**HYBRID MODEL FOR DIABETIC RETINOPATHY CLASSIFICATION FROM FUNDUS IMAGES USING DEEP LEARNING**

**A PROJECT REPORT**

Submitted By

**AARTHE R S (20EC1001)**

**ASWIN V (20EC1006)**

**KARAN J (20EC1038)**

**KEERTHIVASAN R (20EC1041)**

Under the guidance of

**Dr.V.VIJAYLAKSHMI**

**Professor**

Submitted to Puducherry Technological University in partial fulfillment

for the award of the degree of

**BACHELOR OF TECHNOLOGY**

in



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING PUDUCHERRY TECHNOLOGICAL UNIVERSITY**

**PUDUCHERRY – 605014**

**APRIL 2024**

**BONAFIDE CERTIFICATE**

Certified that this project report **HYBRID MODEL FOR DIABETIC RETINOPATHY CLASSIFICATION FROM FUNDUS IMAGES USING DEEP LEARNING** is the bonafide work done by **AARTHE R S [20EC1001], ASWIN V[20EC1006], KARAN J [20EC1038] and KEERTHIVASAN R[20EC1041]** Seventh semester B.Tech. class of **Electronics and Communication Engineering** in the project work (phase I) during the year 2023 – 2024.

**PROJECT GUIDE HEAD OF THE DEPARTMENT**

**Dr.V.VIJAYLAKSHMI DR. V. SAMINADAN**

**ACKNOWLEDGEMENT**

The successful completion of the project is indeed practically incomplete without the mention of all those people who greatly supported and encouraged us throughout the project. We feel grateful to our guide **Dr.V.VIJAYLAKSHMI ,** AssociateProfessor, Department of Electronics and Communication Engineering, Puducherry Technological University, for her encouragement and support. She has been of valuable guidance, suggestions and kindness during the course of project work.

We would like to express our sincere thanks to our Respected VICE-CHANCELLOR **Dr. S. MOHAN,** Puducherry Technological University for providing the college facilities for the completion of this project.

We would like to express our heartfelt gratitude to **Dr. V. SAMINADAN,** Professor and Head of the Department of Electronics and Communication Engineering, Puducherry Technological University for providing all department facilities and guidance which enabled us to complete the project.

We feel obliged to thank the review by the panel members **Dr. G. NAGARAJAN** Professor, **Dr. S. BATMAVADY** Professor, and **Dr. V. VIJAYALAKSHMI** Professorof the Electronics and Communication Engineering Department, for their support, encouragement and guidance.

We express our deep sense of gratitude to all teaching and non-teaching staff of our department, and to our friends for their support and encouragement during the entire course of this dissertation work. We also express our sincere thanks to our parents who motivated us in all means.

**ABSTRACT**

Diabetic retinopathy (DR) remains a major cause of vision impairment worldwide, necessitating the development of accurate and efficient diagnostic tools.

This project introduces a groundbreaking Hybrid Model for Diabetic Retinopathy Classification, a crucial advancement in combating vision impairment worldwide. Diabetic retinopathy (DR) remains a leading cause of vision loss, necessitating innovative diagnostic tools. Our approach combines cutting-edge deep learning frameworks, specifically Convolutional Neural Network (CNN) models and Transformer models, to create a sophisticated system capable of significantly enhancing diagnostic accuracy and efficiency.

The fusion of CNN and Transformer architectures is strategically designed to leverage their respective strengths. CNN models excel in capturing intricate spatial dependencies within medical images, particularly fundus images crucial for DR diagnosis. On the other hand, Transformer models are renowned for their exceptional attention mechanisms, which can discern subtle features vital for accurate classification.

The synergy achieved by integrating these models results in a hybrid system with superior performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score. Our model underwent extensive simulation and validation on diverse datasets to ensure robustness and generalizability across different scenarios. The outcomes of this rigorous testing demonstrate its reliability and effectiveness compared to traditional methods.

.

|  |  |  |
| --- | --- | --- |
| **CHAPTER**  **NO.** | **TITLE** | **PAGE NO.** |
|  | ***BONAFIDE CERTIFICATE*** | ***i*** |
|  | ***ACKNOWLEDGEMENT*** | ***ii*** |
|  | ***ABSTRACT*** | ***iii*** |
|  | ***LIST OF FIGURES*** | ***vi*** |
|  | ***LIST OF ABBREVIATIONS*** | ***vii*** |
| **1** | **INTRODUCTION** | **1** |
|  | 1.1 Prelude | 1 |
|  | 1.2 Motivation | 2 |
|  | 1.3 Literature Survey | 3 |
|  | 1.4 Objective | 4 |
|  | 1.5 Organization of the reports | 4 |
| **2** | **MEDICAL IMAGE CLASSIFICATION USING DEEP LEARNING** | **5** |
|  | 2.1 Prelude | 5 |
|  | 2.2 Medical Image Classification | 5 |
|  | 2.3 Fundus Images | 6 |
|  | 2.4 Diabetic Retinopathy and its Stages | 7 |
|  | 2.5 Why Deep Learning | 8 |
|  | 2.6 Deep Learning | 9 |
|  | 2.7 Image Classification using Deep Learning | 10 |
|  | 2.8 Transfer Learning | 11 |
|  | 2.9 Hybrid Model | 12 |
|  | 2.10 Ensemble Learning | 13 |
|  | 2.11 Evaluation Metrics | 14 |
| **3** | **PROPOSED SYSTEM** | **16** |

|  |  |  |
| --- | --- | --- |
|  | 3.1 Prelude | 16 |
|  | 3.2 Existing Technique | 16 |
|  | 3.3 Dataset (APTOS2019) | 17 |
|  | 3.4 Software and Frameworks Used | 17 |
|  | 3.5 Data Preprocessing | 19 |
|  | 3.6 Conventional Neural Network | 19 |
|  | 3.7 CNN Model- ResNet50 | 22 |
|  | 3.8 Vision Transformer | 24 |
|  | 3.9 Steps Involved in Vision Transformer | 25 |
|  | 3.10 Hybrid Model | 28 |
| **4** | **CODE AND OUTPUT** | **30** |
|  | 4.1 Evaluation of Dataset | 30 |
|  | 4.2 Data Pre-Processing | 31 |
|  | 4.3 Training of ResNet50 | 35 |
|  | 4.4 Training Of ViT | 37 |
|  | 4.5 Training Of Ensemble Model | 42 |
|  | 4.6 Prediction | 45 |
| **5** | **RESULT, CONCLUSION AND FUTURE SCOPE** | **46** |
|  | 5.1 Results | 46 |
|  | 5.2 Conclusion | 47 |
|  | 5.3 Future scope | 47 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| 2.1 | Fundus Image | 7 |
| 2.2 | Stages Of Diabetic Retinopathy | 8 |
| 3.1 | Architecture of CNN Model | 23 |
| 3.2 | Architecture of ViT Model | 25 |
| 3.3 | Patch embedding of input image | 25 |
| 3.3 | Architecture of Hybrid Model | 29 |
| 4.1 | Pie Graph of Dataset | 30 |
| 4.2 | Ba Bar Graph of 5 classes in Train set | 31 |
| 4.3 | Input Image | 31 |
| 4.4 | Cropped and Grayscale converted Input Image | 32 |
| 4.5 | Dark Spots Enhanced Input Image | 33 |
| 4.6 | Augmentation of Input Image | 34 |
| 4.7 | Training of ResNet50 Model | 35 |
| 4.8 | Accuracy vs Epochs & Loss vs Epochs Graph | 36 |
| 4.9 | Trai Training of ViT | 40 |
| 4.10 | Accuracy vs Epochs & Loss vs Epochs Graph | 41 |
| 4.11 | Training of Hybrid Model | 42 |
| 4.12 | Accuracy vs Epochs & Loss vs Epochs Graph | 43 |
| 4.13 | Confusion Matrix | 45 |
| 4.14 | Prediction of DR | 45 |
|  |  |  |

**LIST OF ABBREVIATIONS**

* DR Diabetic Retinopathy
* CNN Convolutional Neural Network
* ViT Vision Transformer
* ResNet50 Residual Network
* GNN Graph Neural Network
* LCA Layer-wise Class token Attention
* GCNN Graph Convolutional Neural Network
* AI Artificial Intelligence
* RNN Recurrent Neural Network
* GELU Gaussian Error Linear Unit
* RELU Rectified Linear Unit
* ELU Exponential Linear Unit
* APTOS [Asia Pacific Tele-Ophthalmology Society](https://www.kaggle.com/c/aptos2019-blindness-detection/overview/aptos-2019)
* VGG16 Visual Geometry Group 16
* IEEE Institute of Electrical and Electronics Engineers

**CHAPTER 1**

**INTRODUCTION**

**1.1 PRELUDE**

Diabetic retinopathy (DR) stands as a significant concern within the realm of ophthalmology and public health, representing a leading cause of vision impairment and blindness worldwide. This progressive diabetic complication primarily affects individuals with diabetes mellitus, particularly those with poorly controlled blood sugar levels over time. The condition manifests as damage to the blood vessels within the retina, the light-sensitive tissue located at the back of the eye crucial for vision.

As diabetes persists and progresses, the delicate blood vessels supplying the retina become compromised, leading to various stages of diabetic retinopathy. These stages can range from mild nonproliferative retinopathy, characterized by microaneurysms and small hemorrhages, to more advanced proliferative retinopathy, marked by the growth of abnormal blood vessels that can leak and cause retinal detachment if left untreated.

This project introduces a groundbreaking Hybrid Model for Diabetic Retinopathy Classification, a crucial advancement in combating vision impairment worldwide. Diabetic retinopathy (DR) remains a leading cause of vision loss, necessitating innovative diagnostic tools. Our approach combines cutting-edge deep learning frameworks, specifically Convolutional Neural Network (CNN) models and Transformer models, to create a sophisticated system capable of significantly enhancing diagnostic accuracy and efficiency.

The fusion of CNN and Transformer architectures is strategically designed to leverage their respective strengths. CNN models excel in capturing intricate spatial dependencies within medical images, particularly fundus images crucial for DR diagnosis. On the other hand, Transformer models are renowned for their exceptional attention mechanisms, which can discern subtle features vital for accurate classification.

The synergy achieved by integrating these models results in a hybrid system with superior performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score. Our model underwent extensive simulation and validation on diverse datasets to ensure robustness and generalizability across different scenarios. The outcomes of this rigorous testing demonstrate its reliability and effectiveness compared to traditional methods.

**1.2 MOTIVATION**

Early detection plays a pivotal role in combating the devastating impact of diabetic retinopathy, a leading cause of blindness among working-age adults. Classification algorithms enable the timely identification of retinopathy, facilitating prompt interventions that can significantly mitigate the risk of vision loss. This emphasis on early detection underscores the importance of leveraging advanced classification models to enhance healthcare outcomes.

Furthermore, the use of classification models offers a pathway to personalized treatment strategies. By categorizing individuals based on the severity of their condition, these models enable healthcare providers to tailor treatment plans according to each patient's unique needs. This personalized approach not only improves patient outcomes but also optimizes resource allocation within healthcare systems, ensuring that interventions are targeted and effective.

In the realm of research, advancements in diabetic retinopathy classification hold immense potential. Continued exploration of classification algorithms can deepen our understanding of disease progression and aid in the discovery of new biomarkers. This scientific progress lays the groundwork for the development of more efficacious treatments and interventions, promising better outcomes for individuals affected by diabetic retinopathy.

