Journal of Business and Social Sciences (ISSN: 2350-8876)
Research Department, Hetauda School of Management and Social Sciences (HSMSS), Hetauda

Classification of Retinal Disorders from OCT Images using Attention based CNN

Er. Gunjan Gupta
Lecturer, Hetauda School of Management & Social Sciences
IT Department
Email: gunjan@hsm.edu.np

Cite this paper:

Gupta, G. (2024). Classification of retinal disorders from OCT images using attention based CNN. *Journal of Business and Social Sciences, 6*(1), 41-56

https://doi.org/10.3126/jbss.v6i1.78753

Abstract

Efficient automated decision support systems for the detection of retinal disorders are crucial in ophthalmology. Optical Coherence Tomography (OCT), a widely used imaging modality, allows visualization and measurement of retinal layer thickness, aiding in the early detection of disorders such as age-related macular degeneration (AMD), diabetic macular edema (DME), and other abnormalities. Despite advancements, existing methodologies often lack generalization and primarily focus on entire retinal images, disregarding the central retinal region where most abnormalities manifest. This study proposes a deep learning model integrating attention mechanisms with a convolutional neural network (CNN) and auto-encoder for OCT image classification into four categories: Choroidal Neovascularization (CNV), DME, Drusen, and Normal. The attention mechanism emphasizes relevant features, while the auto-encoder detects anomalies effectively. Optimized using random search, the model achieves a remarkable accuracy of 97.8%, with precision, recall, and F2-scores of 98.4%, 98.3%, and 98.3%, respectively, demonstrating significant improvement over existing approaches. This model offers enhanced accuracy and efficiency for retinal disorder classification, promising improved diagnostic and treatment planning in clinical applications.

Keywords: Optical Coherence Tomography (OCT), Convolutional Neural Network (CNN), Attention Mechanism, Retinal Disorders, Deep Learning

Introduction

Retinal disorders, including Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen, represent a growing global health challenge, particularly as they contribute significantly to vision impairment and blindness (World Health Organization, 2021). Age-related macular degeneration and diabetic eye diseases are among the leading causes of these disorders, necessitating effective diagnostic tools to mitigate their impact. Early detection is critical, as timely intervention can prevent progression and preserve vision. Optical Coherence Tomography (OCT), a non-invasive imaging modality, has revolutionized ophthalmology by providing high-resolution cross-sectional images of retinal layers. These images allow clinicians to identify structural abnormalities with exceptional precision (Huang et al., 1991).

Despite its utility, manual interpretation of OCT images remains labor-intensive and subject to variability across clinicians. Automated systems for retinal disorder classification address these challenges by reducing diagnostic time and improving consistency (Li et al., 2020). However, existing methods often fail to generalize well to diverse datasets or to focus on the central retinal region, where critical features indicative of disease are frequently observed. These shortcomings highlight the need for advanced models capable of extracting and prioritizing relevant features while maintaining high accuracy across different imaging conditions (Srinivasan et al., 2014).

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown considerable promise in medical image analysis (LeCun et al., 2015). CNNs excel at identifying spatial hierarchies in image data, making them suitable for tasks such as retinal disorder classification. However, their performance can be further enhanced by integrating attention mechanisms and auto-encoders. Attention mechanisms enable the model to focus selectively on critical features within the image, while auto-encoders facilitate the detection of subtle anomalies and enhance feature encoding. This hybrid approach addresses the limitations of traditional CNN models and offers a robust framework for OCT image classification (Chen et al., 2017).

This study aims to develop an attention-based CNN model combined with auto-encoders to classify OCT images into four categories: CNV, DME, Drusen, and Normal. The proposed model is designed to optimize feature extraction, improve classification accuracy, and provide a reliable tool for clinical applications. By leveraging techniques such as random search optimization for hyperparameter tuning and k-fold cross-validation for robustness, the study demonstrates the potential of this hybrid architecture to outperform existing methods. The results validate the model's efficacy and its applicability in real-world diagnostic settings, paving the way for more efficient and accurate automated systems in ophthalmology (Goodfellow et al., 2016).

Literature Review

Advancements in medical imaging, particularly Optical Coherence Tomography (OCT), have revolutionized the diagnosis and management of retinal disorders. OCT provides detailed cross-sectional images of the retina, facilitating early detection of conditions like diabetic macular edema (DME), age-related macular degeneration (AMD), and drusen-related macular degeneration (Pandey, 2023). Recent research has focused on leveraging deep learning models to improve diagnostic efficiency and accuracy, given the limitations of manual image analysis.

Pandey (2023) highlights the potential of convolutional neural networks (CNNs) in medical imaging due to their ability to identify intricate features in images. Incorporating auto-encoders and attention mechanisms enhances this capability. Auto-encoders compress data, focusing on anomalies, while attention mechanisms emphasize relevant features in spatial and channel dimensions, significantly improving classification accuracy for conditions like choroidal neovascularization (CNV), DME, and drusen.

