A Vision Transformer-Based Convolutional Neural Network for the Automated Diagnosis of Eye Diseases Using Self-Attention Mechanisms

Noor Ayesha
Center of Excellence in Cyber Security (CYBEX), Prince Sultan University Riyadh, Saudi Arabia
drnayesha@gmail.com (corresponding author)

Received: 20 February 2025 | Revised: 12 April 2025, 27 April 2025, and 3 May 2025 | Accepted: 4 May 2025

 DOI: https://doi.org/10.48084/etasr.10649

ABSTRACT

Daily life is highly dependent on the eyes, making them one of the most essential organs in the body. This study focuses on four eye conditions: Normal, Diabetic Retinopathy, Cataracts, and Glaucoma. This study presents a Convolutional Neural Network (CNN) model based on a Vision Transformer (ViT) with a Self-Attention Mechanism (SAM) for diagnosing various eye diseases. Initially, the dataset was preprocessed through resizing and normalization to enhance image quality and facilitate feature extraction. The proposed model was evaluated, achieving a commendable accuracy of 94% on test data, with an average AUC of 98.82%. This model effectively diagnoses conditions such as Diabetic Retinopathy, Cataracts, Glaucoma, and normal cases. The GUI-based application was developed and tested, allowing doctors to upload multiple images and analyze eye disease categories, enhancing interpretability and showing promise for clinical applications. The proposed model can assist ophthalmologists in detecting eye disorders, enabling timely treatment of patients and helping to prevent vision loss.

Keywords-eye disease; deep learning; vision transformer; classification; health risks

I. INTRODUCTION

The eyes are the most important organ in the body and are used mostly in daily routine activities. Eye diseases affect the retina, potentially causing permanent damage. Eye disorders can lead to blindness and difficulty in reading, writing, driving, navigating, and facial recognition [1]. Approximately 2.2 billion people face various forms of visual impairment worldwide [2]. The cost of vision impairment is estimated to be almost USD 411 billion annually around the world, imposing a significant burden on economies. Worldwide, only 36% of patients with vision impairment due to refractive errors and 17% of people with cataracts receive adequate treatment. Common eye diseases are cataracts, diabetic retinopathy, and glaucoma [3].

Cataracts cause blindness, affecting people aged 40 and older, with a prevalence of approximately 11.8% to 18.8% [4]. Diabetic retinopathy damages blood vessels in the retina due to high blood sugar content [5]. Glaucoma is a condition that causes vision loss due to the degeneration of retinal ganglion cells and the thinning of the retinal nerve fiber layer [6]. Early detection of eye disease can help with proper treatment and prevent the worst conditions. Computerized screening is critical due to limited awareness, high consultation costs, and the lack of ophthalmologists. Different studies have adopted eye imagery methods to predict and classify healthy versus diseased eyes [7]. In recent years, the application of Deep Learning (DL) has gained notable attention in medical image analysis. DL models, such as MobileNet, have been used in intelligent eye-tracking systems for autism spectrum disorder diagnosis [8]. With the latest advances in DL models, it is possible to design an accurate and prompt EyeState Electroencephalography (EEG) classification problem [9].

In [10], a deep convolutional generative adversarial network was applied to classify four types of eye disease, glaucoma, myopia, diabetic retinopathy, and normal, achieving 80.45% training and 83.74% validation accuracy, highlighting the potential in therapeutic diagnostics. In [11], CNN, VGG16, and Inception V3 models were applied for the same four categories, achieving 51.30%, 79.40%, and 81% accuracy, respectively. In [12], a DL-based EfficientNet-B3 model was used to classify retinal images divided into glaucoma, cataract, diabetic retinopathy, and normal, achieving an overall 92.50% accuracy. In [13], the VGG19 model was used to classify retinal images, achieving 87% accuracy on training data and 88% on testing data. In [7], a transfer learning method and an individual CNN were used to classify eyes with glaucoma, cataract, diabetic retinopathy, and normal, achieving 94% and 84% accuracy, respectively. This study proposes a CNN model based on a ViT with a Self-Attention Mechanism (SAM) to diagnose various eye diseases. The main contributions of this study are:

- The ViT (SAM) based CNN model achieved 94% and 97% accuracy on testing and training data, respectively, demonstrating its effectiveness in diagnosing diabetic retinopathy, cataract, glaucoma, and normal cases.