Moreover, automated classification systems contribute to improving accessibility to eye care, especially in underserved regions where access to specialized healthcare professionals may be limited. By democratizing healthcare through automated screening tools, we can extend the reach of preventive measures and interventions, ultimately preventing vision loss on a larger scale and promoting inclusive healthcare practices.

**1.3 LITERATURE SURVEY**

The most important step in project development is a literature review. The basic idea of this project is obtained from the following papers:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **TITLE** | **AUTHOR** | **PUBLISHED DETAILS** | **INFERENCE** |
| 1 | Grading of Diabetic Retinopathy Images Based on Graph Neural Network | Meiling feng, Jingyi wang, Kai wen and Jing sun | IEEE Access, vol. 11, pp. 98391-98401 ,September 2023 | Integrating graph neural networks (GNN) with convolutional neural networks (CNN) |
| 2 | A Hybrid Convolutional Neural Network Model for Automatic Diabetic Retinopathy Classification From Fundus Images | Ghulam ali ,Aqsa dastgir ,Muhammad waseem iqbal ,Muhammad anwar ,and Muhammad faheem | IEEE Journal of Translational Engineering in Health and Medicine, vol. 11, pp. 341-350, June 2023 | A novel convolutional neural networks (CNN) model, utilizing Resnet50 and Inceptionv3 features |
| 3 | Classification of Diabetic Retinopathy Disease Levels by Extracting Topological Features Using Graph Neural Networks | Sumod sundar and S. Sumathy | IEEE Access, vol. 11, pp.51435-51444, May 2023. | Graph convolutional neural network (GCNN), incorporating variational autoencoder features, enhances diabetic retinopathy diagnosis |
| 4 | Dual Transformer Encoder Model for Medical Image Classification | F. Yan, B. Yan and M. Pei | IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, pp. 690-694 ,September 2023 | A dual transformer encoder model with variable hidden sizes and a Layer-wise Class token Attention (LCA) classification module improves medical image analysis |
| 5 | CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification | Chun-Fu Richard Chen, Quanfu Fan, Rameswar Panda | IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 347-356, February 2022. | CrossViT, a dual-branch transformer with cross-attention token fusion, excels in multi-scale feature learning for image classification, outperforming vision transformers and efficient CNN models. |

**1.4 OBJECTIVE**

To develop a Hybrid Model for Diabetic Retinopathy(DR) classification, using ensemble model, Vision Transformer and Convolutional Neural Network (ViT-CNN) to enhance diagnostic accuracy and other evaluation metrics, making it more precise at discerning the complexities associated with diabetic retinopathy.

**1.5 ORGANISATION OF THE REPORT:**

**Chapter 1** introduces the project and elaborates on the literature review, motivation, and project objective.

**Chapter 2** provides the information about the techniques of image classification using deep

learning.

**Chapter 3** describes in detail about the design methodology of the proposed work and explains in detail about the working of the proposed work.

**Chapter 4** discusses the code and output gives an insight about the work.

**Chapter 5** discusses the result and concludes the project work.

**CHAPTER 2**

**MEDICAL IMAGE CLASSIFICATION USING DEEP LEARNING**

**2.1 PRELUDE**

The realm of medical diagnostics has been revolutionized by advanced technologies, particularly in medical image classification, crucial in ophthalmology for assessing and managing diabetic retinopathy (DR). DR, a complication of diabetes mellitus, progresses from mild to severe stages, requiring distinct clinical interventions.

Traditional methods for DR classification relied on manual interpretation, prone to subjectivity. However, deep learning, inspired by neural networks, has transformed medical image analysis, offering accuracy and efficiency. Deep learning excels in image classification by learning hierarchical representations and extracting meaningful features, especially in fundus images where subtle changes signify disease progression.

Transfer learning enhances deep learning's effectiveness in medical imaging, adapting learned features to new domains with limited data. Hybrid models integrate deep learning with interpretable models, enhancing clinical interpretability. Ensemble learning further improves model robustness by aggregating predictions from multiple learners.

Evaluation metrics like sensitivity, specificity, accuracy, and F1 score assess model performance, guiding clinical decision-making and patient care.

**2.2 MEDICAL IMAGE CLASSIFICATION**

Medical image classification is a type of artificial intelligence (AI) technique used in medical image analysis.expand\_more It involves automatically categorizing medical images, such as X-rays, MRI scans, and CT scans, into predefined groups. The purpose is to develop computer algorithms that can interpret these images and classify their content to aid in tasks like medical diagnosis, treatment planning, and disease monitoring.

Deep learning, a type of machine learning, is particularly prominent in medical image classification.expand\_more Convolutional neural networks (CNNs) are a kind of deep learning model that excel at image recognition. In medical image classification, CNNs are trained on vast amounts of labeled medical images to identify patterns and relationships within the data. This allows them to categorize new images with a high degree of accuracy.

Medical image classification has the potential to revolutionize healthcare by:

**Assisting medical professionals in diagnosis:** By automatically identifying abnormalities in medical images, AI can aid doctors in making more accurate and timely diagnosis.

**Streamlining treatment planning**: AI can analyze medical images to help determine the most effective course of treatment for a particular patient.

**Improving disease monitoring:** AI can track changes in medical images over time to monitor disease progression and treatment response.

However, there are also challenges associated with medical image classification, such as the need for large amounts of labeled data and the potential for bias in algorithms. Overall, medical image classification is a rapidly developing field with the potential to significantly improve healthcare outcomes.

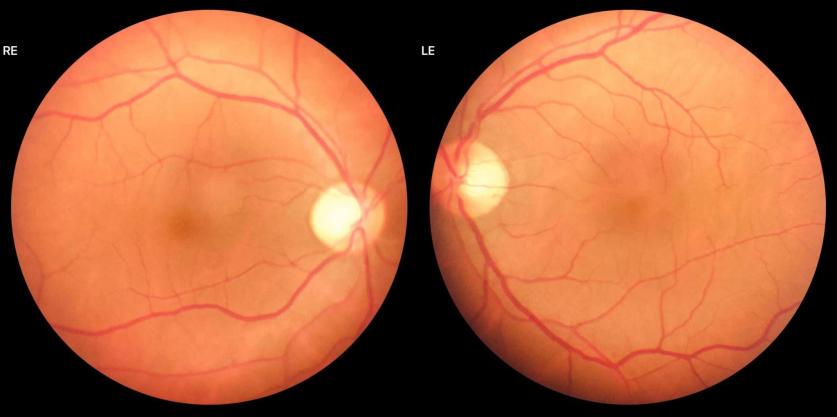
**2.3 FUNDUS IMAGES**

The Main Focus of this project is to correctly classify the Medical Fundus Images according to the severity of Diabetic Retinopathy. Fundus images are photographs of the back of the eye, also known as the fundus. These images are captured using a specialized fundus camera during an ophthalmic exam. The camera uses a bright light and often dilates the pupil to get a clear view of the inner eye.

Fundus images reveal important structures like the retina, optic nerve, and macula. These structures play a vital role in vision, and fundus images allow ophthalmologists to diagnose and monitor various eye conditions, including:

* Diabetic retinopathy
* Age-related macular degeneration
* Glaucoma
* Macular edema

By examining fundus images, doctors can identify abnormalities such as bleeding, swelling, or unusual blood vessel growth. This information is crucial for early detection and treatment of eye diseases, potentially preventing vision loss.



**Figure 2.1 Fundus Image**

**2.4 DIABETIC RETINOPATHY AND ITS STAGES**

Diabetic retinopathy (DR) is a leading cause of blindness in adults, particularly those with diabetes. It's a complication caused by chronic high blood sugar levels damaging the blood vessels in the retina, the light-sensitive layer at the back of the eye. DR progresses through stages, often without noticeable symptoms in the early stages. Early detection and management are crucial to prevent vision loss.

**Understanding the Retina:**

The retina is a complex structure containing millions of photoreceptor cells (rods and cones) that convert light into electrical signals. These signals are then transmitted through the optic nerve to the brain, enabling us to see. The macula, a central region of the retina, is responsible for sharp central vision.

**How Diabetes Affects the Retina:**

* High blood sugar levels weaken and damage the retinal blood vessels. This can lead to:
* Microaneurysms: Weak spots that bulge out in the blood vessels.
* Leakage: Fluid and blood leaking from damaged vessels, causing swelling (macular edema) and vision problems.
* Blockages: Blocked blood vessels deprive the retina of oxygen and nutrients.
* Abnormal new blood vessel growth: In response to insufficient blood flow, the body attempts to grow new blood vessels, which are often weak and fragile.

**Stages of Diabetic Retinopathy:**

DR progresses through five main stages:

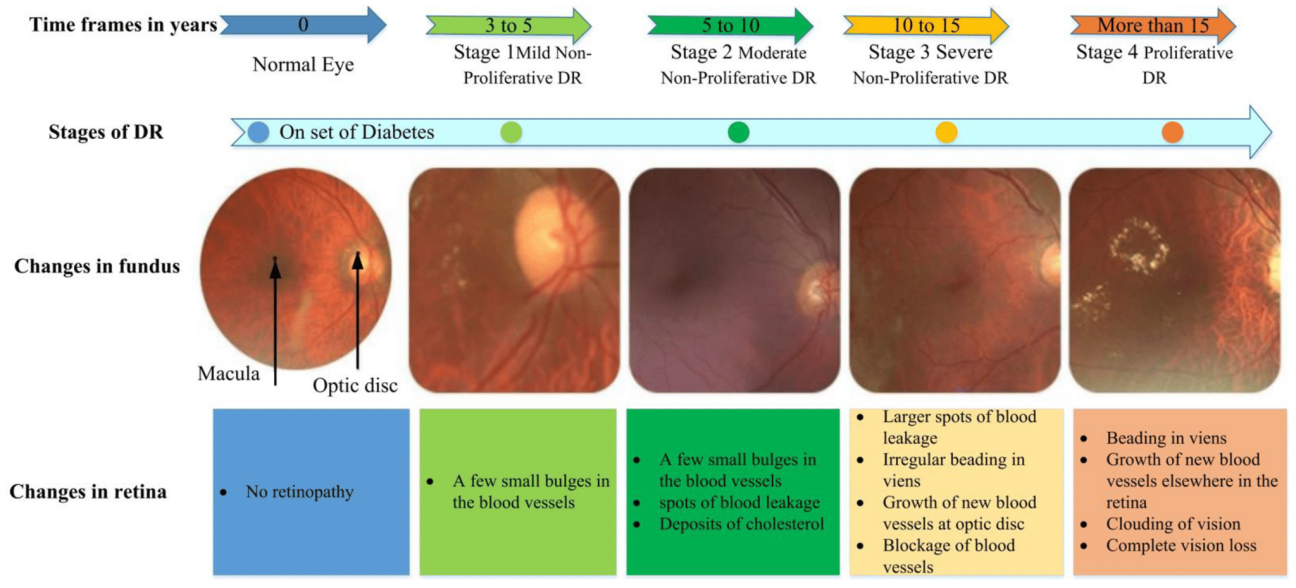
1. **No Apparent Retinopathy(No DR)**: There are no visible Symptoms. The fundus image of the eye looks completely healthy.The blood vessels and retina looks completely normal.

2. **Mild Non-proliferative Diabetic Retinopathy (NPDR)**: This is the early stage of DR with little to no Symptoms. A few small bulges may appear in the blood vessels.