Numerous studies underscore the effectiveness of OCT and CNNs in retinal disease detection. Rajagopalan et al. (2021) achieved 97% accuracy using deep learning but lacked generalization due to dataset biases. Similarly, Upadhaya et al. (2022) utilized coherent CNNs to enhance accuracy but faced challenges with imbalanced datasets. Attention mechanisms have shown promise in overcoming these limitations by selectively amplifying crucial features (Pandey, 2023).

Cross-validation and hyperparameter tuning are pivotal in optimizing model performance, as demonstrated in experiments with different architectures. Models combining CNNs with auto-encoders and spatial-channel attention consistently outperform conventional CNNs, achieving an overall accuracy of 97.8% and precision of 98.4% (Pandey, 2023). Furthermore, preprocessing techniques like center cropping and resizing are essential for focusing on relevant retinal regions, significantly enhancing classification results.

The limitations of single-disease classification models are notable. Real-world scenarios often require multi-label classification to address coexisting conditions. Multi-hybrid models integrating attention mechanisms offer a pathway for future research, enabling comprehensive diagnostic solutions tailored for clinical applications (Pandey, 2023).

The literature underscores the efficacy of deep learning, especially attention-enhanced CNNs, in addressing the challenges of OCT image classification. Despite advancements, future research should focus on multi-label classification and model generalization to tackle coexisting retinal disorders effectively.

Methodology

The very first step is to collect data from different medical & healthcare centers, also some from data available on the internet, then the dataset is preprocessed, segmented and features will be extracted from the data.

The complete system is a combination of three models which optimizes the whole procedure Retinal optical coherence tomography (OCT) technique. The models are:

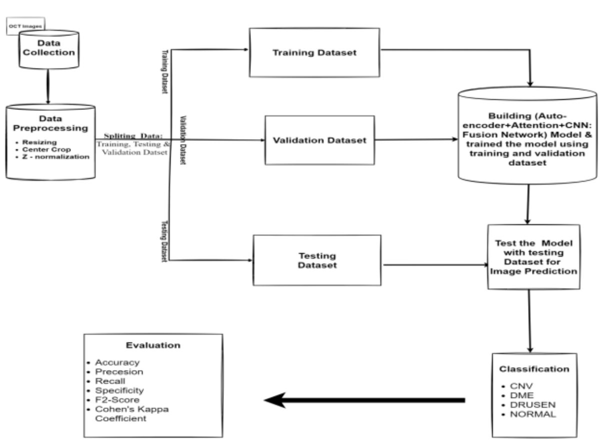

Figure 1: Methodology

Data Collection

The Kaggle Retinal OCT Images dataset, comprising 84,615 images across four categories (Normal, Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen), was utilized for this study (kermany, 2018). This dataset, due to its high-quality and labeled images, supports accurate model training and evaluation. The dataset's distribution and sample images are illustrated in Figure 2: Dataset Overview.

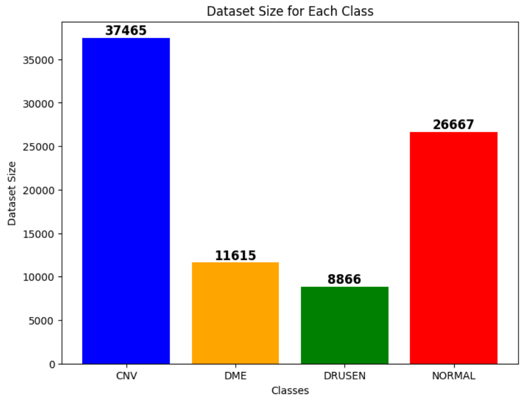

Figure 2: Dataset Overview

Data Preprocessing

To ensure consistency and optimize model performance, the following steps were applied:

Image Resizing and Cropping: All images were resized to 64×64 pixels and center-cropped to emphasize the diagnostically significant central retinal region.

Normalization: Pixel intensities were normalized using:

Data Splitting: Data was divided into training, validation, and testing subsets.

To ascertain the generalization capabilities of the trained model, thereby preventing overfitting or underfitting during training, the input dataset is partitioned into three subsets:

Training Set: Employed to teach the model underlying patterns and correlations.

Validation Set: Facilitates unbiased assessment while fine-tuning hyperparameters.

Test Set: Reserved for unbiased evaluation of the finalized model's performance.

This division is achieved in different ratio between training and test sets. Further enhancing the model's generalization ability, the training set is subjected to k-fold cross-validation. This technique mitigates overfitting concerns and promotes the identification of the most robust and generalized model configuration. Notably, the test set is reserved exclusively for the ultimate assessment of the model.

