STMFNet: Spatial Texture Multi-Scale Feature Fusion Attention Network for Diabetic Retinopathy Classification

MD Ilias Bappi, Md Monir Ahammod Bin Atique, Kyungbaek Kim
Department of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
i_bappi@jnu.ac.kr, monir024@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

***Abstract*—Diabetic Retinopathy (DR) is a major cause of vision loss and blindness, particularly among diabetic patients. Effective and timely treatment of DR relies on precise and automated detection systems that can assess disease severity from retinal fundus images. Traditional clinical approaches are often time-consuming, and earlier texture attention models may struggle to accurately detect subtle features, such as microaneurysms or abnormal blood vessel patterns, which are crucial for early diagnosis. To address this, we proposed the STMFNet model, a hierarchical framework designed for classifying 11 stages of DR severity. The model combines two primary mechanisms: The Texture Spatial Attention Network, which focuses on identifying critical texture features related to DR while minimizing irrelevant background information through attention gating, and the EfficientNet backbone with Multi-Scale Feature Fusion, which captures a wide range of image patterns. By extracting and fusing features from different layers of EfficientNet-V1 B0, the model effectively learns both low (e.g., edges, blobs) and high (e.g., objects, patterns) level representations. These features are further refined through spatial multi-scale attention, and the final classification into 11 DR stages is achieved through a Fully Connected Network (FCN) and SoftMax. Our experimental results show that STMFNet significantly outperforms existing state-of-the-art models on the publicly available Kaggle fundus dataset, demonstrating its potential for reliable DR diagnosis in clinical settings.**

***Index Terms*—Diabetic retinopathy, Texture and Spatial attention, Classification, Multi-scale feature fusion, Detection.**

I. Introduction

Early diagnosis is crucial in healthcare, especially for managing conditions like diabetes, a chronic disease affecting millions worldwide due to insufficient insulin regulation. According to the International Diabetes Federation, over 425 million adults globally have diabetes [1]. Left untreated, it can lead to complications such as diabetic retinopathy (DR), which affects the retina and can cause vision loss. DR damages the blood vessels in the retina, impairing vision and potentially leading to blindness. Early detection is vital to prevent severe outcomes [2].

Automated methods for DR detection not only facilitate early diagnosis but also bring about cost savings, improved efficiency, and enhanced accuracy in clinical surroundings. The role of ComputerAided Diagnosis (CAD) systems has proven to be especially beneficial in medical image processing, as demonstrated in numerous studies [3]. More recently, Deep learning (DL) techniques, particularly convolutional neural networks (CNNs), have revolutionized the field of medical imaging. K. Xu et al [4]. Li et al [5] introduced a CNN-based approach for distinguishing between normal and DR images. utilized a Deep CNN (DCNN) model incorporating fractional max-pooling to improve the extraction of discriminative features, followed by classification using a support vector machine (SVM). R. Pires et al [6]. developed a 16-layer CNN model to classify DR images into referable and non-referable categories, using dropout and L2 regularization to mitigate overfitting.

Despite advancements, current DL models struggle to capture both local textures and global context for accurate DR level recognition. To address this, we propose incorporating texture and spatial attention mechanisms with multi-level CNNs, decomposing the representation space into style and content features. This process involves two key components: (1) the Texture Spatial Attention Network, which highlights important texture features related to DR while suppressing irrelevant background information, and (2) the EfficientNet-V1 B0 backbone with Multi-Scale Feature Fusion, which captures both low- and high-level patterns from retinal images. By extracting features from multiple layers, the model can learn fine-grained representations, which are further refined through spatial multi-scale attention before final classification into 11 DR categories using a FCN and SoftMax.

The key contributions of this paper are summarized as follows:

- The development of an innovative spatial textural attention module in the **STMFNet** model enhances its ability to emphasize salient microaneurysms and abnormal features, enabling the model to automatically classify 11 DR stages, including healthy cases.
- A Multi-Scale Feature Fusion process within the EfficientNet-V1 B0 backbone, enhancing both low and high level pixel information from fundus images.
- Achievement of state-of-the-art results on a public DR classification dataset.