3. **Moderate Non-proliferative Diabetic Retinopathy (NPDR**): More extensive damage compared to mild NPDR.A few small bulges are visible in the fundus images. There are defined spots of blood on the retina and there also exists deposits of cholesterol.

4. **Severe Non-Proliferative Diabetic Retinopathy (PDR)**: A more severe case of DR which can lead to vision loss if untreated. Large spots of blood leakage are visible. Irregular beading in veins is visible. There exists blockage of blood vessels and growth of new blood vessels at the optic disc.

5. **Proliferative Diabetic Retinopathy**: This is the most severe case of DR. The Beadings in veins are profusely visible. The symptoms include clouding of vision. In the fundus images growth of new blood vessels elsewhere in the retina. Complete vision loss may happen in this stage.



**Figure 2.2 Stages of Diabetic Retinopathy**

**2.5 WHY DEEP LEARNING**

Deep learning is favored over traditional machine learning for medical image classification primarily due to its ability to automatically extract intricate features from raw image data. Unlike traditional methods that rely on manual feature engineering, deep learning models, especially convolutional neural networks (CNNs), can learn hierarchical representations directly from images, capturing complex patterns and structures effectively. This automated feature extraction process is particularly beneficial in medical imaging, where images can contain subtle yet critical information for diagnosis and treatment planning.

Furthermore, deep learning excels in handling large and diverse datasets commonly found in medical imaging, allowing for improved generalization and accuracy. The continuous advancements in deep learning architectures, optimization techniques, and the availability of pre-trained models further contribute to its state-of-the-art performance in various image-related tasks.

While traditional machine learning approaches have their merits, such as interpretability and suitability for smaller datasets, they may struggle to capture the nuanced features present in medical images. Therefore, the inherent capability of deep learning to automatically learn relevant features from data makes it a compelling choice for medical image classification, leading to more accurate and reliable diagnostic outcomes in healthcare applications.

**2.6 DEEP LEARNING**

Deep learning has emerged as a transformative technology in medical image classification, revolutionizing how healthcare professionals analyze and interpret imaging data. One of the primary advantages of deep learning in this context is its ability to automatically learn hierarchical features from raw image data. Traditional methods often relied on handcrafted features, which were labor-intensive to design and limited in capturing complex patterns. CNNs, on the other hand, excel at learning intricate features through multiple layers of abstraction, making them highly effective for tasks like lesion detection, organ segmentation, and disease classification.

The application of deep learning in medical image classification has led to significant improvements in diagnostic accuracy and efficiency. For example, CNN-based models have demonstrated remarkable performance in detecting abnormalities in chest X-rays, identifying tumors in MRI scans, and analyzing histopathology images for cancer diagnosis. These advancements not only streamline the diagnostic process but also aid in early disease detection and treatment planning, ultimately benefiting patient outcomes.

Despite its successes, deep learning in medical imaging poses unique challenges. One major concern is the need for large annotated datasets, which are often scarce and costly to acquire. Additionally, ensuring model interpretability and generalization across diverse patient populations and imaging modalities remains a critical area of research.

To address these challenges, researchers have developed techniques such as transfer learning, data augmentation, and model explainability methods. Transfer learning enables leveraging pre-trained models on large datasets to improve performance on smaller medical datasets. Data augmentation techniques artificially increase the diversity of training data, mitigating issues related to dataset scarcity. Model explainability methods help clinicians understand and trust model predictions, enhancing the adoption of deep learning systems in clinical practice.

In conclusion, deep learning has significantly advanced medical image classification by automating feature extraction, improving diagnostic accuracy, and enabling new avenues for research and innovation in healthcare. Continued research and collaboration between AI experts and healthcare professionals will further propel the integration of deep learning into routine clinical workflows, ultimately benefiting patients worldwide.

**2.7 IMAGE CLASSIFICATION USING DEEP LEARNING**

Image classification using deep learning is revolutionizing the way computers understand and categorize visual content. By mimicking the human brain's structure and function, deep learning models excel at automatically identifying objects, scenes, and actions within images. This technology underpins numerous applications, from self-driving cars and facial recognition to medical image analysis and autonomous robots.

**Deep Learning Unveils Hidden Patterns**: Traditional machine learning approaches often require manual feature engineering, a process of hand-picking relevant characteristics from the data. Deep learning takes a different approach. Deep neural networks, the workhorses of deep learning, consist of multiple interconnected layers of artificial neurons. These layers progressively extract features from the raw image data, starting with low-level edges and lines to more complex shapes and objects in higher layers. This hierarchical learning allows the model to capture intricate patterns and relationships within the data, ultimately enabling accurate image classification.

**Convolutional Neural Networks (CNNs)**: The Champs of Image Recognition: Among deep learning models, Convolutional Neural Networks (CNNs) reign supreme in image classification tasks. CNNs are specifically designed to process grid-like data like images. They utilize specialized convolutional layers that efficiently extract features by learning to identify patterns across local regions of the image. These layers are often followed by pooling layers that downsample the data, reducing its dimensionality and computational cost. Through this process, CNNs progressively transform the image into a high-level representation that captures the essence of the content.

**The Deep Learning Image Classification Workflow**:

**Data Preparation:** Images are collected, pre-processed (resized, normalized), and labeled with their corresponding categories.

**Model Architecture Design:** A CNN architecture is chosen or designed, specifying the number and type of convolutional, pooling, and fully-connected layers.

**Model Training:** The model is trained on a large dataset of labeled images. During training, the model adjusts the weights and biases of its connections based on the comparison between its predictions and the correct labels. This iterative process minimizes the error and allows the model to learn effective feature representations.

**Evaluation and Testing:** The trained model's performance is evaluated on a separate test dataset to assess its generalization ability, meaning its accuracy on unseen images.

**Fine-tuning (Optional):** Pre-trained models, CNNs trained on massive datasets for general image recognition, can be leveraged as a starting point. These models can be fine-tuned for specific tasks by retraining them on a smaller dataset tailored to the desired classification problem.

**2.8 TRANSFER LEARNING**

Transfer learning is a pivotal technique in deep learning that involves leveraging knowledge from a pre-trained model to enhance the training of another model for a related task or domain. This approach has gained significant popularity due to its ability to address challenges such as limited labeled data and resource-intensive training processes.

At its core, transfer learning exploits the features and representations learned by a model during its training on a large and diverse dataset, often referred to as the source domain. These learned features, which capture general patterns and structures in the data, are then transferred and fine-tuned to a new model targeting a specific task or dataset, known as the target domain. By doing so, transfer learning offers several advantages.

Firstly, it accelerates the training process by providing a head start to the new model, as it doesn't have to learn low-level features from scratch. This significantly reduces the computational resources and

time required for training, making it feasible to train deep learning models on smaller datasets or in resource-constrained environments.

Secondly, transfer learning enhances the generalization capability of the model. The pre-trained model's knowledge serves as a strong initial representation that captures a broad understanding of the data, allowing the new model to better adapt and generalize to new and unseen data instances in the target domain.

Additionally, transfer learning facilitates domain adaptation, where the source and target domains may exhibit different characteristics or distributions. Fine-tuning the pre-trained model on the target domain helps in aligning the learned representations to better suit the nuances and complexities of the new data, leading to improved performance and accuracy.

Overall, transfer learning plays a pivotal role in overcoming data scarcity, reducing training complexity, improving generalization, and enabling efficient model adaptation across various deep learning applications, making it a fundamental technique in the field.

**2.9 HYBRID MODEL**

The decision between using a single deep learning model or a hybrid model hinges on several factors that influence model performance, complexity, and interpretability. A hybrid model combines the strengths of multiple models or modalities, offering a versatile approach to address complex tasks and leverage diverse data sources effectively.

A single model approach is straightforward and effective for tasks with well-defined objectives and homogeneous data. It is particularly suitable when computational resources are limited or when the task can be adequately addressed by a specific model architecture. Single models are easier to train, interpret, and deploy, making them preferable for scenarios where simplicity and efficiency are prioritized over intricate modeling.

On the other hand, hybrid models introduce a level of flexibility and capability to handle more complex scenarios. By combining different models or modalities, such as integrating convolutional neural networks (CNNs) with recurrent neural networks (RNNs) or fusing information from multiple sensors in multimodal tasks, hybrid models can capture diverse aspects of the data and extract complementary features. This approach is beneficial for tasks involving diverse data sources, transfer learning across domains, or feature fusion to enhance model performance.

The choice between a single model and a hybrid model depends on the specific requirements of the task at hand. For straightforward tasks with homogeneous data and limited computational resources, a single model approach may suffice. However, for tasks involving diverse data sources, complex patterns, or the need for enhanced performance through feature fusion and model combination, a hybrid model can provide a more comprehensive and effective solution. Striking a balance between model complexity, interpretability, and performance goals is key when deciding between these two approaches in deep learning applications.

**2.10 ENSEMBLE LEARNING**

Ensemble learning is a powerful technique that can be applied to hybrid models in deep learning to further improve performance, robustness, and generalization.

Hybrid models combine multiple models or modalities to leverage their complementary strengths, such as combining convolutional neural networks (CNNs) with recurrent neural networks (RNNs) in image captioning tasks or integrating information from different sensors in multimodal applications. Ensemble learning takes this idea a step further by combining predictions from multiple individual models, often referred to as base learners, to produce a final prediction that is more accurate and reliable than any single model's prediction.

Model Diversity: Ensemble learning works best when the individual models (base learners) in the ensemble are diverse yet accurate. In the context of hybrid models, each base learner can represent a different aspect or perspective of the data, enhancing overall model diversity.

Reduced Overfitting: By combining predictions from multiple models, ensemble learning helps reduce overfitting, especially in complex models like hybrid models. It promotes generalization by leveraging the collective wisdom of diverse models, thereby improving performance on unseen data.

Improved Robustness: Ensemble learning enhances model robustness by mitigating the impact of individual model biases or errors. Even if some base learners make incorrect predictions, the ensemble's combined prediction tends to be more reliable and robust.

Boosted Performance: Ensemble learning often leads to boosted performance compared to individual models, especially when base learners are well-trained and diverse. It can push the performance boundaries of hybrid models, achieving higher accuracy and better results across different tasks and datasets.

Common ensemble learning techniques include bagging (Bootstrap Aggregating), boosting (e.g., AdaBoost, Gradient Boosting), and stacking, each with its approach to combining base learners' predictions. These techniques can be applied to hybrid models by treating each component model or modality as a base learner in the ensemble.

**2.11 EVALUATION METRICS**

Evaluation metrics play a crucial role in assessing the performance of deep learning models, particularly in image classification tasks.These evaluation metrics are essential in understanding how well a deep learning model performs in classifying images and are often used together to gain comprehensive insights into the model's strengths and weaknesses, especially in scenarios with multiple classes or imbalanced datasets.

**1.Accuracy:** Accuracy measures the proportion of correctly classified instances among all instances in the dataset. It is calculated as the ratio of correct predictions to the total number of predictions made by the model.

**Accuracy= [True Positives+True Negatives]/**

**[(True Positives+True Negatives+False Positives+False Negatives]**

**2. Precision:** Precision measures the proportion of true positive predictions among all positive predictions made by the model. It focuses on the correctness of positive predictions and is calculated as the ratio of true positives to the sum of true positives and false positives.

**Precision= True Positives/(True Positives+False Positives)**

**3. Recall (Sensitivity or True Positive Rate):** Recall measures the proportion of true positive predictions among all actual positive instances in the dataset. It focuses on the model's ability to correctly identify positive instances and is calculated as the ratio of true positives to the sum of true positives and false negatives.