Model Architecture

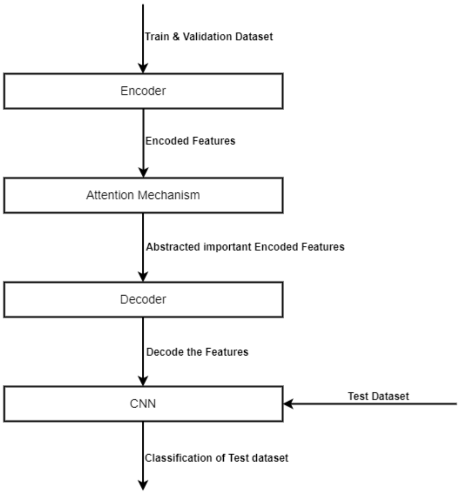

Figure 3: Model View of Attention Based CNN

The model architecture of system contains three parts: Auto-encoder, attention and CNN. The main aim of using Auto-encoder is to encode-decode features in efficient manner and detect anomaly between the features. While making model to extract useful and important features present in an image, the attention mechanism can be combined with CNN-based Auto-encoder architecture. Typically, the attention mechanism is incorporated into the decoder part of the network in this architecture. The most important features of an image are identified and extracted through the attention mechanism, making it possible to efficiently use parameters such as time, cost, and memory in the classification of CNV, DME, DMD abnormalities of eyes. For this architecture we design a model with 30 layers and perform the research task.

The layers that are present to process the whole models are mainly combination of following:

Convolutional-2D: A 2D Convolution operation is a widely used operation in computer vision and deep learning. It is a mathematical operation that applies a filter to an image, producing a filtered output. PyTorch provides a convenient and efficient way to apply 2D Convolution operations. It provides functions for performing operations on tensors and it also provides functions for building deep learning models. Convolutions are a fundamental concept in computer vision and image processing. They are mathematical operations that take an input signal and produce a transformed output signal that highlights certain features of the input. Convolutional neural networks are deep learning models that are built using convolutions as a core component.

ReLU: ReLU stands for rectified linear activation unit and is considered one of the few milestones in the deep learning revolution. It's simple but really better than the trigger functions of its predecessors like sigmoid or tanh.

Maxpool-2D: Max pooling is a sample-based discretization process. The objective is to down-sample an input representation (image, hidden-layer output matrix, etc.), reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned. This is done to in part to help over-fitting by providing an abstracted form of the representation. As well, it reduces the computational cost by reducing the number of parameters to learn and provides basic translation invariance to the internal representation. Max pooling is done by applying a max filter to (usually) non-overlapping subregions of the initial representation.

Adaptiveavgpool-2D: Applies a 2D adaptive average pooling over an input signal composed of several input planes. The output is of size H x W, for any input size. The number of output features is equal to the number of input planes.

Self Attention: Self-attention mechanism is also called as intra-attention due to it uses attention mechanism in same channel. It is a mechanism that captures the relationships between different elements within the same input sequence or feature map. It computes different attention weights for each element based on its relation to other elements. In the context of computer vision, self-attention can be applied to the feature maps to capture dependencies between different spatial positions within each channel.

Channe lAttention: Channel-wise attention focuses on capturing the importance of different channels in the feature maps. It learns to assign different attention weights to different channels to emphasize relevant channels while suppressing less relevant ones.

Spatial Attention: Spatial attention focuses on capturing the importance of different spatial positions within each channel of the feature maps. It learns to assign different attention weights to different spatial locations to emphasize relevant spatial positions while suppressing less relevant ones.

Linear: The linear layer is used in the final stages of the neural network. It is also called a fully connected layer or dense layer in Keras. This layer helps in changing the dimensionality of the output from the preceding layer so that the model can easily define the relationship between the values of the data in which the model is working.

*Dropou*t: The Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting. Inputs not set to 0 are scaled up by 1/(1 - rate) such that the sum over all inputs is unchanged.