II. Related Work

A. CNN based approaches of DR classification

DL methods, particularly CNNs, have achieved significant progress in recent years for the classification of DR. To enhance CNN performance in DR screening DL techniques, especially CNNs, have brought significant advancements in medical imaging. Xu et al. [4] and Li et al. [5] introduced CNN-based approaches for distinguishing between normal and DR-affected images. Li's work, in particular, utilized a deep CNN (DCNN) architecture combined with fractional max-pooling to enhance the extraction of discriminative features, followed by classification using SVM. Pires et al [6]. developed a 16-layer CNN model to classify DR images into referable and non-referable categories, incorporating dropout and L2 regularization to prevent overfitting. However, these methods primarily rely on conventional feature extraction from fundus images, focusing on general patterns without fully capturing detailed textural features. In contrast, our multi-scale feature fusion combination block with the **EfficientNet-V1 B0** backbone is chosen for its optimal balance between performance and computational efficiency, effectively addressing these limitations.

B. Attention based approaches of DR classification

To make up for the deficiency of the CNN-based method in capturing long-distance and ground details some attention methods (Luo, Xiaoling, et al [7].; Alahmadi, Mohammad D [8].; Wang, Xiaofei, et al [9].) based on attention have been proposed. Luo, Xiaoling, et al [7]. tackled the limitations of single-view methods in DR detection by introducing a multi-view model integrated with a Cross-Interaction Self-Attention Module. By leveraging multiple retinal views, their method captures cross-view pathological relationships, offering a more thorough retinal analysis. Meanwhile, another approach Alahmadi, et al [8]. improved DR classification by introducing a recalibration mechanism that prioritizes critical areas of retinal images. This approach separates features into texture and semantic components and uses texture attention to enhance classification accuracy. Another study Wang, Xiaofei, et al [9]. proposed a deep multi-task learning framework that focuses on DR grading using low-resolution fundus images. This method combines image super-resolution and lesion segmentation. Despite its innovations, it leaves room for improvement in effectively capturing fine details through attention mechanisms. Nevertheless, there is a reduced footprint and a more precise identification of the ground textures. By learning to highlight important regions in retinal pictures, particularly the posterior pole and peripheral fundus, our spatial textural attention mechanism fills this gap and enables more precise identification of crucial diagnostic regions.

III. Method

In clinical DR diagnosis, ophthalmologists examine fundus images to assess lesions' location, size, and number, determining disease stages. In contrast, our model detects 10 conditions, including a healthy retina, by extracting texture features from retinal images. Fundus images are individually fed into the EfficientNet and the Spatial Texture Attention module to extract pattern and texture-based features. In this context, EfficientNet serves as the backbone, learning complex representations from Blocks 2 to 7, with the features being fused through integration. The texture attention module focuses on fine-grained details. After concatenating the outputs from both modules, a spatial multi-scale attention mechanism generates an attention map to highlight important regions. These features are passed through a FCN with dense and dropout layers, followed by SoftMax for final DR classification. An overview of the proposed model is shown in "Fig. 1".

A. Feature Extraction

1) *Spatial Texture Attention*

In order to emphasize informative areas in each texture frame from retinal image, we develop a spatial texture attention module to assign higher weights to crucial areas, while assigning lower weights to areas containing less information. The architecture of the texture attention module is illustrated in "Fig. 2". Motivated by [10], [11], we employ both max and global average pooling, alongside a 3×3 conv instead of a 1×1 conv, to enhance cross-channel interaction and capture the spatial structure of the textural feature F obtained from the CNN module. This combination helps to create robust spatial context descriptors along the channel axis, refining the feature representation [12]. The global average-pooling is used to effectively learn tactile information (with output FS_{max}), whereas max-pooling is used to preserve prominent features (with output FS_{avg}). FS_{max} and FS_{avg} are then concatenated and convolved with a 3×3 kernel, followed by activation with a sigmoid function to produce a 2D spatial texture attention map $A_S(F)$:

$$\begin{aligned} A_S(F) &= \sigma\left(f^{3\times3}\left([\text{MaxPool}(\mathbf{F}); \text{GAvgPool}(\mathbf{F})]\right)\right) \\ &= \sigma\left(f^{3\times3}\left(\left[\mathbf{F}^S_{\max}; \mathbf{F}^S_{\text{avg}}\right]\right)\right) \end{aligned} \tag{1}$$

where σ denotes the sigmoid function. Then, we get the output feature map $F^S = A_S(F) \otimes F$ from the spatial attention module, where $\otimes$ refers to element-wise multiplication.

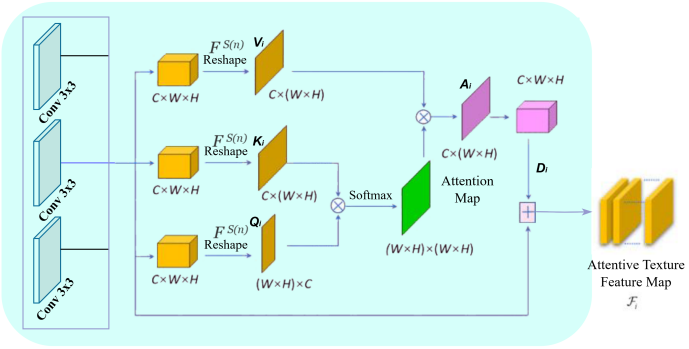

Fig. 2. Architecture of Spatial Textural Attention.

After obtaining the extracted features from the spatial attention module of each texture frame, we concatenate all