- The SAM enhances global context understanding, enabling the CNN model to capture subtle disease-specific patterns in imbalanced retinal images.
- A GUI-based app was developed using the proposed model, allowing doctors to upload multiple images and analyze different eye diseases.

II. PROPOSED FRAMEWORK

This study used the eye disease dataset in [14], which consists of four types: Normal, Cataract, Glaucoma, and Diabetic Retinopathy. The data consists of 4217 images in PNG format, where 1074 images belong to normal eyes, 1038 belong to Cataracts, 1007 belong to Glaucoma, and 1098 belong to Diabetic Retinopathy. All images have resolutions from 256×256 to 800×533 pixels.

A. Data Preprocessing

The following steps were used to preprocess images:

- Resizing: Images were resized into 100×100 pixels for better feature selection.
- Normalization: The resized images were normalized between 0 and 1 for a more balanced presentation.
- Partition: The preprocessed data was split into training and testing subsets: 70% of the images of the four classes were split into a training subset and 30% into a testing subset for the diagnosis process.

B. CNN

CNN models have achieved remarkable success in many computer vision applications, especially in image classification [15], and are considered a powerful method to learn image features for classification tasks. However, they are weak in operating with local neighbors and miss global information.

C. Self-Attention Mechanism (SAM)

SAM can deal with global information and capture long-range interactions, which is the key component of ViT. The attention mechanism was introduced in [16] for a transformer model that integrated with various neural network models to enhance their classification performance. SAM enables models to capture appropriate long-range relationships by giving weight importance to different elements. It is considered a central part of different deep learning models and has gained popularity in computer vision tasks [17, 18].

Many studies integrated SAM with CNN models, achieving remarkable performance. In [19], a SAM was integrated with a CNN model, improving the classification accuracy of speaker identification systems. In [20], an SAM was used with a multigram CNN model, achieving improved performance in relation classification tasks. In [21], SAM and CNN models were integrated to improve the diagnostic utility of low-dose CT scans. In [22], a hybrid ViT model was used for the detection of cervical cancer, achieving remarkable accuracy compared to EfficientNetB, DenseNet, Xception, and ResNet50. In [23], a SAM was combined with a CNN model, achieving better results compared to traditional methods.

D. Proposed Model Architecture

This study employed a CNN model based on a ViT (SAM) to identify and classify eye diseases. An input of 100×100×3 was fed into five convolutional layers with the same 3×3 kernel and 16, 32, 32, 64, and 64 filter sizes, respectively. The five convolutional layers used the ReLU activation function. Similarly, five pooling layers had the same 2×2 kernel size, and the SAM layer had 16 number heads and 64 key dimensions. Dropouts of 0.10 and 0.40 were established for the max-pooling and the dense layers, respectively, to avoid overfitting. Last, two dense layers with 64 and 128 filter sizes were added, and one output layer was connected for four types of eye images using the softmax activation function.

The model was compiled with the Adam optimizer and sparse categorical cross-entropy for loss, and accuracy was used to evaluate the training and test procedures. The model was trained for 25 epochs using early stopping to avoid overfitting. Table I details layer types, output shapes, and parameters of the proposed model. A total of 459,278 (1.75 MB) parameters were estimated, where 153,092 (598.02 KB) were trainable parameters and 306,186 (1.17MB) were optimizer parameters. These parameters optimize the performance of the proposed model for diagnosing four categories of eye images.