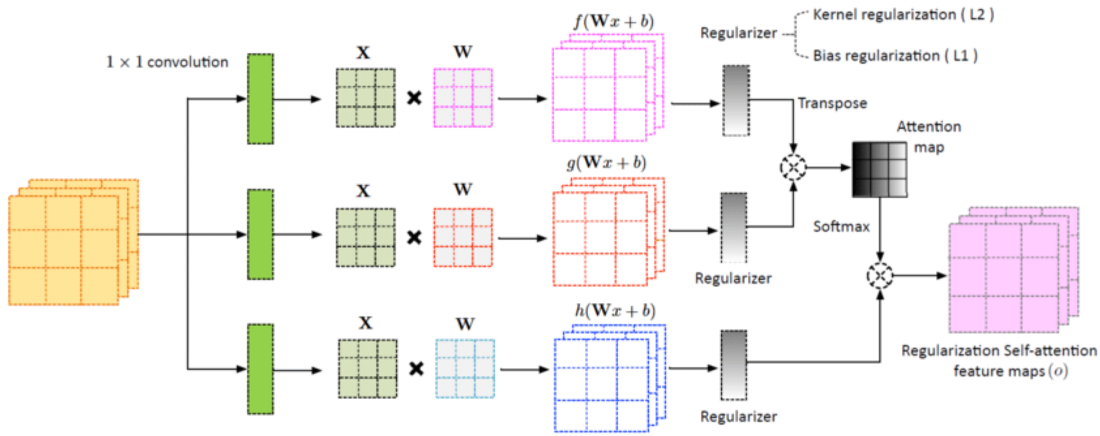

Figure 4: Self Attention Mechanism (Zhou et al., 2019)

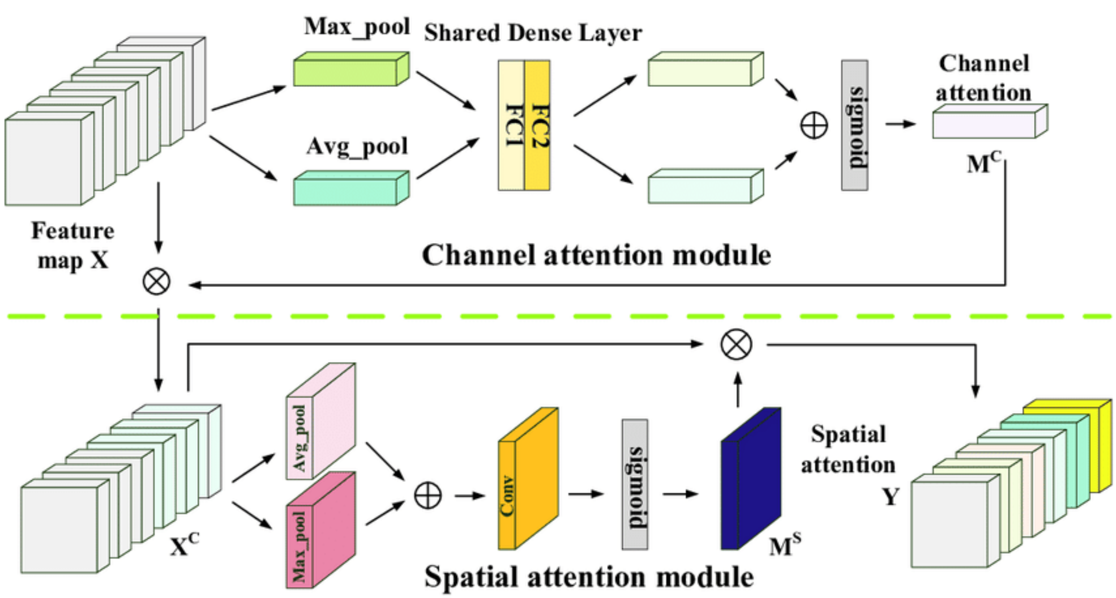

Figure 5: Channel-wise and Spatial Attention Mechanism (Xu et al., 2020)

Analysis Metrics

Many classification metrics can be used to analyze the system. The system developed is used to discriminate between CNV, DME, DMD and Normal. Confusion matrix for a multiclass classifier can be created as Table 1.

Table 1: Confusion Matrix

		Predicted			
		CNV (Class-0)	DME (Class-1)	DRUSEN (Class-2)	NORMAL (Class-3)
True	CNV (Class-0)	TP	FP	FP	FN
	DME (Class-1)	FP	TP	FP	FN
	DRUSEN (Class-2)	FP	FP	TP	FN
	NORMAL (Class-3)	FP	FP	FP	TN

From the confusion matrix, other performance metrics are calculated to evaluate the performance of the proposed system including Accuracy, Precision, Recall, F1-Score, Specificity and Error Rate.

Accuracy is the measurement of closeness, the mathematical formula for accuracy is:

$$\mathrm{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

Precision is the ratio of correctly classified positive examples to the no. of examples labeled by the system, given as:

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall, is the probability outcomes of model to correctly identifies the threat detection and also called sensitivity or true positive rates and calculated by:

$$\mathrm{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-Score is the harmonic average of precision & recall. It gives the single measure of comparison and higher is better but in case of this research recall is more important and hence FN assumes

higher priority so $F - \beta score$ can calculate where $\beta > 1$.

$$F - \beta\ \mathrm{score} = \frac{(1+\beta^2)\times \mathrm{Precision}\times \mathrm{Recall}}{\beta^2\times \mathrm{Precision}+\mathrm{Recall}} \quad (4)$$

When $\beta = 2$,

$$\text{F2} - \text{score} = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

Specificity, also known as true negative rate, is a metric used in binary classification to measure the probability that the model correctly identifies the absence of the threat or negative class. It is the ratio of correctly predicted negative instances (true negatives) to the total number of actual negative instances.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (5)$$

The Receiver Operator Characteristic (ROC) curve is a classification problem for evaluation metrics. It plots TPR vs FPR at different threshold values. Area under the curve (AUC) is the classifier's ability to distinguish classes and summarize the ROC curve. The higher the AUC value, the better the performance of the model for the classification task. The higher the AUC value, the more powerful the system. The main idea behind its use is that it can analyze the problem of class imbalance. It is not necessary to analyze ROC/AUC in the balanced classification.