**Sensitivity=True Positives/(True Positives+False Negatives)**

**4. F1 Score:** The F1 score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It combines both precision and recall into a single metric and is particularly useful when dealing with imbalanced datasets where one class dominates over the other.

**Recall= True Positives/(True Positives+False Negatives)**

**F1 Score=2× [(Precision+Recall)/(Precision×Recall)]**

**5. Confusion Matrix:** A confusion matrix is a table that summarizes the performance of a classification model by presenting the counts of true positive, true negative, false positive, and false negative predictions. It provides a detailed view of the model's performance across different classes and aids in calculating metrics like accuracy, precision, recall, and F1 score.

**CHAPTER-3**

**PROPOSED SYSTEM**

**3.1 PRELUDE**

The existing system in diabetic retinopathy (DR) classification has made significant strides by leveraging a combination of Convolutional Neural Networks (CNN) and Graph Neural Networks (GNN). This hybrid architecture has shown promise in automating DR detection from fundus images. However, our proposed system aims to further enhance these capabilities using the APTOS2019 dataset, known for its comprehensive and annotated fundus images representing various DR stages.

To achieve this, we utilize industry-leading frameworks such as PyTorch and TensorFlow, along with the collaborative platform Google Colab, for seamless development and experimentation. Data pre-processing plays a crucial role in optimizing image quality and feature standardization, ensuring robust model performance.

Our methodology incorporates CNN ResNet50 and Vision Transformer architectures for feature extraction and representation learning. These models are renowned for their effectiveness in capturing intricate features crucial for DR diagnosis. By combining these techniques, our hybrid model aims to improve diagnostic accuracy and generalization across diverse datasets.

The proposed system not only enhances feature extraction but also improves interpretability and model transparency, critical factors in medical diagnostics. This approach is expected to contribute significantly to the field of ophthalmic diagnostics, offering scalable and adaptable solutions for DR classification.

**3.2 EXISTING SYSTEM**

The existing system utilizes the robust relationship-capturing capabilities of graph neural networks (GNNs) and introduces a novel DR (Diabetic Retinopathy) intelligent classification model. This model is composed of two cascaded networks: a convolutional neural network (CNN) to extract deep features from DR images and a GNN to further capture relationships among these features. Subsequently, the outputs from these networks are fused using adaptive weighting to provide the overall grading result.

This system has been evaluated on the APTOS2019 dataset, demonstrating strong performance metrics: accuracy (85%), precision (82%), recall (85%), and F1 score (84%). These results underscore the effectiveness of the model in accurately classifying DR images, highlighting its potential to enhance diagnostic capabilities in the medical domain.

**3.3 DATASET (APTOS2019)**

Millions of people suffer from [diabetic retinopathy](https://nei.nih.gov/health/diabetic/retinopathy), the leading cause of blindness among working aged adults. Aravind Eye Hospital in India hopes to detect and prevent this disease among people living in rural areas where medical screening is difficult to conduct. Further, the solutions will be spread to other Ophthalmologists through the [4th Asia Pacific Tele-Ophthalmology Society (APTOS) Symposium](https://www.kaggle.com/c/aptos2019-blindness-detection/overview/aptos-2019).  
 The APTOS2019 dataset, also known as the "APTOS Blindness Detection" dataset, is a widely used benchmark dataset in the field of ophthalmology and medical image analysis, specifically for the task of diabetic retinopathy (DR) detection. It was released as part of a Kaggle competition hosted by the American Society of Ophthalmic and Reconstructive Surgery (ASORS).

The dataset comprises high-resolution retinal images obtained using fundus photography. These images are categorized into different severity levels of diabetic retinopathy, ranging from 0 (no DR) to 4 (severe DR with proliferative diabetic retinopathy). The primary goal of the dataset is to facilitate the development and evaluation of machine learning and deep learning models for automated DR detection and grading. The dataset contains a diverse collection of retinal images captured using various imaging devices, reflecting real-world clinical scenarios. Each image in the dataset is labeled with its corresponding DR severity level, enabling supervised learning tasks for classification and regression. The dataset is relatively large, containing thousands of images, which is crucial for training robust and generalizable machine learning models.

**EVALUATION OF DATASET:** The dataset consists of a total of 2930 fundus image samples for training and a total of 366 fundus image samples for testing, each belonging to five classes.

**3.4 SOFTWARE AND FRAMEWORKS USED**

This project utilized a suite of powerful software tools and frameworks to develop and implement our deep learning-based solution for diabetic retinopathy detection and grading. The key software components used are as follows:

**1. TensorFlow:**

We leveraged TensorFlow, an open-source machine learning framework developed by Google, for building and training deep neural network models. TensorFlow provides a rich set of tools and APIs that facilitate the development of complex deep learning architectures. Its flexibility and scalability make it a popular choice for deep learning projects.

**2. PyTorch:**

PyTorch, another open-source deep learning framework, was also instrumental in this project. Known for its dynamic computational graph feature and intuitive API, PyTorch enabled us to experiment with various model architectures, implement custom layers, and optimize training processes efficiently. We leveraged PyTorch's capabilities to fine-tune pre-trained models, perform data augmentation, and handle complex data pipelines.

**3. OpenCV (Open Source Computer Vision Library):**

OpenCV played a crucial role in image preprocessing, manipulation, and feature extraction tasks within this project. As a powerful computer vision library, OpenCV provided us with a wide range of functions and algorithms for image processing, such as image resizing, color space conversions, edge detection, and feature extraction. These capabilities were essential for preparing the retinal images and extracting relevant features before feeding them into our deep learning models.

**4. Google Colab:**

Google Colab served as our cloud-based development environment for training deep learning models and running experiments. With its integration with Google Drive and access to GPU and TPU resources, Google Colab provided us with a scalable and cost-effective platform for prototyping, training models at scale, and collaborating with team members. Its Jupyter Notebook interface allowed us to write and execute code seamlessly while leveraging Google's infrastructure for computation-intensive tasks.

By harnessing the capabilities of TensorFlow, PyTorch, OpenCV, and Google Colab, able to build and deploy a robust deep learning solution for diabetic retinopathy classification, leveraging state-of-the-art techniques and tools in the field of machine learning and computer vision.

**3.5 DATA PRE-PROCESSING:**

Data preprocessing is a crucial step in preparing raw data for machine learning and deep learning tasks. It involves transforming and organizing data into a format that is suitable for training and analysis. Data preprocessing refers to the cleaning, transformation, and organization of raw data to make it suitable for machine learning algorithms. It involves various techniques such as data cleaning, feature scaling, encoding categorical variables, handling missing values, and more. Preprocessing helps improve the quality of data, reduces noise, and enhances the performance and accuracy of machine learning models.

**BGR to Grayscale Conversion:** Color images captured using fundus cameras are represented in Blue-Green-Red (BGR) channels. Converting these color images to grayscale simplifies the data, reduces dimensionality, and retains essential information for diagnosis. Grayscale images combine the three color channels into a single channel, eliminating color variations that may not be relevant for identifying DR-related features.

**Image Cropping to Target Size (224x224):** Standardizing image dimensions to a fixed size (e.g., 224x224 pixels) is crucial for consistency during model training. Cropping or resizing images to a uniform size facilitates batch processing and ensures the model learns consistent features across all samples.

**Image Augmentation:** Image augmentation is used to increase the diversity of the training dataset. It involves applying transformations such as rotation, flipping, zooming, shifting, brightness, and contrast adjustments to create variations of existing images. Augmentation helps prevent overfitting, improves model generalization, and enhances the robustness of the trained model.

**Lightning Effect for Dark Spots Enhancement:** In retinal fundus images, dark spots such as lesions, microaneurysms, and hemorrhages are critical indicators of diabetic retinopathy. Enhancing the visibility of these dark spots through contrast enhancement techniques improves the accuracy of diagnosis. Techniques like adaptive histogram equalization are used to enhance dark spots, contributing to early detection and better patient care**.**

**3.6 CONVOLUTIONAL NEURAL NETWORK**

A Convolutional Neural Network (CNN) is a type of Deep Learning neural network architecture commonly used in Computer Vision. Computer vision is a field of Artificial Intelligence that enables a computer to understand and interpret the image or visual data.

When it comes to Machine Learning, Artificial Neural Networks perform really well. Neural Networks are used in various datasets like images, audio, and text. Different types of Neural Networks are used for different purposes, for example for predicting the sequence of words we use Recurrent Neural Networks more precisely an LSTM, similarly for image classification we use Convolution Neural networks. In this blog, we are going to build a basic building block for CNN.

These are the the layers commonly used in CNN architectures:

**Convolutional Layer:** The convolutional layer is the fundamental building block of CNNs. It applies convolution operations to the input data using learnable filters (kernels) to extract spatial features. These filters slide across the input image, performing element-wise multiplications and summations to produce feature maps. Convolutional layers help capture local patterns and spatial hierarchies in the input data.

**Activation function:** The activation function introduces non-linearity into the network by applying an activation function to the output of the convolutional layer. For example,Rectified Linear Unit (ReLU) is a commonly used activation function that sets negative values to zero and keeps positive values unchanged. ReLU activation helps introduce non-linearities, making the network capable of learning complex relationships and improving gradient flow during training.

**Pooling Layer (Max Pooling or Average Pooling):** Pooling layers are used to downsample feature maps spatially, reducing computational complexity and controlling overfitting. Max pooling and average pooling are two common pooling operations. Max pooling retains the maximum value within each pooling window, emphasizing dominant features. Average pooling computes the average value within each window, providing a smoothed representation of features.

**Fully Connected (Dense) Layer:** Fully connected layers, also known as dense layers, are traditional neural network layers where each neuron is connected to every neuron in the previous layer. These layers help in learning global patterns and making final predictions. In classification tasks, a softmax activation function is often used in the final dense layer to produce class probabilities.

**Dropout Layer:** Dropout is a regularization technique used to prevent overfitting in deep learning models. During training, dropout layers randomly deactivate a fraction of neurons, forcing the network to learn more robust and generalized features. Dropout helps improve model generalization and reduces the risk of memorizing noise in the training data.

**Batch Normalization Layer:** Batch normalization is a technique used to stabilize and accelerate training in deep neural networks. It normalizes the activations of each layer by adjusting and scaling them using the mean and variance of the mini-batch. Batch normalization helps in mitigating issues like vanishing or exploding gradients and improves the convergence speed of the network.

These layers, along with appropriate configurations and hyperparameters, collectively form the architecture of a CNN. By stacking and connecting these layers in a sequential manner, CNNs can effectively learn hierarchical representations of visual data and perform tasks like image classification with high accuracy.

Pretrained models and transfer learning are powerful techniques in deep learning that leverage pre-existing knowledge from large datasets to improve the performance of models on new, related tasks.

**1. Pretrained Models:** Pretrained models are deep learning models that have been trained on large-scale datasets, such as ImageNet, using vast computational resources. These models have learned rich and generalized features from diverse data and are capable of recognizing a wide range of patterns and objects within images. Pretrained models are often available for various architectures like ResNet, VGG, Inception, and more.

Using pretrained models saves significant time and computational resources that would otherwise be required for training a deep learning model from scratch. Pretrained models capture general knowledge about features in images, which can be transferred and fine-tuned for specific tasks, leading to faster convergence and better performance.

