16]. Traditionally, only scalar waves derived from a single measurement function were recognized and utilized. However, more recently, the identification of multiple unit measurement functions has resulted in the emergence of multiple waves [17]. These multiple waves possess several advantages over scalar waves, including symmetry, fading moments, orthogonality, and short support [18]. When comparing these waves with scalar waves, the latter lack all of these properties simultaneously.

Also, the multi-wave system can provide perfect reconstruction and at the same time maintain the orthogonality as well as the arrangement, have high approximation and good performance through linear length symmetry [19]. With all these advantages that multi-wave conversion possesses, its performance exceeds that of scalar wavelets in image processing applications [20].

Multiple waves in which each channel has an input of a vector value and the outputs also have a vector value. One of the important differences between it and scalar waves is that the input signal whose value is scalar must be converted into a signal with an appropriate vector value. This conversion is called preprocessing [8].

Some reasons for potentially choosing multiwavelets can be summarized as follows:

i- The extra degrees of freedom inherent in multiwavelets can be used to reduce restrictions on the filter properties.

ii- The support length and the number of vanishing moments is directly linked to the filter length for scalar Wavelets.

iii- Multiple waves are able to have the best properties at the same time and this is the opposite of scalar waves.

iv- Multiple waves have good energy compression properties, as it relates the input signal to a small number of measurement parameters that contain most of the energy, and it is considered one of the desirable properties in image compression.

v- Previous literature has shown promising results in the application of multiwavelets to image compression.

vi- Finally, the issue of computational complexity is effective even though each branch of a multi-wave filter contains 2-input and 2-output filters and two channels. However, the filters in a symmetric multi-filter have the same kind of symmetry, the Figure 2 show Modify multiwavelet.

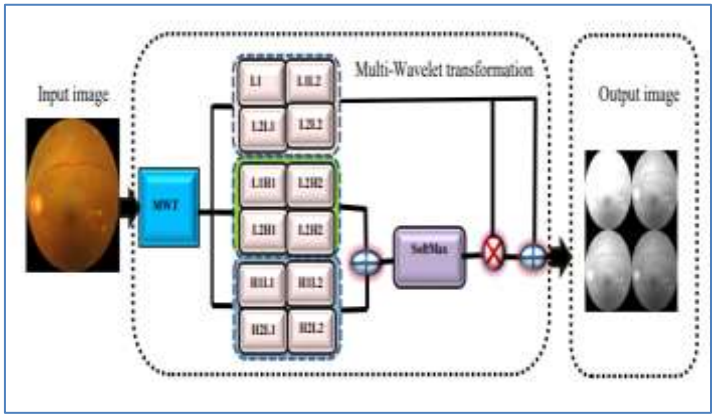


Fig 2. Modified multiwavelet.

In the subsequent feature extraction stage, the multi-Wavelet transform is employed. This transformation generates four copies of the image, namely LL, LH, HL, and HH, each containing four subparts. However, the HH subpart, which contains unwanted details, is disregarded. The horizontal and vertical details are combined, and the resulting output is then subjected to a SoftMax function. This output is further multiplied with the approximation (LL) copy and combined with the approximation (LL) copy, resulting in enhanced image details for clearer visualization.

C. The Shifted Window Transformer (ST)

After the pre-processing procedure, the Shifted Window Transformer or Swin Transformer (ST) was utilized to create hierarchical feature maps of the image. This involved combining patches into deeper layers. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows as well as allowing cross-window connection. It has a linear computational complexity proportional to the size of the input image due to self-attention processing occurring only within each local window. Thus, it generates feature maps with a single low resolution in comparison with earlier vision transformers. As a result, it can be used as a general-purpose backbone for image and texture classification and speech processing applications. There are many Challenges in Vision Applications using the Shifted Window Transformer. Transformer-based models in which all tokens are fixed in size, and thus are not suitable for computer vision applications, and since the pixels in the image have very high accuracy, and this is another difference between images and text segments, and images need a dense prediction at the pixel level, and to overcome all these challenges presented Microsoft's research team in Asia is the Swin converter, which is the backbone of computer vision applications and its computational complexity is linear.

Architecture

The basic design of ST is a transformation consisting of several successive layers of subjective attention which are connected to the last layer, thus increasing the modeling power. This technique is very effective in terms of computational complexity and latency in the real world.[21]. Figure 2 illustrates the architecture in its smallest configuration. Initially, the input RGB image is divided into non-overlapping patches using the ViT splitter.