Cohen's Kappa Coefficient is a statistical measure used to assess the agreement between two raters or evaluators when dealing with categorical data. It takes into account the possibility of agreement occurring by chance and provides a more robust evaluation of inter-rater agreement than simple percentage agreement. A higher value of Cohen's Kappa indicates better agreement between the raters, while a value close to 0 indicates poor agreement or agreement equivalent to chance. It is commonly used in inter-rater reliability studies, especially in fields where subjective judgments are involved, such as medical diagnoses or image classifications.

$$k = \frac{(P_0 - P_e)}{(1 - P_e)} \qquad (6)$$

where,

P_0 is the observed agreement between the raters, which is the proportion of cases where both raters agree.

P_e is the expected agreement between the raters under the assumption of chance agreement.

Tools used

The system is implemented and evaluated in a LINUX platform and created using python programming language. This research has also used additional libraries for python. The following tools are used:

- Kaggle: online platform for data science and machine learning datasets collaboration
- Google Drive: for storing models and history log of trained model

- Numpy, Pandas, Sci-kit Learn, Matplotlib: are python library used for scientific computation, manipulations and analysis, various classification, regression and clustering algorithms, and plotting purposes
- Keras, Pytorch: an open-source library that provides a Python interface for neural networks.
- Jupyter Notebook: Jupyter notebook allows us to easily write and execute Python codes and widely used in scientific computing and research purposes.

Results

Dataset Preprocessing Outcomes

The dataset was preprocessed to standardize images, resize them to 64×64 pixels, and focus on central retinal regions through cropping. This preprocessing improved the clarity and relevancy of image features used for training. Figure 6: Preprocessing Workflow illustrates the transformations from raw images to normalized inputs.