TABLE I. PARAMETERS OF THE PROPOSED MODEL

Layer type	Output shape	Parameters
InputLayer	None, 100, 100, 3	0
Conv2D	None, 100, 100, 16	448
Conv2D	None, 100, 100, 32	4,460
MaxPooling2D	None, 50, 50, 32	0
Dropout	None, 50, 50, 32	0
Conv2D	None, 50, 50, 32	9,248
MaxPooling2D	None, 25, 25, 32	0
Dropout	None, 25, 25, 32	0
Conv2D	None, 25, 25, 32	18,496
MaxPooling2D	None, 12, 12, 64	0
Dropout	None, 12, 12, 64	0
Conv2D	None, 12, 12, 64	36,928
MaxPooling2D	None, 6, 6, 64	0
Dropout	None, 6, 6, 64	0
SelfAttention	None, 6, 6, 64	66,496
Flatten	None, 576	0
GlobalAverage	None, 64	0
Dense	None, 128	8,320
Dropout	None, 128	0
Dense	None, 64	8,256
Dropout	None, 64	0
Dense	None, 4	260

E. Performance Metrics

The performance of the proposed model was evaluated using accuracy, precision, recall, f1-score, and Receiver Operating Characteristic - Area Under the Curve (AUC-ROC).

$$Accuracy = \frac{\sum(TPi)}{\sum(TPi+FPi+FNi)} \quad (1)$$

$$Precision = \frac{TPi}{TPi + FPi} \quad (2)$$

$$Recall = \frac{TPi}{TPi+FNi} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$AUC - ROC = \int_0^1 TPR(t)\, dFPR(t) \quad (5)$$

where TPi denotes True Positives for class i, FPi denotes False Positives for class i, and FNi denotes False Negatives for class i.

III. RESULTS AND DISCUSSION

A. Accuracy and Losses

Figure 1 presents the classification accuracy and loss performance of the proposed model across 150 epochs in the training and testing datasets. It can be observed that the accuracies of both datasets were increased and losses decreased with epochs.

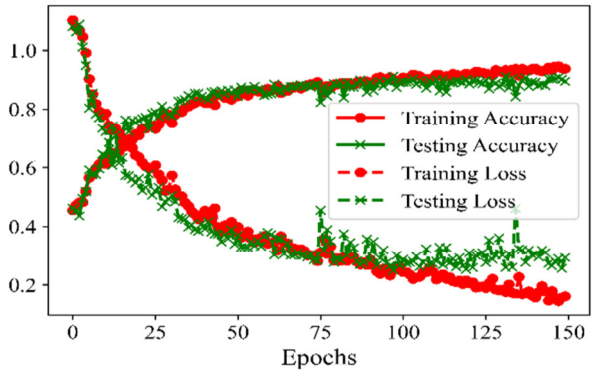

Fig. 1. Accuracy and loss of the proposed model.

B. Classification Report

In training, the proposed model achieved an accuracy of 97% with 0.93, 0.97, 1, and 0.99 precision, 0.99, 0.99, 0.90, and 1 recall, and 0.96, 0.98, 0.95, and 1 F1-score for Normal, Cataract, Glaucoma, and Diabetic Retinopathy classes, respectively. In testing, the proposed model achieved 94% accuracy, with 0.87, 0.93, 0.96, and 1 precision, 0.96, 0.97, 0.81, and 1 recall, and 0.91, 0.95, 0.88, and 1 F1-score for the Normal, Cataract, Glaucoma, and Diabetic Retinopathy classes, respectively.

TABLE II. CLASSIFICATION REPORT OF TRAINING DATA

Class	Precision	Recall	F1	Accuracy
Normal	0.93	0.99	0.96	97%
Cataract	0.97	0.99	0.98	
Glaucoma	1	0.90	0.95	
Diabetic Retinopathy	0.99	1	1	

TABLE III. CLASSIFICATION REPORT OF TESTING DATA

Class	Precision	Recall	F1	Accuracy
Normal	0.87	0.96	0.91	
Cataract	0.93	0.97	0.95	94%
Glaucoma	0.96	0.81	0.88	
Diabetic Retinopathy	1	1	1	

C. Confusion Matrices

Figure 2 shows the confusion matrix for the training data. The 759 images of Normal eye (0), 705 images of Cataract disease (1), 641 images of Glaucoma (2), and 760 images of Diabetic Retinopathy (3) were correctly classified by the model.