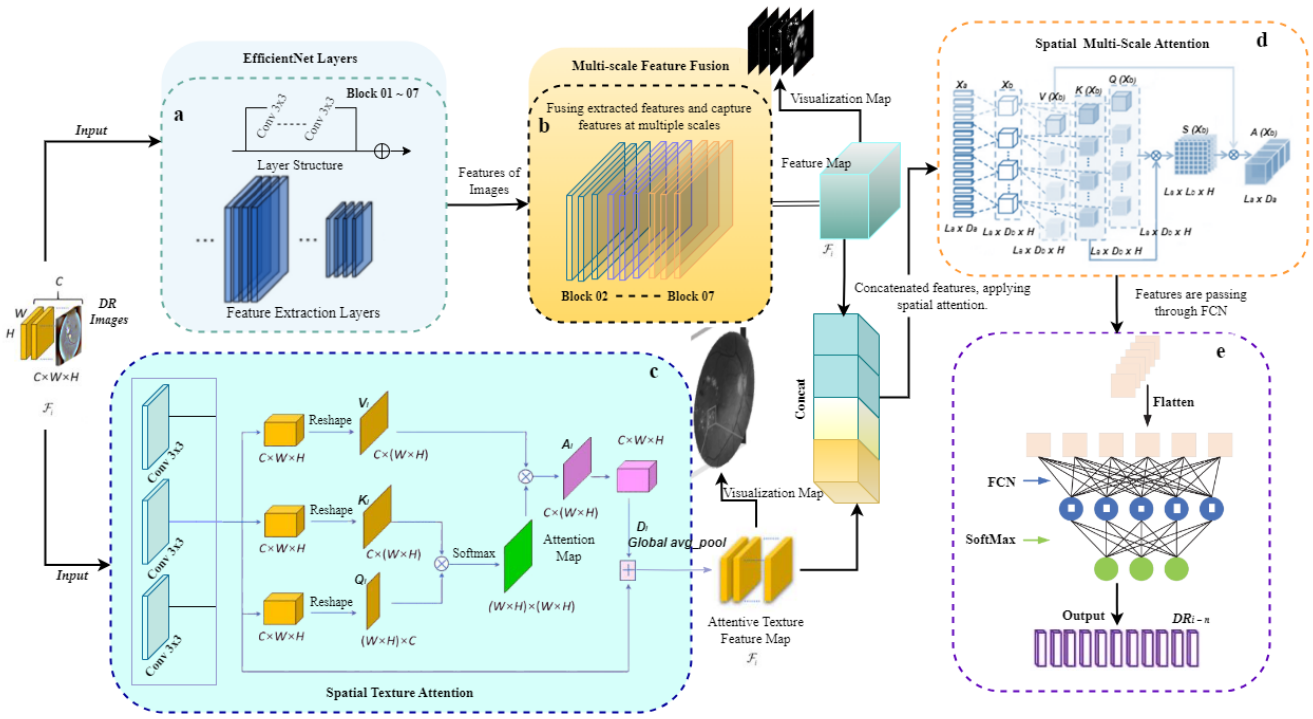

Fig. 1. Framework of the proposed STMFNet DR multiple stages classification.

the features together to achieve a sequence of texture features, represented by $F^S(n)$. To model the long-distance dependency in the tactile sequence, we developed a textural attention module on top of the spatial attention layer. As illustrated in "Fig. 2", this module estimates the salience and relevance of all regions in a tactile sequence over time, regardless of distance. $F^S(n)$ is first converted into two feature spaces, $q(F^S(n))$ and $k(F^S(n))$, using two sets of 3×3 convolutions, where $q(F^S(n)) = W_q F^S(n)$ and $k(F^S(n)) = W_k F^S(n)$ (W_q and W_k are trainable weight matrices). Subsequently, we reshape both $q(F^S(n))$ and $k(F^S(n)) \in \mathbb{R}^{m \times c}$, where $m = n \times h \times w$, to calculate the attention map of any pairs of regions through the time dimension. The attention map $A_T(F^S(n))$ is given as follows:

$$\mathbf{A}_T\left(\mathbf{F}^S(n)\right)_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{m} \exp(s_{ij})}, \tag{2}$$

where $s_{ij} = q(F^S(n)_i) k(F^S(n)_j)^\top$. $A_T(F^S(n))_{j,i}$ demonstrates how much $F^S(n)_i$ correlates with $F^S(n)_j$. The output feature map of the textural attention is $F_T = (F_T^1, F_T^2, \ldots, F_T^j, \ldots, F_T^m)$, where

$$\mathbf{F}_j^T = \sum_{i=1}^{m} \mathbf{A}_T\left(\mathbf{F}^{S(n)}\right)_{j,i} v\left(\mathbf{F}_i^{S(n)}\right) + \mathbf{F}_j^{S(n)}, \tag{3}$$

$v(F^S(n)) = W_v F^S(n)$ (where W_u is a learnable matrix) and $F^S(n)_j$ is added back to retain more information.

We incorporate concatenation and spatial attention, to allow the model to synthesize information jointly from different representation feature spaces [13]. Ultimately, the learned representations are aggregated and passed to the spatial multi-scale attention phase. The outcomes from this phase are then fed into a fully connected layer to perform a classification task that calculates the probability of the predicted label DRi. Microaneurysms visualization comparison of different recent attention with ours is indicated in "Fig. 3"

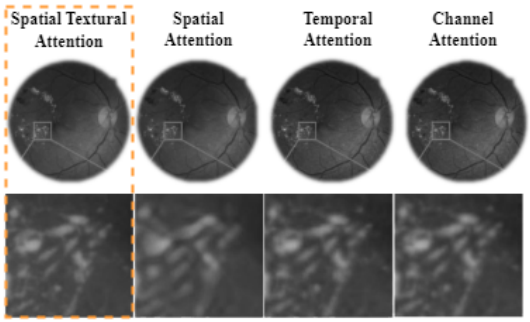

Fig. 3. Comparison of feature maps from various attention mechanisms with the proposed method.