**2. Transfer Learning:** Transfer learning is a technique that involves taking knowledge learned from a pretrained model and transferring it to a new, related task or domain. In transfer learning, the pretrained model's weights and learned features are used as a starting point for training a new model on a smaller or different dataset. Transfer learning can be achieved through feature extraction or fine-tuning.

Transfer learning addresses challenges such as limited labeled data and computational constraints by leveraging knowledge from large pretrained models. It improves model generalization, reduces overfitting, and speeds up the training process, especially in scenarios where training from scratch is impractical or inefficient.

By combining pretrained models with transfer learning, practitioners can build and deploy deep learning models more effectively for specific tasks. For example, in image classification tasks, a pretrained model like ResNet or VGG can be fine-tuned on a medical imaging dataset for diagnosing diseases like diabetic retinopathy, leveraging the learned features and optimizing the model's performance for the new medical domain.

In this project, the following pre-trained models have been tried for the CNN stage.

**ResNet50 (Residual Network):** ResNet50 is a deep convolutional neural network architecture that was introduced by Microsoft Research. It is part of the ResNet family of models known for introducing the concept of residual learning, which addresses the degradation problem in very deep networks. ResNet50 specifically has 50 layers and is designed to learn high-level features from images. It includes shortcut connections (skip connections) that enable the network to bypass certain layers, allowing for smoother gradient flow during training.

**VGG16 (Visual Geometry Group 16-layer):** VGG16 is a convolutional neural network architecture developed by the Visual Geometry Group at Oxford University. It is characterized by its simplicity and uniform architecture, consisting of 16 layers with trainable parameters. VGG16 follows a stack of convolutional layers, interspersed with max-pooling layers, and culminating in fully connected layers. Despite its depth, VGG16's straightforward architecture makes it easy to understand and implement.

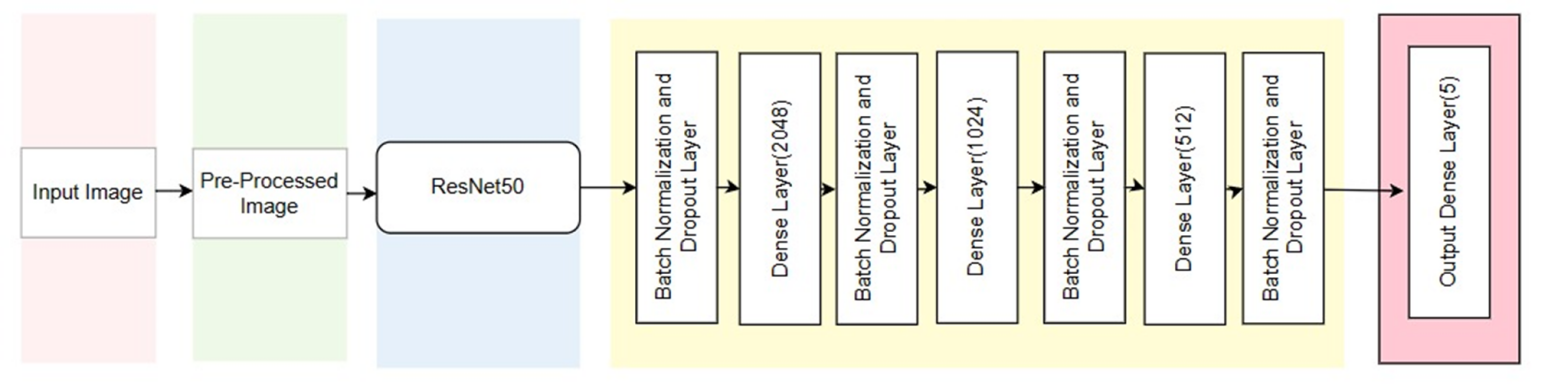
**EfficientNet:** EfficientNet is a family of convolutional neural network architectures developed by Google AI. It introduces a novel compound scaling method that balances model depth, width, and resolution to achieve better performance with fewer parameters. EfficientNet models are designed to be computationally efficient while maintaining high accuracy across various tasks. They achieve state-of-the-art performance on image classification benchmarks with significantly fewer parameters compared to other architectures.

**DenseNet (Densely Connected Convolutional Networks):** DenseNet is a convolutional neural network architecture that emphasizes dense connectivity between layers. In DenseNet, each layer is connected to every other layer in a feed-forward fashion, resulting in densely connected blocks. This dense connectivity promotes feature reuse, reduces vanishing gradient problems, and encourages feature propagation throughout the network. DenseNet architectures typically consist of dense blocks followed by transition layers to control model complexity and size.

Out of these 4 pre-trained models, based on multiple trial and errors, ResNet50 has given better results and better accuracy for the ensemble approach compared to other pre-trained models. ResNet50 also required a lesser number of parameters to train, compared to the other models for the same results.

**3.7 CNN MODEL - ResNet50**

In this project, the CNN part of the ensemble model consists of a few custom layers on top of the pre-trained ResNet50 model.



**Figure 3.1 Architecture of CNN Model**

The Steps involved in this CNN stage are:

**Input Image:** The input image is passed onto the next layer for pre-processing.

**Image Pre-processing:** The input image is preprocessed to an input size of 224x224, so that it is compatible with the next layer. The images are augmented at this step by extending, rotating, flipping to make the model robust to the input image. This step is done to ensure that the model does not memorize the input training images.

**ResNet50:** The 50-layer ResNet architecture includes the following elements:

1. A 7×7 kernel convolution alongside 64 other kernels with a 2-sized stride.
2. A max pooling layer with a 2-sized stride.
3. 9 more layers—3×3,64 kernel convolution, another with 1×1,64 kernels, and a third with 1×1,256 kernels. These 3 layers are repeated 3 times.
4. 12 more layers with 1×1,128 kernels, 3×3,128 kernels, and 1×1,512 kernels, iterated 4 times.
5. 18 more layers with 1×1,256 cores, and 2 cores 3×3,256 and 1×1,1024, iterated 6 times.
6. 9 more layers with 1×1,512 cores, 3×3,512 cores, and 1×1,2048 cores iterated 3 times.(up to this point the network has 50 layers)
7. Average pooling, followed by a fully connected layer with 1000 nodes, using the softmax activation function

In this project, the last softmax layer is removed from the ResNet50 architecture and a custom set of layers are added on top to it.

**Custom layers:** Followed by the ResNet50, are a layer of Batch Normalization and Dropout immediately followed by a dense fully connected layer of neurons of number 2048.

This is again followed by a layer of Batch Normalization and Dropout immediately followed by a dense fully connected layer of neurons of number 1024.

And again this is followed by a layer of Batch Normalization and Dropout immediately followed by a dense fully connected layer of neurons of number 512.

**Output Layer:** Finally there is a layer of 5 neurons with softmax activation function pertaining to the five classes of the image classification.

**3.8 VISION TRANSFORMER:**

Vision Transformers (ViTs) have been chosen for this project due to their cutting-edge capabilities and their status as a trending topic in the field of deep learning and computer vision. In recent years, ViTs have garnered significant attention and acclaim for their remarkable performance in image classification tasks, including those involving medical imaging such as diabetic retinopathy (DR) detection.

The decision to utilize ViTs stems from several key advantages that align with the project's objectives and requirements:

**1. Global Context and Long-Range Dependencies:**

ViTs excel in capturing global context and understanding long-range dependencies within images. This is crucial in medical image analysis, where detecting subtle patterns and considering the overall context can significantly impact diagnostic accuracy, especially in conditions like DR.

**2. Adaptability to Different Resolutions and Modalities:**

Unlike traditional convolutional neural networks (CNNs) that may require specific input resolutions, ViTs are less sensitive to image resolution variations. This adaptability makes them well-suited for handling diverse medical imaging modalities and varying image qualities commonly encountered in real-world medical datasets.

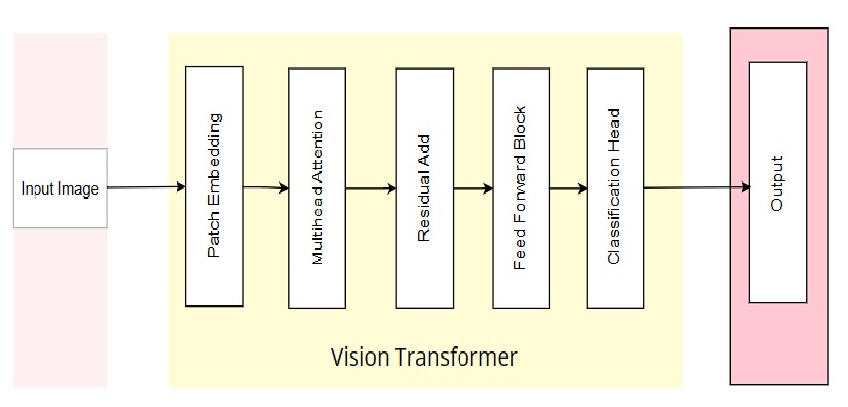
**3. State-of-the-Art Performance and Trending Advancements:**

ViTs have demonstrated state-of-the-art performance in various image classification benchmarks and are actively researched and improved upon by the deep learning community. Leveraging cutting-edge techniques aligns with the project's goal of employing the most advanced tools and methodologies available.

**4. Generalization and Transfer Learning:**

Pretrained ViT models, often trained on large-scale datasets like ImageNet, offer a wealth of learned features and representations that can be fine-tuned for specific tasks. This facilitates transfer learning, where the model can leverage pre-existing knowledge and adapt it to the nuances of medical image classification, potentially reducing the need for extensive labeled medical datasets.

By leveraging the strengths of ViTs and their current prominence in the deep learning landscape, the project aims to achieve not only high accuracy in DR classification but also contribute to the ongoing advancements and exploration of state-of-the-art techniques in medical image analysis.

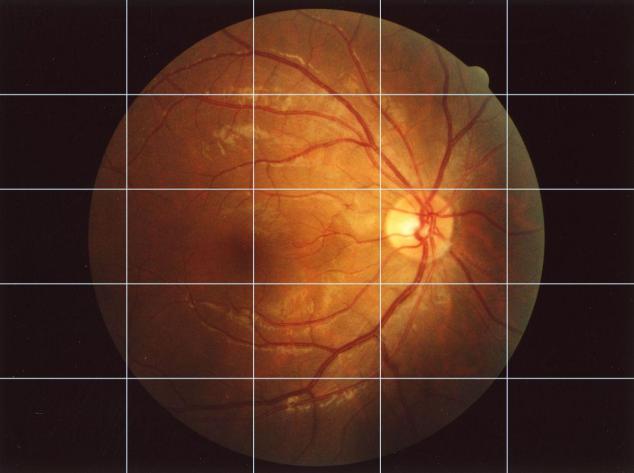


**Figure 3.2 Architecture of ViT Model**

**3.9 STEPS INVOLVED IN VISION TRANSFORMER:**

**1.PATCH EMBEDDING:**

In DR detection using ViTs, patch embedding refers to breaking down retinal fundus images into smaller patches or regions. Each patch is then transformed into numerical embeddings, capturing local features such as lesions, hemorrhages, or microaneurysms. Patch embedding allows the ViT model to analyze detailed information from different parts of the retina and learn discriminative features relevant to DR diagnosis.



**Figure 3.3 patch embedding of input image**

**Image Segmentation into Patches:** The input image, which is 224x224 pixels, is divided into non-overlapping patches each of size 16x16 pixels. Given that the full image dimension is divisible evenly by the patch size, this results in (224 / 16) x (224 / 16) = 14 x 14 = 196 patches per image.