The input image is passed through the patch partion layer, and the image is divided into 4×4 patches, thus creating patch codes (W/4 channel, H/4 and 4×4).These generated tokens pass through the linear modulation stage in the first stage, after which they are fed through two Swin switches in which the tokens are of size (W/4, H/4, C), the third and fourth stages consist of (patch merging) and (Swin block) Since the 2, 3 and 4 are phases, the symbols are (W/8, H/8, 2C), (W/16, H/16, 4C), and (W/32, H/32, 8C) respectively.

Swin Transformer can be implemented by replacing the standard multi-head self-attention module in a transformer block by using shifted windows keeping the other layers the same. Swin Transformer block consists of a shifted window based on multi-head self-attention module. Usually, it will be followed by a 2-layer multi-layer perceptron network with Gaussian Error Linear Unit nonlinearity in between. Next, Layer normalization is applied before each multi-head self-attention module and each multi-layer perceptron. Finally, a residual connection is applied after each of the above modules mentioned before.

ST is created by replacing (MSA) with (SW-MSA) with the rest of the layers left unchanged, so that each block of ST contains (SW-MSA) and (MLP) and (LN) nonlinear window-based self-attention module lacks connections through windows, which limits its modeling ability. Equations (2-5) provide the mathematical expression for W-MSA and SW-MSA as follows [22].

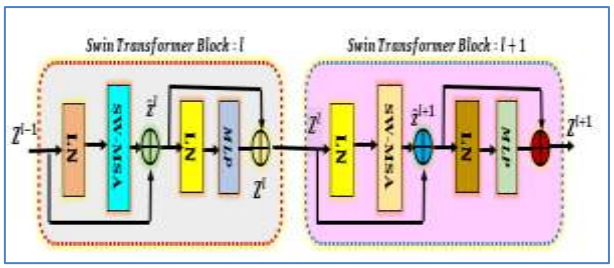


Fig 3. Swin transformer block.

$$\hat{z}^{l} = W - MSA\big(LN(Z^{l-1})\big) + Z^{l-1} \quad (2)$$

$$z^{l} = MLP\big(LN(Z^{l-1})\big) + \hat{z}^{l} \quad (3)$$

$$\hat{z}^{l+1} = SW - MSA\big(LN(z^{l})\big) + z^{l} \quad (4)$$

Where (z ^ l) is the item in the current block, (zl – 1) is the item in the previous block, (LN) is layer-norm, (MLP) is multi-layer perceptron, (W1-1MSA) is window self-attention, and (SW1-1MSA) is shift window self-attention [13].

IV. RESULT

The experimental data used in this study was introduced, and the performance of the proposed approach was evaluated.

A. Datasets

The results obtained from the APTOS 2019 Kaggle benchmark dataset are presented. The dataset comprises retina images captured using fundus imaging, with a wide range of imaging conditions. The challenge involved detecting blindness based on these images. The dataset has been carefully categorized into five classes (0 to 4) by domain specialists. Each class represents a different severity level of DR: "0" indicates no DR, "1" represents mild, "2" stands for moderate, "3" signifies severe, and "4" indicates proliferative DR [23], [24].

B. Swin transformer implementation

Firstly, the performances of various transforms, such as the multi-Wavelet transform and Swin Transformer, are compared in the following sections.

i. Swin transformer with multi-Wavelet for binary-class:

The use of Swin-T with multi-wave for feature extraction is a promising approach for image classification tasks where in this case take the batch size=64 based on the train time (3111.27s) is less other batch size with the same accuracy in the test and best validation accuracy of 0.9891 with 100 Epoch show in Figure 4.