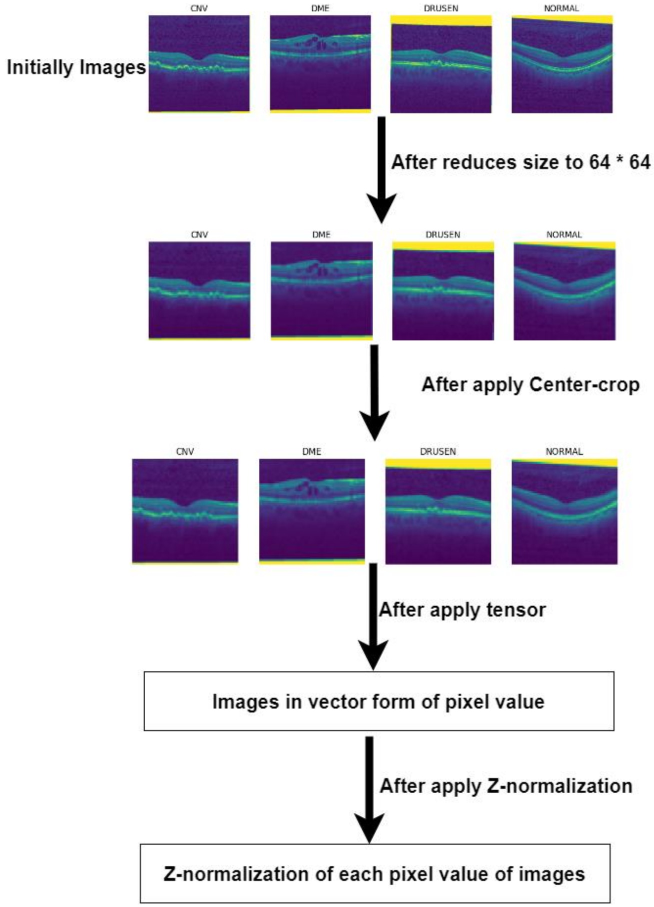

Figure 6: Preprocessing of Data

Model Performance

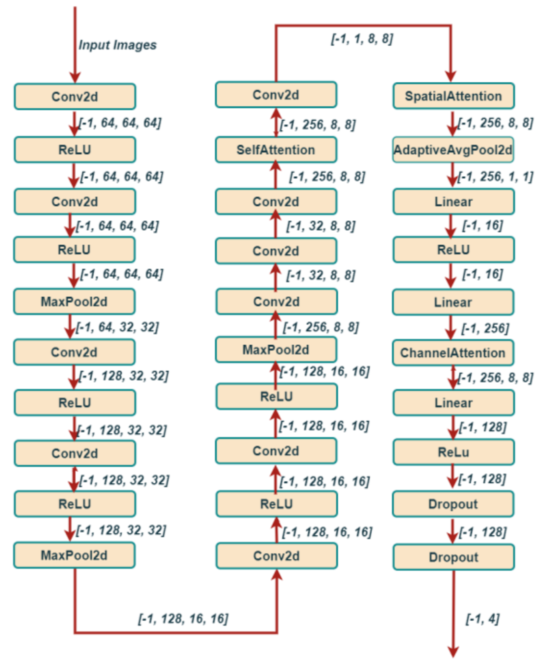

Figure 7: Model Layer Architecture

The proposed model, integrating auto-encoders, attention mechanisms, and CNNs, achieved superior performance metrics. The training process optimized the model for high classification accuracy and reduced generalization errors.

Table 2: Performance in Different Architecture

Architecture	Accuracy	Precision	Recall	F-2 Score	Specificity
CNN	0.927	0.929	0.928	0.928	0.928
Auto-encoder	0.942	0.943	0.941	0.942	0.942
Auto-encoder & CNN	0.946	0.949	0.948	0.948	0.948
Auto-encoder with self-attention & CNN	0.953	0.937	0.929	0.931	0.933
Auto-encoder with spatial + channel-wise attention & CNN	0.954	0.954	0.954	0.954	0.954

During the research we firstly, train the model with different architecture and hence achieve the better performance using auto-encoder and attention mechanism with CNN which is shown in above figure. The model evaluation in different architecture was conducted with adding self, spatial and channel attention the model has high accuracy with recall and precision values. Adam optimizer is used for training. The best among them is combination of Auto-encoder with Self, Spatial & Channel attention followed by CNN for classification. After finding of best model architecture, to make the system more robust and efficient the k-fold cross validation and hyperparameter optimization process is done randomly.

To ensure the system stability and validation the k-fold cross validation techniques is apply. In cross-validation approaches accuracy remains 96.22%. The system exposes the high accuracy in k-fold which can be shown in table 2. The k-fold generalizes the model as it is validate. Also it overcomes the problem of overfitting and unbalances structure of datasets. Finally, calculation of average among all folds gives the value of 96.22% of accuracy of system.

Table 3: k-fold cross validation with k=5

Fold	Accuracy	Precision	Specificity	Recall	F2-Score	Cohen Kappa's Coefficient
1	0.9434	0.9434	0.9435	0.9435	0.9435	0.9247
2	0.9551	0.9572	0.9570	0.9570	0.9587	0.9310
3	0.9557	0.9555	0.9555	0.9555	0.9553	0.9351
4	0.9746	0.9734	0.9731	0.9731	0.9721	0.9573
5	0.9824	0.9815	0.9814	0.9814	0.9807	0.9752
Average	**0.9622**	**0.9622**	**0.9621**	**0.9621**	**0.9621**	**0.9447**

The system is compared to pretrained model as to find the weather system achieves better efficiency or not. As we found result are high difference in terms of accuracy and other evaluation parameters. The results of pretrained model are shown in figure 8.

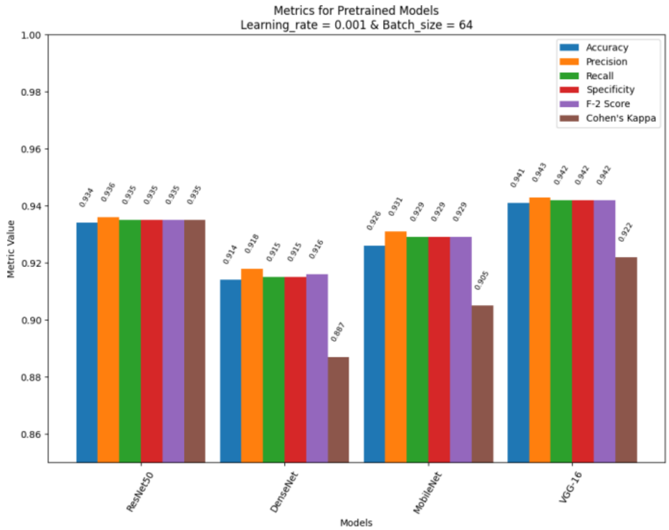

Figure 8: Pretrained Model Results

As the aim of study is to find optimal values of hyperparameters the ratio on which data split is change in different values and results of such process is quite different as we use high ratio for train dataset and low ratio for test dataset which lead the better results in terms of accuracy whereas increase in test value and decrease in train values leads the system less accurate. The experimental of different ratio of data are as shown in table.