**Patch Conversion:** Each of these 16x16 patches contains 16x16x3 = 768 pixel values (since each pixel has 3 channels: Red, Green, and Blue). These patches are then linearly transformed into a higher-dimensional space through a trainable embedding matrix.

**Patch Embedding:** The transformation typically involves flattening each 16x16x3 patch into a 1-dimensional vector of 768 elements and then multiplying this vector by a trainable matrix to obtain an embedded vector of desired dimensionality, say D. This D-dimensional vector represents the embedded patch. Common dimensions for D are 768, 1024, etc., depending on model design and complexity.

**Positional Encodings:** Since the transformer architecture does not inherently account for the order or position of input tokens (in this case, patches), positional encodings are added to each patch embedding. This ensures that the model retains information about the relative or absolute position of the patches within the image. Positional encodings can be either learned or fixed and are element-wise added to the embedded patch vectors.

**Input to Transformer:** The resultant embeddings, each now a D-dimensional vector representing a corresponding 16x16 patch, along with their positional encodings, are then fed into subsequent layers of the Vision Transformer.

**2.MULTI-HEAD ATTENTION:**

Multi-head attention extends the basic self-attention mechanism by allowing the model to jointly attend to different parts of the input representation. It operates by splitting the input embeddings into multiple heads and computing attention independently for each head. These attention heads capture diverse aspects of the input data, providing a richer and more comprehensive understanding. For DR detection, multi-head attention helps the model understand relationships between various elements such as blood vessels, optic disc, and lesions. It allows the model to attend to important regions while considering the context of the entire image, aiding in the accurate detection of DR-related abnormalities.

**Key Components of Multi-Head Attention:**

**Query, Key, and Value Projections:** In multi-head attention, the input embeddings are linearly projected into separate query, key, and value vectors for each attention head. This projection allows the model to learn different representations for attending to different aspects of the input.

**Scaled Dot-Product Attention:** Each attention head computes attention scores by taking the dot product between the query and key vectors, scaled by a factor to control the magnitude of gradients. These scores determine the importance of different elements (patches in the case of ViTs) and are used to weight the corresponding value vectors.

**Concatenation and Linear Projection:** After computing attention independently in each head, the outputs are concatenated and linearly transformed to restore the original dimensionality. This process allows the model to capture diverse patterns and relationships from multiple perspectives.  
In medical image analysis, multi-head attention plays a crucial role in understanding complex structures and features within images:

**Capturing Global Context:** Multi-head attention enables ViTs to capture global context and long-range dependencies across different patches in medical images. This is essential for analyzing subtle abnormalities, identifying relevant features related to DR, and making accurate diagnostic decisions.

**Localized Feature Extraction:** Different attention heads may focus on specific regions or patterns within the image, such as lesions, blood vessels, or optic disc abnormalities. This localized attention facilitates detailed feature extraction and aids in the interpretation of medical imaging data.

**Improved Robustness and Generalization:** By attending to multiple aspects of the input data concurrently, multi-head attention enhances model robustness, reduces overfitting, and improves generalization capabilities. This is particularly beneficial in medical imaging tasks where data variability and complexity are common challenges.

**3.FEED FORWARD LAYER:**

Within a Transformer encoder block, feedforward layers are responsible for processing and transforming the patch embeddings. Typically, a feedforward layer consists of two linear transformations separated by an activation function (GELU). These layers introduce non-linearity and help the model learn complex mappings between input patches and their representations.  
**Gaussian Error Linear Unit Activation Function:**

GELU is an activation function that introduces non-linearities into neural network architectures. It is defined as GELU(x)=x⋅Φ(x), where Φ(𝑥) is the cumulative distribution function of the standard normal distribution.GELU has gained popularity due to its smoothness, non-saturation, and resemblance to the ReLU (Rectified Linear Unit) activation function. It mitigates the vanishing gradient problem to some extent and encourages faster convergence during training.GELU is often used as the activation function in the feedforward layers within Transformer-based models, including ViTs. Its non-linearity aids in capturing complex patterns from patch embeddings and contributes to the overall expressiveness of the model.

**4.RESIDUAL CONNECTIONS AND LAYER NORMALIZATION:**

Residual connections and layer normalization are vital for stable training and effective gradient flow in ViTs. In the context of DR detection, these components ensure that the model can learn from the intricate details of retinal images without encountering vanishing gradients or training instabilities. The residual connections facilitate the flow of information through the layers, while layer normalization maintains consistent activations and speeds up convergence during training.The Layer Normalization layer performs layer normalization on the input tensor. This linear layer maps the normalized embeddings to the output space with dimensions corresponding to the number of classes.

**5.CLASSIFICATION HEAD:**

The classification head in a Vision Transformer (ViT) is the final layer or set of layers responsible for generating predictions or classifications based on the features learned by the ViT model.The classification head in a Vision Transformer (ViT) is the final layer that transforms learned features into predictions. It typically includes dense layers after global average pooling for dimensionality reduction.

During training, the head's parameters are learned alongside the ViT model, ensuring optimal predictions. For tasks like diabetic retinopathy detection, the output of the classification head, using functions like sigmoid or softmax, provides probabilities or class predictions, making it a crucial component for accurate medical image classification.The Reduce layer performs a reduction operation on the input tensor.

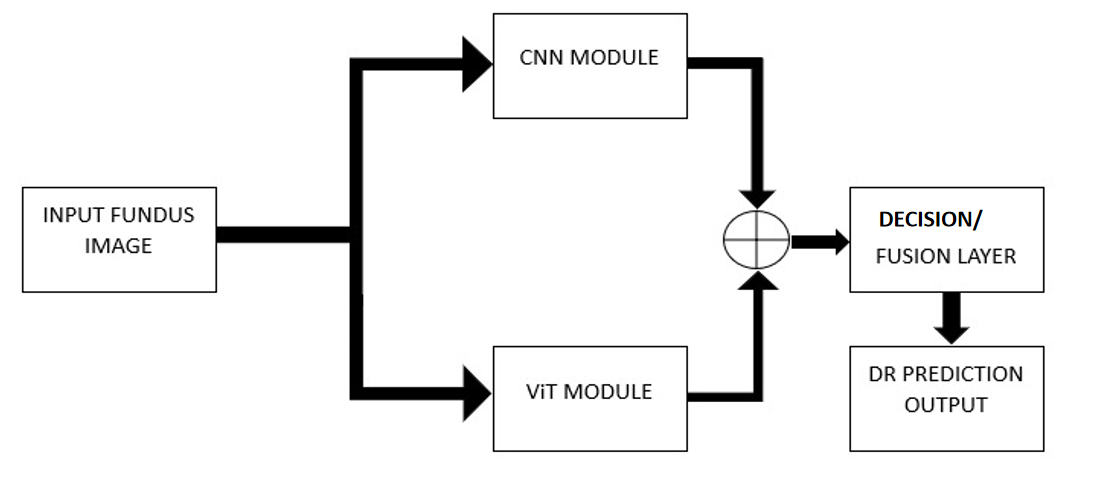
**3.10 HYBRID MODEL:**

The architecture of our ensemble learning approach involves combining the capabilities of a Convolutional Neural Network (CNN) model, specifically ResNet, and a Vision Transformer model. This combination aims to enhance image classification accuracy by leveraging the strengths of both architectures.

We start by loading pre-trained models from HDF5 files: model\_1 represents the ResNet CNN model, while model\_2 represents the Vision Transformer model. These models are then converted into functional models, specifying their input and output layers for further processing.

In the ensemble model setup, we define an input layer (model\_input) that matches the input shape expected by both model\_1 and model\_2. This input image is simultaneously passed through both models using a list comprehension, generating their respective classification outputs.

The ensemble layer, implemented as an averaging mechanism, computes the average of the outputs from model\_1 and model\_2. This ensemble output represents a fused classification prediction derived from the combined insights of the CNN and Vision Transformer models.



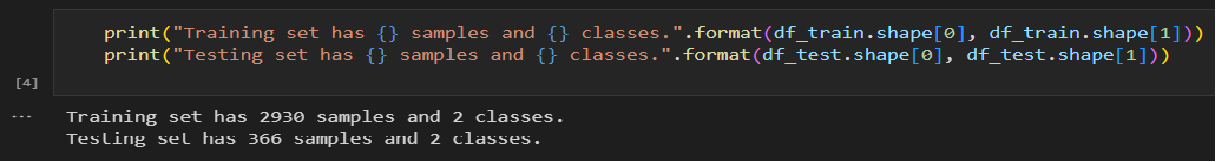
**Figure 3.4 Architecture of hybrid model**

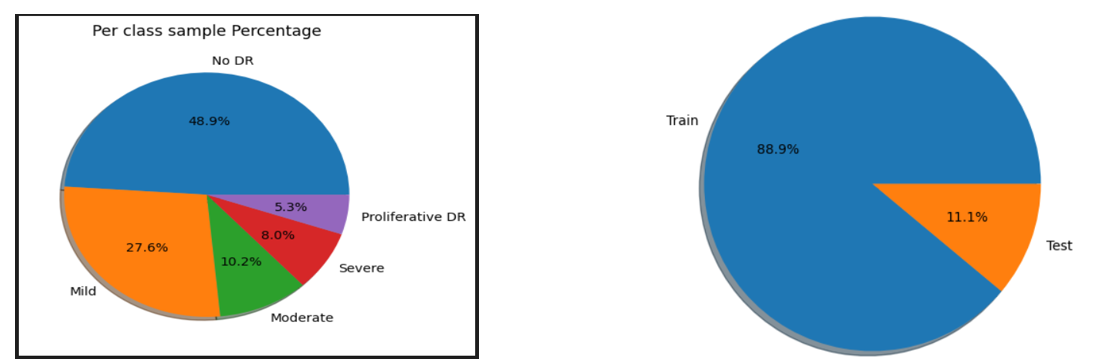
**CHAPTER 4**

**RESULTS AND DISCUSSIONS**

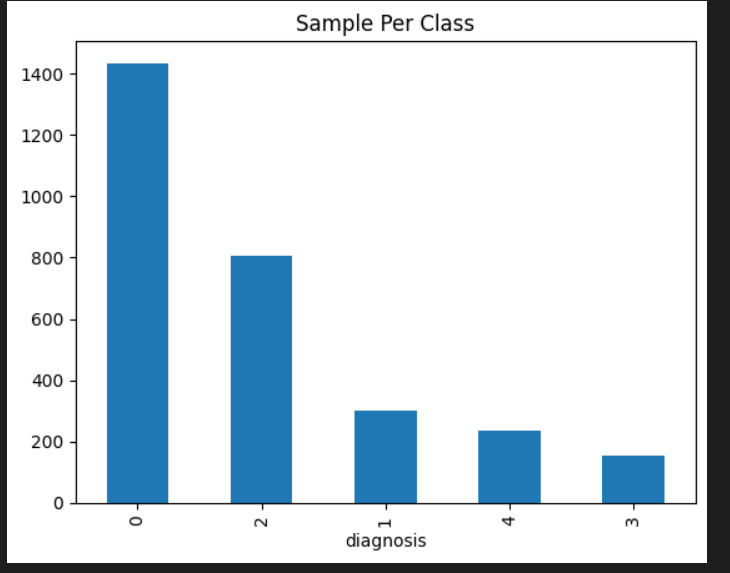
**4.1 EVALUATION OF DATASET**

The dataset used in this project is APTOS2019. This dataset consists of a total of 2930 fundus image samples for training and a total of 366 fundus image samples for testing, each belonging to five classes.