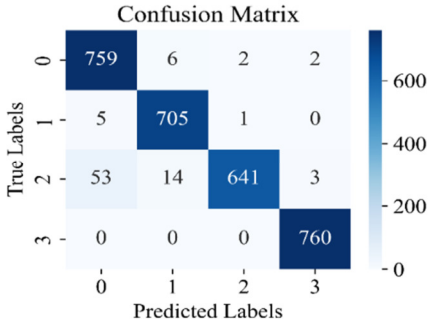

Fig. 2. Confusion matrix on training data.

In testing, 293 images of Normal eye (0), 316 images of Cataract disease (1), 239 images of Glaucoma (2), and 338 images of Diabetic Retinopathy (3) were correctly classified by the proposed model. Figure 3 shows details of correct and incorrect predicted test images. The proposed model incorrectly classified 86 images out of 2951 train images and 80 out of 1266 testing images.

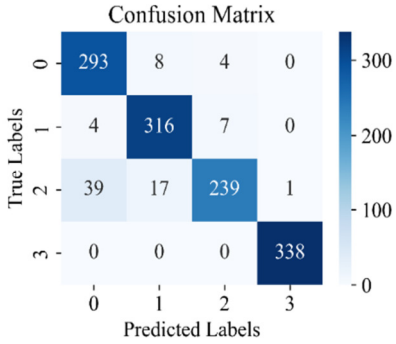

Fig. 3. Confusion matrix on testing data.

D. Sensitivity Analysis

AUC-ROC was used to measure the performance of the proposed model in multiclass classification. Figure 4 shows the training AUC in solid blue, green, red, and black lines for Normal eye (0), Cataract disease (1), Glaucoma (2), and Diabetic Retinopathy (3), respectively, and the testing AUC in dashed blue, green, red, and black lines, respectively. The training data achieved a 99.92% average AUC and the testing data achieved a 98.82% average AUC, which shows an optimal performance for the proposed model.

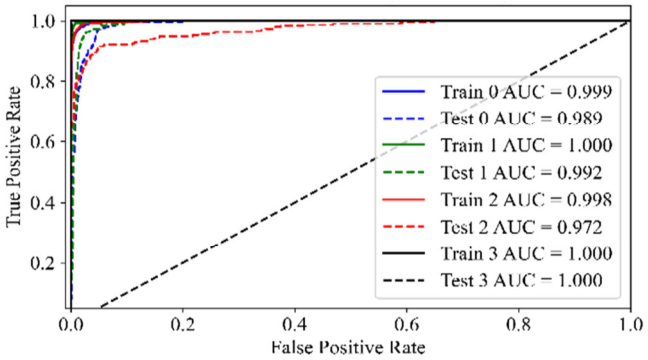

Fig. 4. AUC-ROC of the proposed model on training and testing data.

E. Comparative Analysis

The proposed model achieved higher accuracy than models presented in previous studies, as shown in the detailed comparative analysis in Table IV.

TABLE IV. COMPARATIVE ANALYSIS

Ref.	Year	Disease	Model	Best accuracy
[7]	2023	Cataracts, Diabetic Retinopathy, Glaucoma, and Normal	Transfer learning	94%
[10]	2021	Glaucoma, Myopia, Diabetic Retinopathy, and Normal	Deep convolutional generative adversarial network	83.74%
[11]	2019	Diabetic Retinopathy, Glaucoma, Myopia, and Normal	CNN, VGG16, and InceptionV3	81%
[12]	2024	Glaucoma, Cataract, Diabetic Retinopathy, and Normal	EfficientNet-B3	92.5%,
[13]	2023	Five classes	VGG19	88%
[24]	2021	Cataracts, Diabetic Retinopathy, Glaucoma, and Normal	DenseNet169-LSTM	69.50%

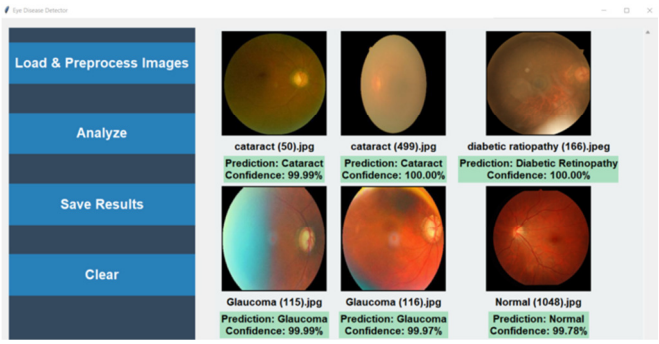

Fig. 5. Prediction confidence for testing data using the proposed model.