2) *Extraction with EfficientNet*

EfficientNet employs a compound scaling method that balances model depth, width, and input resolution, enabling effective feature extraction from retinal fundus images [14]. The architecture utilizes serves as the primary feature extractor for 224×224 fundus images, with its initial layers capturing low-level features like edges and textures, essential

for detecting early signs of DR. These layers work similarly to traditional Conv1 and Conv2 but are optimized by its compound scaling, balancing depth, width, and resolution. Blocks 2 and 3 capture more complex features as the image advances through deeper layers, such as small blood vessels and microaneurysms for comprehensive analysis [15]. The depth-wise separable convolutions allow it to preserve low- and high-level features efficiently. Blocks 4 and 5 extract mid-level patterns, focusing on medium-sized lesions like exudates, while Blocks 6 and 7 capture high-level semantic information, such as lesion distribution and morphology. This enhances feature extraction through depth-wise convolutions and Squeeze-and-Excitation (SE) modules, which focus on important regions. The extracted features are then combined in a multi-scale feature fusion network, integrating outputs from various layers for a comprehensive analysis, improving classification accuracy. As represented in **a** under "Fig. 1".

B. Multi-Scale Feature Fusion

In the multi-scale feature fusion process, the retinal fundus image is passed through the EfficientNet backbone, extracting features from Blocks 2 to 7. The feature maps are upsampled and refined using a deconvolution operation to ensure consistent resolution. A 3×3 convolution is applied to remove redundancies and mitigate aliasing effects. Then each feature from different blocks is concatenated, forming a comprehensive feature vector that integrates both semantic and local details, essential for identifying microstructural anomalies in the retina. To further refine the fused features, a spatial multi-scale attention mechanism is applied, which highlights critical regions, enhancing the model's sensitivity to subtle DR related features. The resulting feature vector, enriched with multi-scale information, is passed through a FCN for classification, ultimately improving the model's robustness and accuracy across various DR stages. As shown in **b** under "Fig. 1."

C. Spatial Multi Scale Attention

At this point inspired by [7], as shown in"Fig. 4" after receiving the vector features from the unification stage and adapting the mechanism of multi-head scale attention the input patches $\mathbf{X}_a$ are initially and randomly divided into multiple heads $\mathbf{X}_b \in \mathbb{R}^{L_a \times D_b \times H}$, where $\mathbf{X}_b = [\mathbf{X}_{b1}, \mathbf{X}_{b2}, \ldots, \mathbf{X}_{bm}, \ldots, \mathbf{X}_{bH}]$ to learn multiple local and global disease features. The number of scale heads H can be regarded as the number of feature groups:

$$H = \frac{D_a}{D_b}. \tag{4}$$

The three generators $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ are employed to convert $\mathbf{X}_b$ to query $Q(\mathbf{X}_b)$, key $K(\mathbf{X}_b)$, and value $V(\mathbf{X}_b)$, respectively. We consider the operations of the three generators, which are defined as:

$$Q(\mathbf{X}_b) = \mathbf{X}_b \cdot \mathbf{w}_Q, \tag{5}$$
$$K(\mathbf{X}_b) = \mathbf{X}_b \cdot \mathbf{w}_K, \tag{6}$$
$$V(\mathbf{X}_b) = \mathbf{X}_b \cdot \mathbf{w}_V, \tag{7}$$

where $\mathbf{w}_Q, \mathbf{w}_K$, and $\mathbf{w}_V$ are learnable parameters. Specifically, the vector $Q(\mathbf{X}_b)$ can be regarded as a feature selector for the channels of the matrix $K(\mathbf{X}_b)$.

In the process of spatial-attention calculation, we define the pairwise function of $Q(\mathbf{X}_b)$ and $K(\mathbf{X}_b)$ as a matrix multiplication:

$$S(\mathbf{X}_b) = Q(\mathbf{X}_b)K(\mathbf{X}_b)^T, \tag{8}$$

where the T operation means matrix transpose. Moreover, the generated $S(\mathbf{X}_b) \in \mathbb{R}^{L_a \times L_b \times H}$ also plays the role of feature selector for the value $V(\mathbf{X}_b)$. Then, global attention can be defined as:

$$A(\mathbf{X}_b) = \text{softmax}(G(\mathbf{X}_b))V(\mathbf{X}_b), \quad A(\mathbf{X}_b) \in \mathbb{R}^{L_a \times D_b \times H}, \tag{9}$$

where the goal of the softmax function is to normalize $S(X_b)$. Next, the output $DR_{eit} \in \mathbb{R}^{L_a \times D_a}$ of the matrix can be roughly described as the splicing of attention maps of multi-scale feature groups:

$$DR_{eit} = \text{Linear}(\text{reshape}(A(X_b))). \tag{10}$$

Specifically, the Linear and reshape functions are designed to ensure that the output is concatenated from the group of the obtained attention maps and has dimension $L_a \times D_a$. Finally,

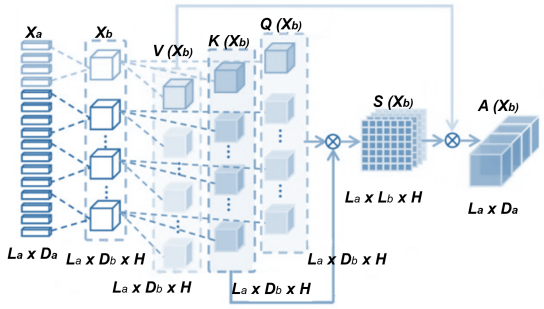

Fig. 4. The details of Spatial multi-scale Attention mechanism.

the attention-mapped features are distributed to the fully connected layers for further processing in DR classification.

D. Fully Connected Network

The FCN transforms the spatially attended feature map into a final feature vector for classification after paying spatial multi-scale attention to the extracted features. It consists of several dense layers, interspersed with dropout layers to mitigate overfitting, which allows the model to learn complex mappings from the input features to the classification outputs. The final layer of the FCN typically employs a SoftMax activation function, which normalizes the output into a probability distribution over multiple classes, enabling the classification of diabetic retinopathy into 10 categories of disease stages including a healthy class. As illustrated in **e** under "Fig. 1".

IV. Experiment

We conducted experiments on the Retinal Fundus Images dataset from Kaggle [16] for DR disease detection, which is currently the only publicly available large-scale dataset of fundus images for DR. The dataset contains 25,452 color retinal images classified into 11 categories, representing 10 diseases and healthy retinas. Ophthalmologists used these images to classify the DR stages of each subject following international standards. The training, test, and validation sets were distributed in an 80:20 ratio, and augmentation techniques such as zooming, scaling, padding, and background noise reduction were considered before feeding the data into the model. For the training setup, we used the PyTorch framework, with a GPU (NVIDIA GeForce RTX 3070), 64 GB RAM, and an Intel(R) i9-10900. Finally, for model evaluation and comparison, we adopted commonly agreed-upon evaluation metrics, including accuracy, precision, recall, and F1 score.

A. Result

Our proposed **STMFNet** model for DR classification performed remarkably well in its final training cycle, with a train and validation loss of 0.0184, 0.0065, the learning rate of 0.0001 after 100 epochs ("Fig. 5"). The robustness of the feature learning was demonstrated by the model's impressive adaptability on the validation dataset. The evaluation results highlight the model's strong predictive capabilities across all DR stages. As shown in Table I, the STMFNet achieved an accuracy of 97.72%, with precision, recall, and F1-score values of 99.72%, 99.72%, and 98.72%, respectively. The model's area under the ROC curve was 99.98%, indicating excellent performance in distinguishing between different DR phases, including healthy and diseased cases. These results underscore the effectiveness of the proposed multi-scale feature fusion including textural and spatial multi-scale attention mechanisms, which enhance the model's ability to capture critical features from retinal fundus images and lead to accurate classification.