**Figure 4.1 Pie Graph of Dataset**

****

**Figure 4.2 Bar Graph of 5 classes in Train set**

The dataset consists of around 48.9% of images belonging to the No DR category, 27.6% of images belonging to the Mild DR category, 10.2% of images belonging to the Moderate DR category, 8.0% belonging to the Severe DR category and 5.3% of images belonging to the Proliferative DR category.

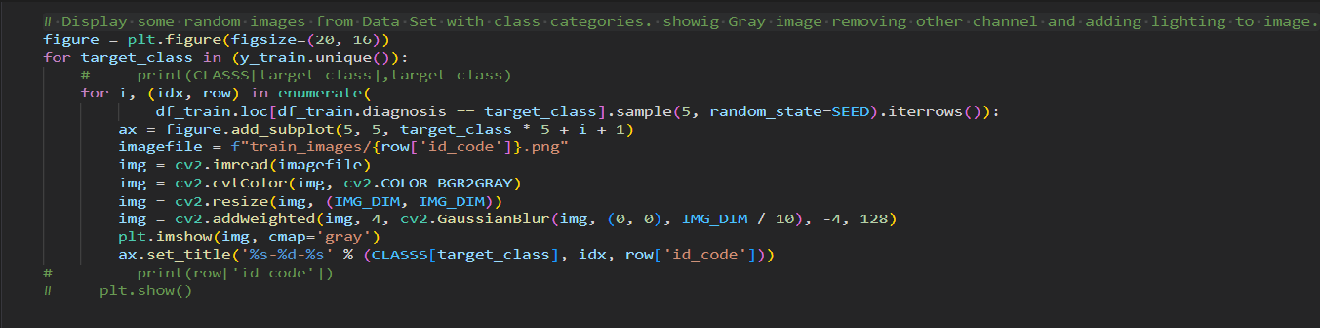
**4.2 DATA PRE-PROCESSING  
 Input Image(Sample of CLASS 4):**

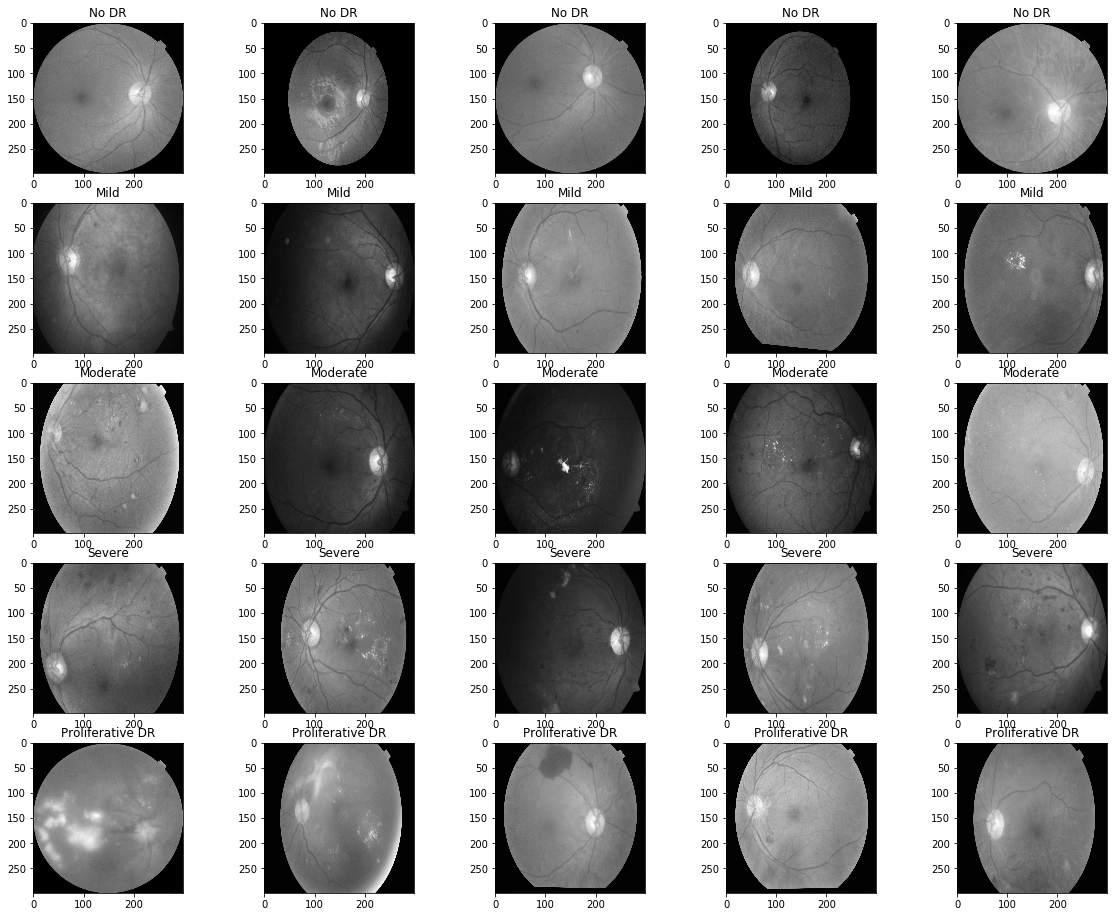
****

**Figure 4.3 Input Imag**e

This figure shows the sample of input fundus images belonging to Class 4. Class 4 belongs to the category of Proliferative DR.

**Cropped To (224x224) & BGR TO Grayscale Image :**

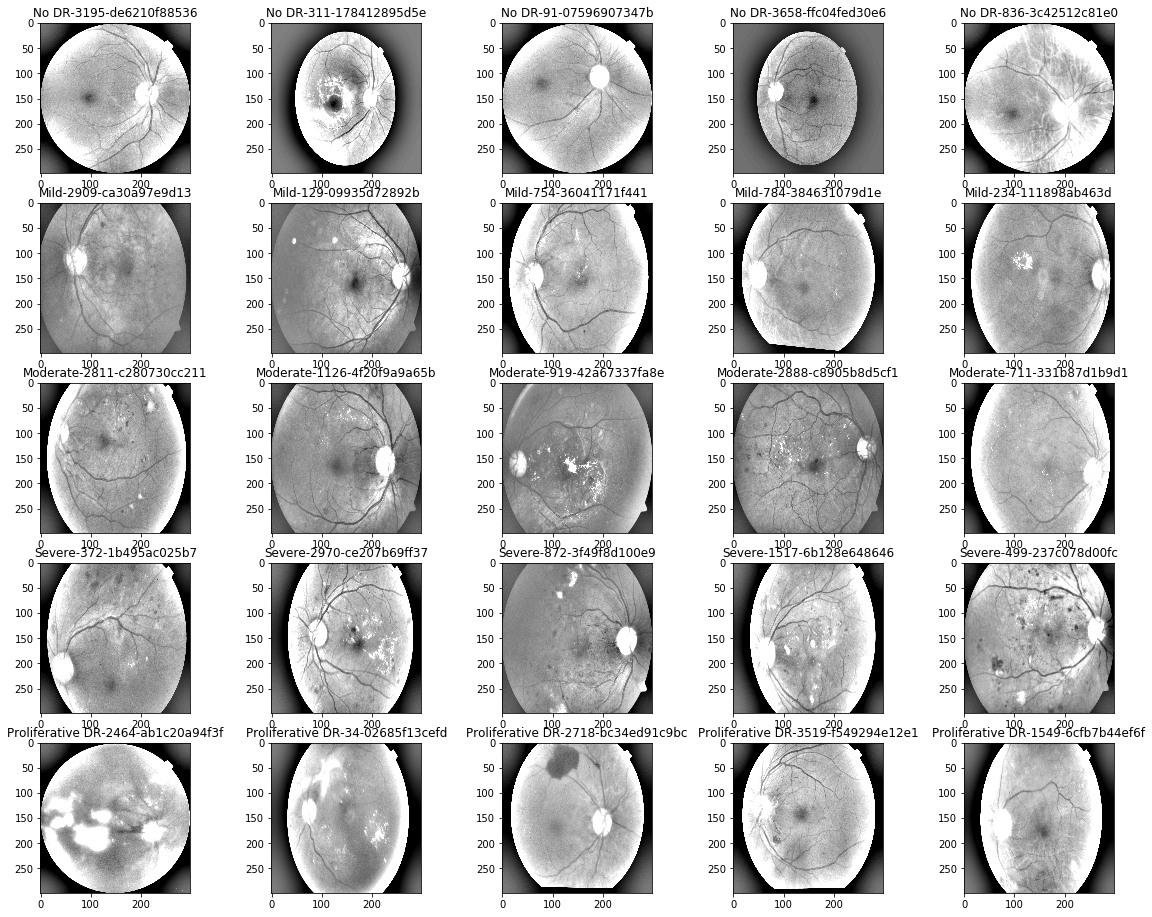
****

****

**Figure 4.4 Cropped and Grayscale converted Input Image**

This figure shows the input fundus image after the beginning steps of Pre-processing. The fundus images are converted from their BGR representation to Grayscale representation, in order to reduce the training load on the CNN model. The Grayscale converted images are again cropped to the size of 244x244, so that the input image is compatible with the input tensor size of the pre-trained CNN model.

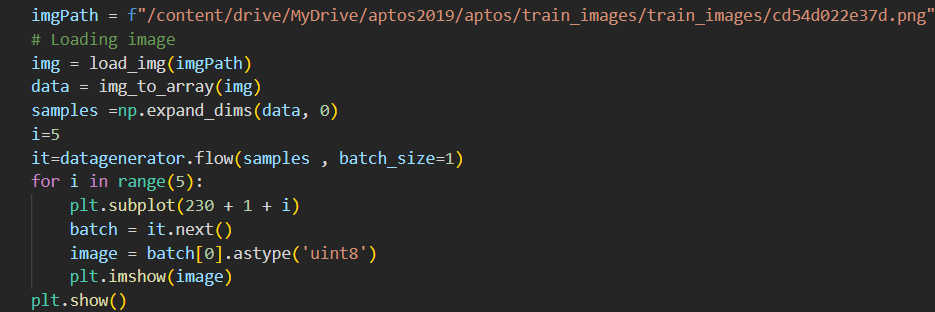
**Lightning Effect for Dark Spots Enhancement:**

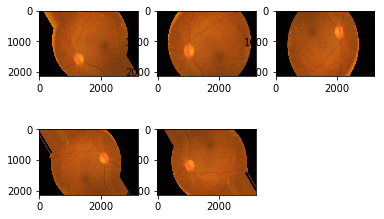
****

**Figure 4.5 Dark Spots Enhanced Input Image**

The Cropped and Grayscale converted fundus images are processed further by adding a lightning effect which enhances the region of interest in the fundus images. Dark Spots, lesions, blood leakage are much more pronounced due to the addition of lightning effect. This can greatly help the model to accurately identify the regions of interest.