F. Eye Disease Detector App

A GUI application was built using the proposed model, and a graphical presentation showing the prediction confidence rate is shown in Figure 5. Eight random images from each category were selected to test the prediction confidence of the proposed model. This application can enhance interpretability and show promise for clinical applications. The ophthalmologist simply uploads multiple images and analyzes eye diseases, while he can also save them for future comparison. The GUI was tested for functionality, but clinical validation with real-world patient data is planned as future work to ensure practical adoption.

IV. CONCLUSION

This study aimed to identify eye diseases that impact human lives, focusing on four specific conditions: Diabetic Retinopathy, Cataracts, Glaucoma, and Normal. A CNN model based on ViT (SAM) was developed to diagnose four types of eye diseases. Initially, the dataset was preprocessed through resizing and normalization to enhance image quality and facilitate feature extraction. The proposed model was evaluated using standard approaches and achieved a testing accuracy of 94% with an average AUC of 98.82%. A GUI application was developed to effectively diagnose conditions such as Diabetic Retinopathy, Cataracts, Glaucoma, and Normal cases. The proposed model can improve interpretability and shows promise for clinical applications, assisting ophthalmologists in detecting eye disorders, enabling timely treatment for patients, and helping to prevent vision loss.

ACKNOWLEDGMENT

The author acknowledges Prince Sultan University, Riyadh, Saudi Arabia, for providing the Article Processing Charges (APC) for this publication.

LIMITATIONS AND FUTURE WORK

This study is limited by the size of the dataset due to data availability, which may affect the model's ability to generalize across diverse real-world cases. This study also used imbalanced data, which can cause overfitting problems. Future work could incorporate techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), class weighting, or focal loss to enhance model performance.

REFERENCES

[1] M. Hussain *et al.*, "An Enhanced Convolutional Neural Network (CNN) based P-EDR Mechanism for Diagnosis of Diabetic Retinopathy (DR) using Machine Learning," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 19062–19067, Feb. 2025, https://doi.org/10.48084/etasr.8854.

[2] S. A. Hassan, S. Akbar, A. Rehman, T. Saba, H. Kolivand, and S. A. Bahaj, "Recent Developments in Detection of Central Serous Retinopathy Through Imaging and Artificial Intelligence Techniques–A Review," *IEEE Access*, vol. 9, pp. 168731–168748, 2021, https://doi.org/10.1109/ACCESS.2021.3108395.

[3] A. Jabbar *et al.*, "Deep Transfer Learning-Based Automated Diabetic Retinopathy Detection Using Retinal Fundus Images in Remote Areas," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, May 2024, Art. no. 135, https://doi.org/10.1007/s44196-024-00520-w.

[4] X. Chen, J. Xu, X. Chen, and K. Yao, "Cataract: Advances in surgery and whether surgery remains the only treatment in future," *Advances in Ophthalmology Practice and Research*, vol. 1, no. 1, Nov. 2021, Art. no. 100008, https://doi.org/10.1016/j.aopr.2021.100008.

[5] D. Kothadiya, A. Rehman, S. Abbas, F. S. Alamri, and T. Saba, "Attention-based deep learning framework to recognize diabetes disease from cellular retinal images," *Biochemistry and Cell Biology*, vol. 101, no. 6, pp. 550–561, Dec. 2023, https://doi.org/10.1139/bcb-2023-0151.

[6] H. Naz, R. Nijhawan, N. J. Ahuja, T. Saba, F. S. Alamri, and A. Rehman, "Micro-segmentation of retinal image lesions in diabetic retinopathy using energy-based fuzzy C-Means clustering (EFM-FCM),"

Microscopy Research and Technique, vol. 87, no. 1, pp. 78–94, 2024, https://doi.org/10.1002/jemt.24413.