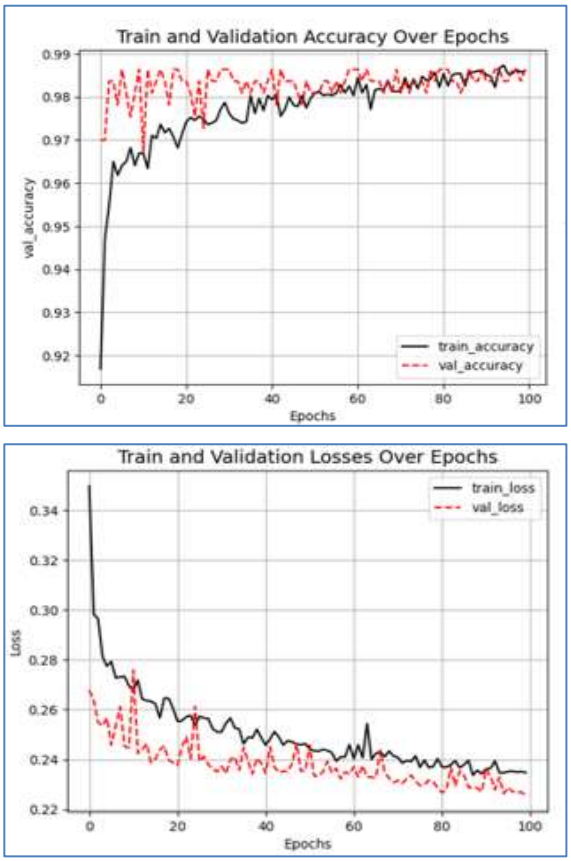


Fig 4: Training and validation over epoch for APTOS 2019 dataset, (a)loss, (b) accuracy, (epochs=100), multi-Wavelet transform to binary class.

The results show that the Swin-T model with multi- Wavelet transformation for feature extraction achieved an overall test accuracy of 97% and an average F1 score of 0.9873 for binary classification of No-DR and DR cases. The model performed better in detecting the No-DR cases, achieving a sensitivity of 0.97 and specificity of 0.9867, while for DR cases, it achieved a sensitivity of 0.9798 and specificity of 0.96 shows in Figure5 and Table1 shows the testing the swin-T with multi- Wavelet transformation for binary class.

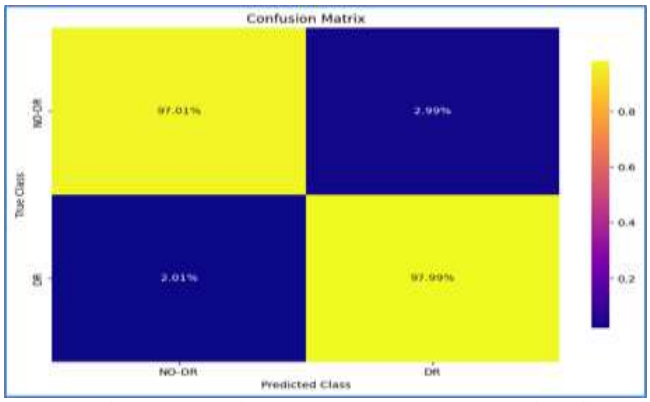


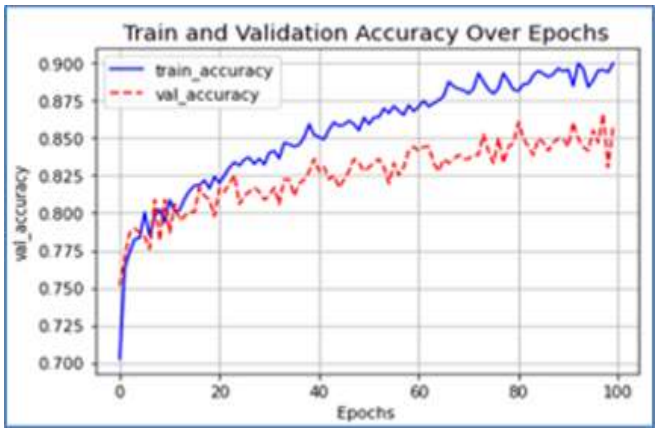
Fig 5: confusion matrix of Swin-T with multi-Wavelet transform for binary class.

TABLE 1: TESTING FOR SWIN-T MULTI-WAVELET FOR BINARY CLASS

Class	Test Accuracy	Test loss	Sensitivity	specificity	F1 Score
No-DR	98%	0.0328	0.9798	0.9700	0.9867
DR	96%		0.9700	0.9798	0.9748
Average	97%		0.9520	0.9520	0.9873

ii. Swin transformer with multi-Wavelet for multi-class:

The results show that using Swin-T transformer with multi-wavelet transformation for feature extraction leads to relatively high accuracy rates ranging from 81% to 82%. as the highest accuracy was achieved with a batch size of 16.
The Figure 6 indicate the fluctuations in accuracy and loss values of the training and validation sets during a 100-epoch.
By using Swin-T transformer with multi-wavelet transformation for feature extraction, the DR multi-classification achieved an average testing accuracy of 82%, with a testing loss of 0.4626, sensitivity of 0.8017, and specificity of 0.9022.