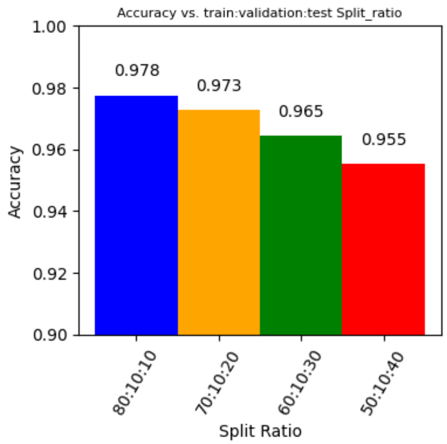

Figure 9: Accuracy vs. Split Ratio

The bar graph of figure 9 shows that while increasing the testing ratio with decreasing the training ratio the accuracy of the system also decreases. As we know that neural network becomes more accurate and robust using high data as train, so the above figure also illustrates that. In another hand, to regulate the overfitting of data the dropout is use in model. The different dropout rates encompass the different accuracy such that curves between dropout & accuracy can be explained by figure.

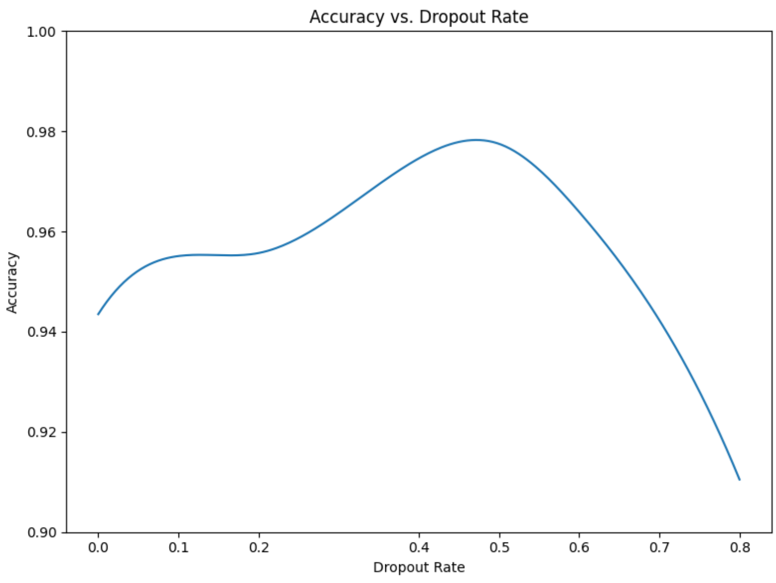

Figure 10: Dropout vs. Accuracy Curve

Table 4:Results of Auto-encoder+Attention+CNN Model

Auto-encoder + (Self + Channel-wise + Spatial) attention + CNN						
Class	***Accuracy***	***Precision***	***Recall***	**F-2 Score**	**Specificity**	**Cohen's Kappa Coefficient**
CNV		0.972	0.988	0.971	0.984	
DME		0.975	0.983	0.975	0.982	
DRUSEN		0.992	0.979	0.992	0.982	
NORMAL		0.996	0.983	0.996	0.986	
Overall	0.978	0.984	0.983	0.983	0.983	0.967

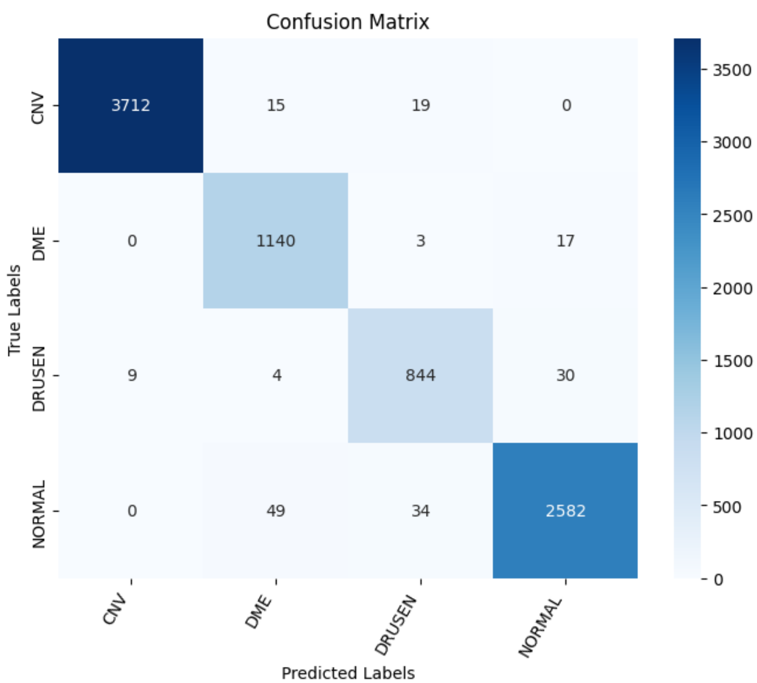

Figure 11: Confusion Matrix

The dataset contains total images of 84615 including CNV data 37465, DME data 11615, DRUSEN data 8866 and NORMAL data 26667 already shown in Section 3.2. While splitting 10% of data are used for testing that contains CNV, DME, DRUSEN & NORMAL as: 3746, 1161, 886 & 2665 respectively. To overcome the overfitting problem "early stopping" function is used.