TABLE I
RESULTS OF OUR PROPOSED **STMFNET** METHOD. QUANTITATIVE RESULTS OF ACCURACY, PRECISION, RECALL, AND F1 SCORE IN DR. THE RESULTS ARE HIGHLIGHTED IN BOLD. (UNIT: %)

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*ROC AUC*
STMFNet	**97.72**	**99.72**	**99.72**	**98.72**	**99.98**

We have also laid out the confusion matrix. As stated in "Fig. 6", the suggested mechanism can effectively classify the stages comprising healthy classes of DR images (with 97.72% confidence) from the disease classes. The model demonstrates strong confidence in accurately classifying non-healthy samples from the healthy class. However, its performance diminishes when distinguishing between Glaucoma and Moderate DR stages. This decline is primarily attributed to the high degree of feature similarity among diabetic stages that are closely related, making it particularly challenging for the deep learning model to differentiate between these subtle variations.

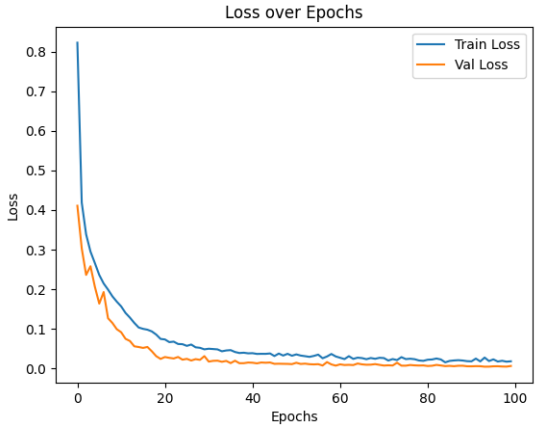

Fig. 5. Validation loss progression during model training.

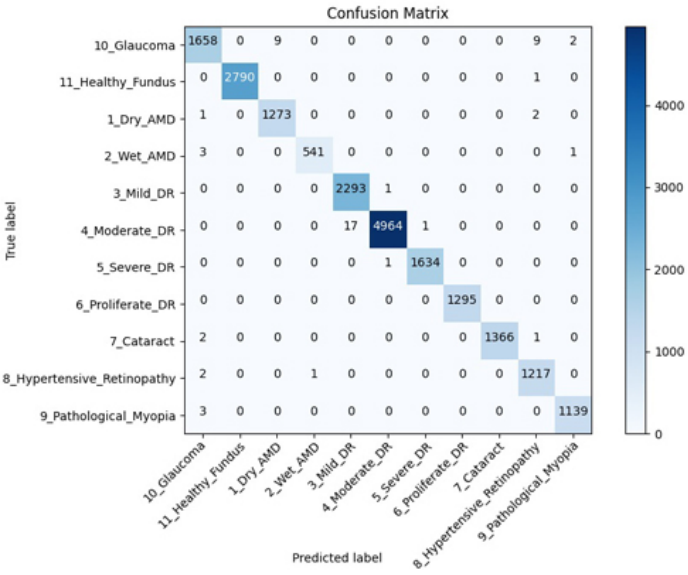

Fig. 6. Confusion matrix (CM) map by applying the proposed STMFNet on the dataset [16]. The CM shows how well the technique can distinguish between images of healthy and unhealthy retinas.

B. Comparison With Recent SOTA Classification Model

Several recent SOTA methods for DR classification have been adopted, which can be broadly categorized into CNN and Transformer models. These models were evaluated on the same dataset we used, highlighting the superior performance of our proposed **STMFNet** model. As shown in Table II, STMFNet achieved the highest overall accuracy (97.72%), significantly outperforming models like **U-Net** (80.69%) and **MVCINN** (78.69%). Furthermore, it delivered the best precision, recall, and F1-score (99.72%, 99.72%, and 98.72%, respectively), surpassing both **ViT** and **ResNet50**, which achieved high but comparatively lower scores. These results demonstrate STMFNet's effectiveness in leveraging multi-scale feature fusion with texture and spatial multi-scale attention, enabling it to extract critical DR-related features from retinal fundus images and establishing it as a leading model

in this domain. All baseline models have been pre-trained on the ImageNet dataset.