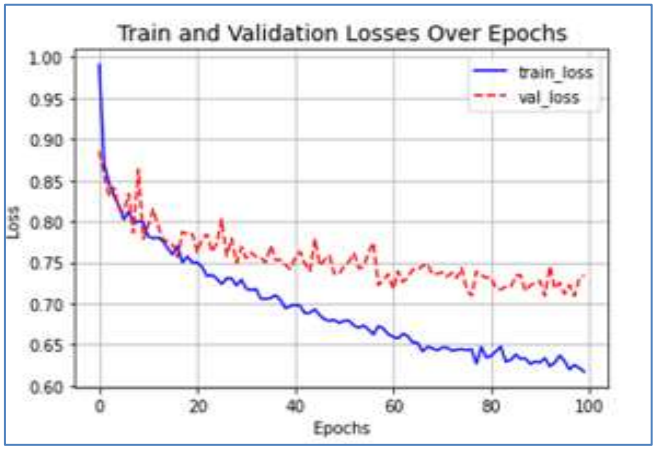


Fig 8: Training and validation over epoch for APTOS 2019 dataset, (a) loss, (b) accuracy, (epochs=100), multi-Wavelet transform to multi-class.

The confusion matrix shows in Figure 9 the performance of the Swin-T transformer with multi-wavelet transformation for feature extraction model on the DR multi-classification task. Table 2 shows the testing the swin-T with multi- Wavelet transformation for multi class.

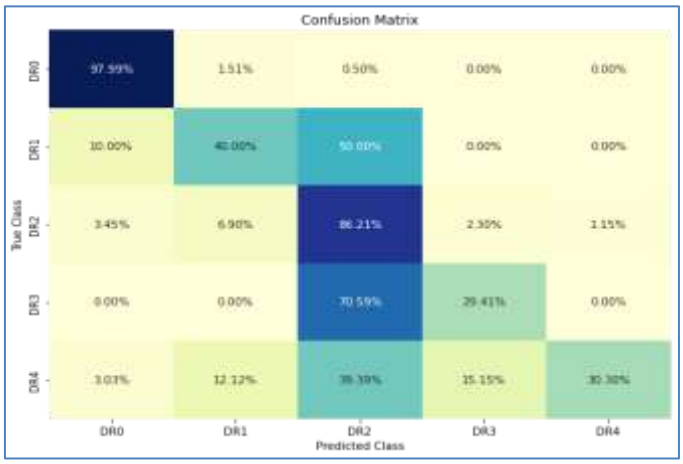


Fig 9: confusion matrix of Swin-T with multi-Wavelet transform for multi-class.

TABLE 2: TESTING FOR SWIN-T MULTI-WAVELET FOR MULTI- CLASS

Class	Test Accuracy	Test loss	Sensitivity	specificity	F1 Score
DR0	97%	0.1233	0.9798	0.9580	0.9829
DR1	41%		0.4	0.9613	0.4878
DR2	85%		0.8620	0.8530	0.8904
DR3	33%		0.2941	0.9799	0.4000
DR4	33%		0.3030	0.9969	0.4000
Average	82%		0.4626	0.8017	0.9022

V. CONCULSION

In this work, the multi-Wavelet transform with Swin-T is presented for DR classification. The performance of Swin-T with multi-Wavelet was evaluated through experiments. The Swin transform was applied to the APTOS 2019 Kaggle dataset in this study. In conclusion, the proposed that used multi-Wavelet transform with Swin-T achieves better performance for DR binary classification, where the accuracy 0.9778% for training and the loss is 0.2537% and the accuracy is 0.9754for validation and the best accuracy is 0.9809%, and the test accuracy is 97% for binary class while used the multi- Wavelet to classification DR to multi class is the loss is 0.5901% and the accuracy is 0.9160% for training and the loss is 0.8250% and the accuracy is 0.8242% for validation and the test accuracy is 82%. The integration of the Swin Transformer with multi-wavelet analysis presents a promising direction for the classification of diabetic retinopathy. By exploring the aforementioned future research directions, such as augmentation techniques, fusion strategies, transfer learning, handling class imbalance, interpretability, and real-world evaluations, we can enhance the accuracy, robustness, and clinical relevance of the proposed method. These advancements will contribute to the development of effective tools for automated DR diagnosis and aid in the early detection and prevention of vision loss in diabetic patients.

REFERENCES

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] A. Grzybowski, P. Brona, G. Lim, and P. Ruamviboonsuk, "Artificial intelligence for diabetic retinopathy screening: a review," Springer Nat., 2020.