[7] T. Babaqi, M. Jaradat, A. E. Yildirim, S. H. Al-Nimer, and D. Won, "Eye Disease Classification Using Deep Learning Techniques." arXiv, Jul. 19, 2023, https://doi.org/10.48550/arXiv.2307.10501.

[8] H. Naz, T. Saba, F. S. Alamri, A. S. Almasoud, and A. Rehman, "An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection," *IEEE Access*, vol. 12, pp. 78611–78623, 2024, https://doi.org/10.1109/ACCESS.2024.3392032.

[9] "Compact Bat Algorithm with Deep Learning Model for Biomedical EEG EyeState Classification," *Computers, Materials and Continua*, vol. 72, no. 3, pp. 4589–4601, Apr. 2022, https://doi.org/10.32604/cmc.2022.027922.

[10] M. Smaida, S. Yaroshchak, and Y. El Barg, "DCGAN for Enhancing Eye Diseases Classification," in *Computer Modeling and Intelligent Systems*, 2021, vol. 2864, pp. 22–33, https://doi.org/10.32782/cmis/2864-3.

[11] M. Smaida and D. Y. Serhii, "Comparative Study of Image Classification Algorithms for Eyes Diseases Diagnostic," *International Journal of Innovative Science and Research Technology*, vol. 4, no. 12, pp. 40–48, 2019.

[12] A. Ramis Uulu, G. Gimaletdinova, and Z. Orozakhunov, "Eye Disease Classification Using Deep Learning Approaches: A Case Study on Retinal Images." Preprints.org, Dec. 25, 2024, https://doi.org/10.20944/preprints202412.1996.v2.

[13] I. Topaloglu, "Deep Learning Based Convolutional Neural Network Structured New Image Classification Approach for Eye Disease Identification," *Scientia Iranica*, vol. 30, no. 5, pp. 1731–1742, Oct. 2023, https://doi.org/10.24200/sci.2022.58049.5537.

[14] "Eye diseases classification." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification.

[15] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention Augmented Convolutional Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 3285–3294, https://doi.org/10.1109/ICCV.2019.00338.

[16] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[17] T. Saba, "Automated lung nodule detection and classification based on multiple classifiers voting," *Microscopy Research and Technique*, vol. 82, no. 9, pp. 1601–1609, 2019, https://doi.org/10.1002/jemt.23326.

[18] A. Rehman and T. Saba, "Features extraction for soccer video semantic analysis: current achievements and remaining issues," *Artificial Intelligence Review*, vol. 41, no. 3, pp. 451–461, Mar. 2014, https://doi.org/10.1007/s10462-012-9319-1.

[19] N. N. An, N. Q. Thanh, and Y. Liu, "Deep CNNs With Self-Attention for Speaker Identification," *IEEE Access*, vol. 7, pp. 85327–85337, 2019, https://doi.org/10.1109/ACCESS.2019.2917470.

[20] C. Zhang *et al.*, "Multi-Gram CNN-Based Self-Attention Model for Relation Classification," *IEEE Access*, vol. 7, pp. 5343–5357, 2019, https://doi.org/10.1109/ACCESS.2018.2888508.

[21] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising With Self-Supervised Perceptual Loss Network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020, https://doi.org/10.1109/TMI.2020.2968472.

[22] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model," *IEEE Access*, vol. 10, pp. 73079–73086, 2022, https://doi.org/10.1109/ACCESS.2022.3189374.

[23] B. L. Chen, J. J. Wan, T. Y. Chen, Y. T. Yu, and M. Ji, "A self-attention based faster R-CNN for polyp detection from colonoscopy images," *Biomedical Signal Processing and Control*, vol. 70, Sep. 2021, Art. no. 103019, https://doi.org/10.1016/j.bspc.2021.103019.

[24] M. Londhe, "Classification of Eye Diseases using Hybrid CNN-RNN Models," M.S. Thesis, Dublin, National College of Ireland, 2021.