Among the data, the model classifies 8278 data correctly and remaining 180 data has misclassification so model achieved accuracy of 97.75% (Early Stopped at 23 Epochs), precision of 98.40%, recall of 98.30% and F2-score of 98.30%. Calculation of its Cohen's Kappa coefficient is 0.967 which shown observed agreement and expected agreement between diseases are highly coupled.

Conclusion and Discussion

The auto-encoder with an attention mechanism followed by a CNN model was trained on a real dataset, achieving an impressive accuracy of 97.75%, precision of 98.40%, recall of 98.30%, and an F2-score of 98.30%. This performance highlights the model's ability to effectively learn and capture patterns and features from the dataset, enabling accurate predictions on unseen data. When trained on a large dataset, the model demonstrated significantly improved accuracy and performance compared to training on a smaller dataset. Additionally, the system achieved a Cohen’s Kappa coefficient of 0.967, indicating a strong observed and expected agreement between disease classifications. This underscores the close coupling of disease-related patterns. The attention mechanism proved highly effective for extracting localized and critical features from central regions of images, while the auto-encoder was instrumental in identifying anomalies.

Notably, the model performed poorly when images were processed without central cropping, whereas focusing on central image regions yielded superior results. This observation suggests that retinal disorders primarily affect the central part of the retina. Therefore, it is more effective to analyze and extract features from the central region of images rather than considering the entire image. The attention mechanism excels at isolating important features, while the auto-encoder reliably detects anomalies.

In summary, the developed model demonstrated strong potential for accurately classifying CNV, DME, DRUSEN, and NORMAL conditions from retinal OCT images. Future optimization of parameters will further enhance the system's performance, making it well-suited for real-world implementation.

Future Directions

Future research should focus on multi-label classification to handle coexisting retinal disorders, as real-world cases often involve multiple conditions. Incorporating explainable AI (XAI) techniques could enhance the interpretability of attention maps, providing insights into the decision-making process. Additionally, testing the model on diverse datasets and developing cross-domain applications would improve generalizability.

References

Ai, Z., Huang, X., Feng, J., Wang, H., & Y. T. (2022). FN-OCT: Disease detection algorithm for retinal optical coherence tomography based on a fusion network. *Frontiers in Neuroinformatics*. [Online]. Available at: https://www.kaggle.com/datasets/paultimothymooney/kermany2018.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2017). Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis, 48*, 41–52. https://doi.org/10.1016/j.media.2017.10.012

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Huang, D., Swanson, E. A., Lin, C. P., Schuman, J. S., Stinson, W. G., Chang, W., ... & Fujimoto, J. G. (1991). Optical coherence tomography. *Science, 254*(5035), 1178–1181. https://doi.org/10.1126/science.1957169

Li, T., Zhang, Z., & Liang, P. (2020). Automated diagnosis of retinal diseases using deep learning: A survey of the state-of-the-art. *Journal of Medical Systems, 44*(5), 78–92. https://doi.org/10.1007/s10916-020-01580-8

Pandey, T. R. (2023). *Classification of retinal disorders from OCT images using attention-based CNN* (Master's thesis). Nepal College of Information Technology, Pokhara University, Nepal.

Rajagopalan, N., Narasimhan, V., Vinjimoor, S. K., & Aiyer, J. (2020). Deep CNN framework for retinal disease diagnosis using optical. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-02095-2

Upadhyay, P. K., Rastogi, S., & Kumar, K. (2022). Coherent convolution neural network-based retinal disease detection using optical coherence tomographic images. *Journal of King Saud University – Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2022.10.001

World Health Organization. (2021). *World report on vision*. Retrieved from WHO website.

Xu, Q., Xiao, Y., Wang, D., & Luo, D. B. (2020). CSA-MSO3DCNN: Multiscale octave 3D CNN with channel and spatial attention for hyperspectral image classification. *Remote Sensing*. https://doi.org/10.3390/rs12091312

Zhou, J., He, Z., Song, Y. N., Wang, H., Yang, X., Lian, W., & Dai, H.-N. (2019). Precious metal price prediction based on deep regularization self-attention regression. *IEEE Access, 4*. https://doi.org/10.1109/ACCESS.2019.2941223