[3] S. Natarajan, A. Jain, R. Krishnan, and A. Rogye, "Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone," JAMAOphthalmology, 2019.

[4] M. Wintergerst, D. Mishra, L. Hartmann, and F. Holz, "Diabetic Retinopathy Screening Using Smartphone-Based Fundus Imaging in India," Am. Acad. Ophthalmol., vol. 127, no. 0161-6420/20, 2020.

[5] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," arXiv Prepr. arXiv2105.05537, 2021, [Online]. Available: http://arxiv.org/abs/2109.04335.

[6] W. Wang, E. Xie, X. Li, and D.-P. Fan, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions arXiv:2102.12122v2," arXiv:2102.12122v2 [cs.CV], 2021.

[7] H. Song, D. Sun, S. Chun, and V. Jampani, "An Extendable, Efficient and Effective Transformer-based Object Detector," arXiv:2204.07962v1, 2022.

[8] L. Wang, R. Li, C. Duan, C. Zhang, and X. Meng, "A Novel Transformer based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images," Geosci. Remote Sens. Lett., 2021.

[9] Y. Gu, Z. Piao, and S. J. Yoo, "STHarDNet: Swin Transformer with HarDNet for MRI Segmentation," Appl. Sci., 2022.

[10] Z. Liao, N. Fan, and K. Xu, "Swin Transformer Assisted Prior Attention Network for Medical Image Segmentation," Appl. Sci., 2022.

[11] J. Liang, J. Cao, G. Sun, and K. Zhang, "SwinIR: Image Restoration Using Swin Transformer," arXiv:2108.10257v1, 2021.

[12] S. Hao, B. Wu, K. Zhao, and Y. Ye, "Two-Stream Swin Transformer with Differentiable Sobel Operator for Remote Sensing Image Classificatio," Remote Sens., 2022.

[13] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images." 2022, [Online]. Available: http://arxiv.org/abs/2201.01266.

[14] A. HM Al-Helali, H. Ali, B. Al-Dulaimi, D. Alzubaydi, and W. Mahmoud, A, "Slantlet transform for multispectral image fusion," J. Comput. Sci., vol. 5, no. 4, p. PP. 263-267, 2009.

[15] A. Al-Helali, W. A. Mahmoud, and H. Ali, "A Fast personal palm print authentication Based on 3d-multi Wavelet Transformation," Transnatl. J. Sci. Technol., vol. 2, no. 8, 2012.

[16] H. Al-Taai, W. A. Mahmoud, and M. Abdulwahab, "New fast method for computing multiWavelet coefficients from 1D up to 3D," Proc. 1st Int. Conf. Digit. Comm. Comp. App., Jordan, no. PP. 412-422, 2007.

[17] A. H. Kattoush and W. Ameen Mahmoud Al-Jawher, "A radon-multiWavelet based OFDM system design and simulation under different channel conditions" Journal of Wireless personal communications," J. Wirel. Pers. Commun., vol. 71, 2017

[18] W. A. Mahmoud Al-Jawher and T. Abbas, "Feature combination and mapping using multiWavelet Transform," IASJ, AL-Rafidain, 2005.

[19] W. A. Mahmoud Al-Jawher, "A Smart Single Matrix Realization of Fast Walidlet Transform," Int. J. Res. Rev., vol. 2, no. 2, pp. 144–150, 2011.

[20] W. A. Mahmoud, A. S. Hadi, and T. M. Jawad, "Development of a 2-D Wavelet Transform based on Kronecker Product," Al-Nahrain J. Sci., vol. 15, no. 4, pp. 208–213, 2012.

[21] J. Ahn, J. Hong, J. Ju, and H. Jung, "Rethinking Query, Key, and Value Embedding in Vision Transformer under Tiny Model Constraints," arXiv:2111.10017v1, 2021, [Online]. Available: http://arxiv.org/abs/2111.10017.

[22] E. In, "Lmsa: Low-Relation Mutil-Head Self- Attention Mechanism in Visual Transformer," pp. 1–11, 2022.

[23] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep Learning Approach to Diabetic Retinopathy Detection Borys," arXiv:2003.02261v1, 2020.

[24] J. D. Bodapati et al., "Blended multi-modal deep convnet features for diabetic retinopathy severity prediction," Electron., vol. 9, no. 6, 2020, doi: 10.3390/electronics9060914