TABLE II
COMPARISON OF SOTA METHODS AND OUR PROPOSED **STMFNET** METHOD. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. (UNIT: %)

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*
U-Net [17]	80.69	73.77	80.69	76.01
MVCINN [7]	78.69	76.90	80.69	81.61
ViT [18]	94.33	94.50	94.33	94.14
ResNet50 [19]	91.04	91.05	91.04	91.03
InceptionV4 [20]	92.54	92.64	92.54	92.56
MobileNetV3 [21]	89.12	89.17	89.12	89.13
ConvNeXt-S [22]	93.66	93.67	93.66	90.99
STMFNet	**97.72**	**99.72**	**99.72**	**98.72**

V. CONCLUSION

In this paper, we introduced an innovative spatial-textural attention mechanism that effectively captures fine-grained textures from retinal fundus images for accurate classification of DR. Our method integrates a multi-scale feature fusion process using EfficientNet-B0, which captures low and high-level patterns from the images. Additionally, spatial multi-scale attention refines these features to enhance classification accuracy. The proposed **STMFNet** model achieved a remarkable 97.72% accuracy, outperforming other models. One of its key strengths is its ability to finely capture and represent subtle textures from each pixel of the fundus images, enabling enhanced feature extraction and classification. Our combination of convolutional network and attention mechanisms accurately identifies regions of interest in the fundus, surpassing many state-of-the-art approaches. In future work, we plan to incorporate self-supervised learning to improve performance on unlabeled images and implement continual learning strategies to enable our model to adapt to new clinical data in real-time, enhancing its practical applicability.

ACKNOWLEDGMENT

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2022-00156287, 34%). This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through the Agriculture and Food Convergence Technologies Program for Research Manpower development, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(project no. RS-2024-00397026, 33%). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437718, 33%) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] . Int. Diabetes Fed. Brussels Belgium, "Idf diabetes atlas, 7th ed.." https://diabetesatlas.org/atlas/seventh-edition/.

[2] R. R. Bourne, G. A. Stevens, R. A. White, J. L. Smith, S. R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, K. Naidoo, *et al.*, "Causes of vision loss worldwide, 1990–2010: a systematic analysis," *The lancet global health*, vol. 1, no. 6, pp. e339–e349, 2013.

[3] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019.

[4] K. Xu, D. Feng, and H. Mi, "Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image," *Molecules*, vol. 22, no. 12, p. 2054, 2017.

[5] Y.-H. Li, N.-N. Yeh, S.-J. Chen, and Y.-C. Chung, "Computer-assisted diagnosis for diabetic retinopathy based on fundus images using deep convolutional neural network," *Mobile Information Systems*, vol. 2019, no. 1, p. 6142839, 2019.

[6] R. Pires, S. Avila, J. Wainer, E. Valle, M. D. Abramoff, and A. Rocha, "A data-driven approach to referable diabetic retinopathy detection," *Artificial intelligence in medicine*, vol. 96, pp. 93–106, 2019.

[7] X. Luo, C. Liu, W. Wong, J. Wen, X. Jin, and Y. Xu, "Mvcinn: multi-view diabetic retinopathy detection using a deep cross-interaction neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 8993–9001, 2023.

[8] M. D. Alahmadi, "Texture attention network for diabetic retinopathy classification," *IEEE Access*, vol. 10, pp. 55522–55532, 2022.

[9] X. Wang, M. Xu, J. Zhang, L. Jiang, and L. Li, "Deep multi-task learning for diabetic retinopathy grading in fundus images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 2826–2834, 2021.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[11] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal attention model for tactile texture recognition," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9896–9902, IEEE, 2020.

[12] Z. Tao, T. Wei, and J. Li, "Wavelet multi-level attention capsule network for texture classification," *IEEE Signal Processing Letters*, vol. 28, pp. 1215–1219, 2021.

[13] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] K. S. S. Nithish, "Retinal fundus images." https://www.kaggle.com/datasets/kssanjaynithish03/retinal-fundus-images, 2021.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241, Springer, 2015.

[18] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[21] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[22] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.