**Data Augmentation:**

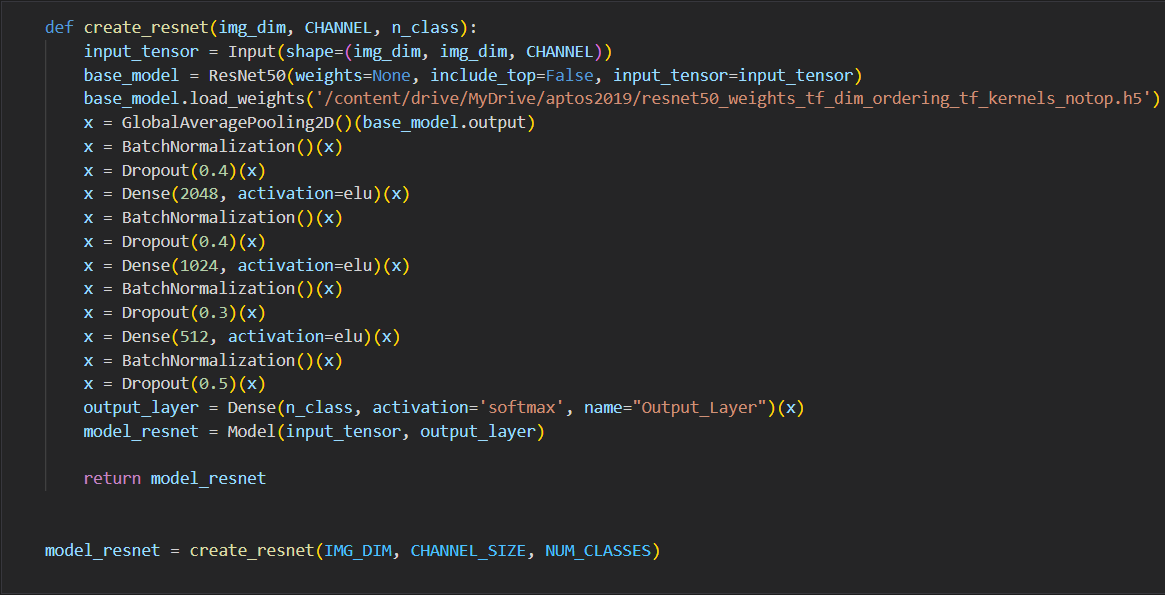
****

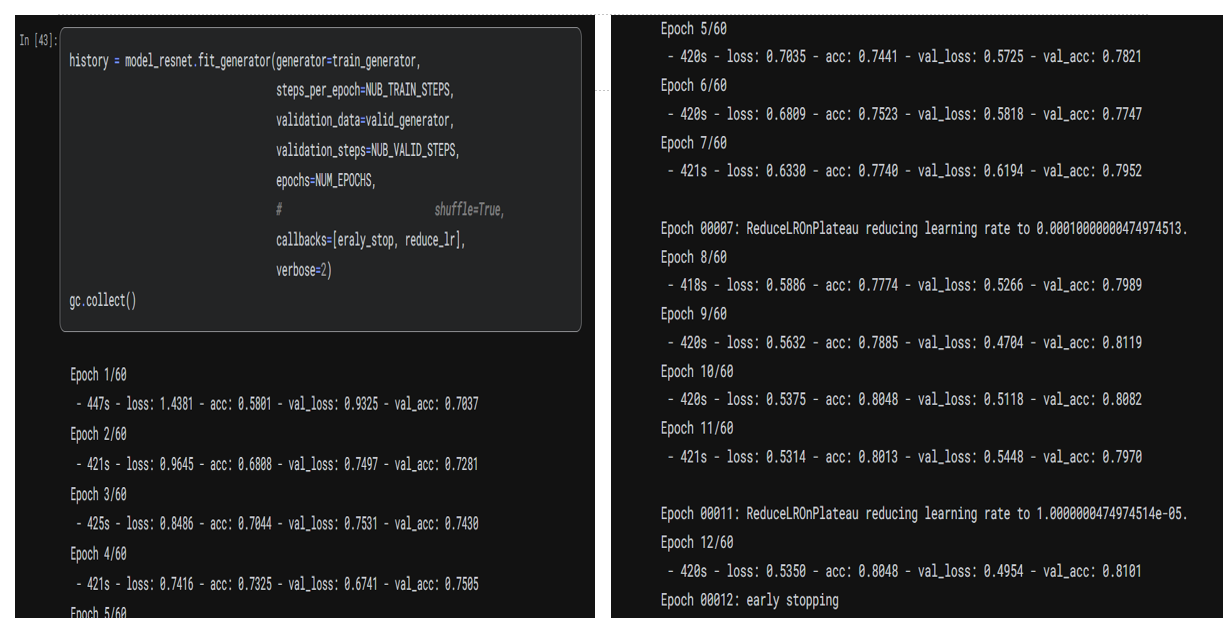
****

**Figure 4.6 Augmentation of Input Image**

This is the last step of pre-processing before sending the fundus images as input to the CNN model. The input images are augmented during this step. Image augmentation is a technique that is used to artificially expand the data-set. The images are twisted, rotated, expanded, sheared. Data augmentation can help to reduce overfitting by providing the model with more data to learn from. This can help the model to generalize better to new data.

**4.3 TRAINING OF RESNET50:**

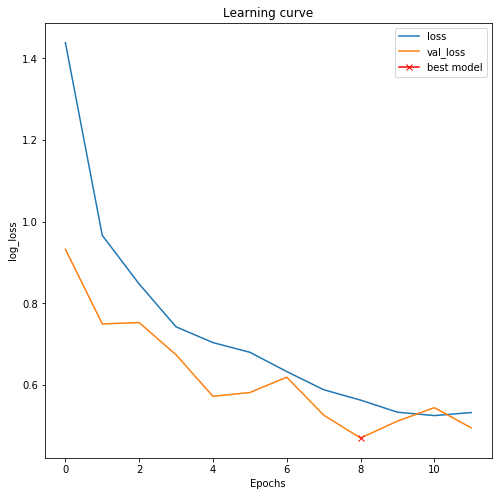
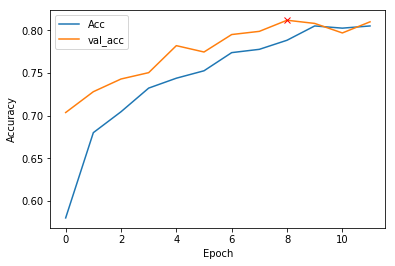
****

****

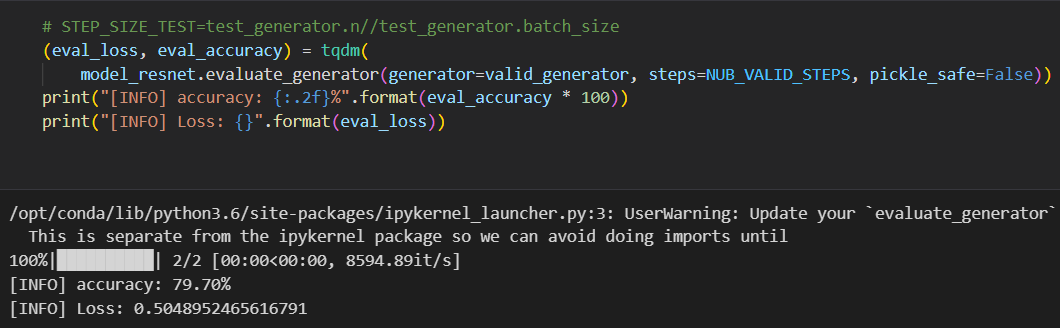
**Figure 4.7 Training of ResNe50 Model**

This step involves the creation and training of the model on the input dataset. The pretrained ResNet50 model is downloaded and all the layers of the model except the last layer is included. Followed by this pretrained model, we add a series of custom layers involving Batch Normalization layers, Dropout layers and Fully Connected layers. The final layer consists of 5 Dense neurons with softmax activation function to aid in classification. Then the model is compiled using the input preprocessed images and is run for a total of 60 epochs. Due to early callback, the model is trained for a total of 12 epochs.

**Evaluation metrics:**

****

**Figure 4.8 Accuracy vs Epochs & Loss vs Epochs Graph.**

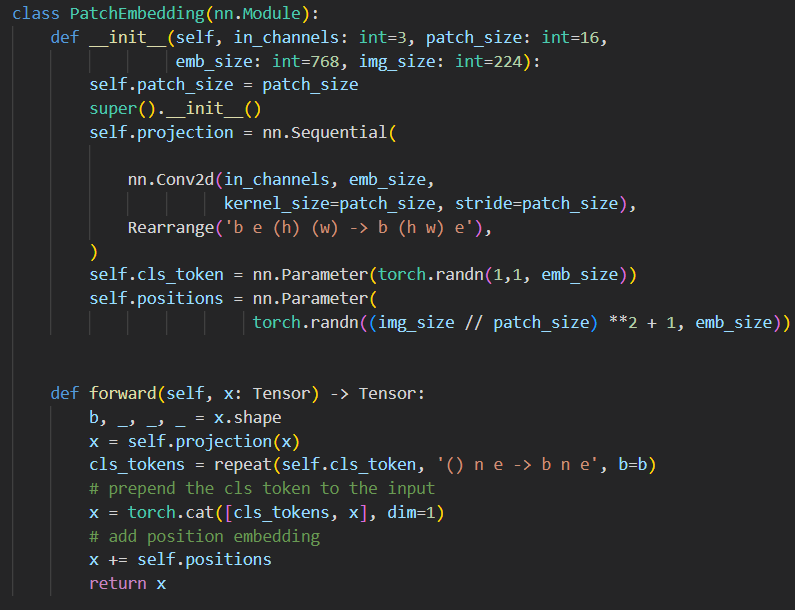


The CNN model is then used to predict in order to extract the evaluation metrics. The accuracy and Loss are recorded against the number of epochs.

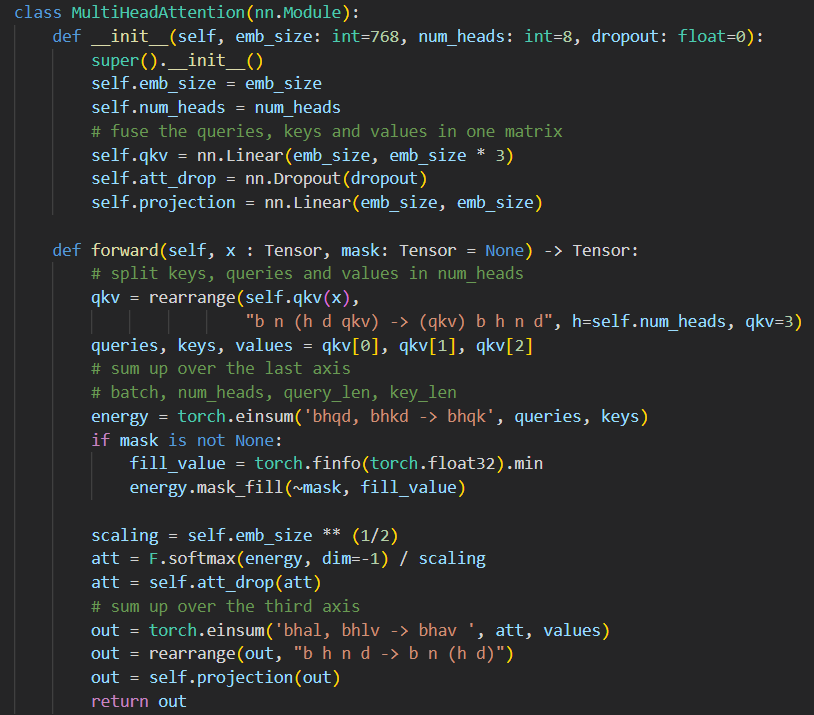
**4.4 TRAINING OF ViT:**

The Vision Transformer is customly built from scratch, including different blocks of its architecture.

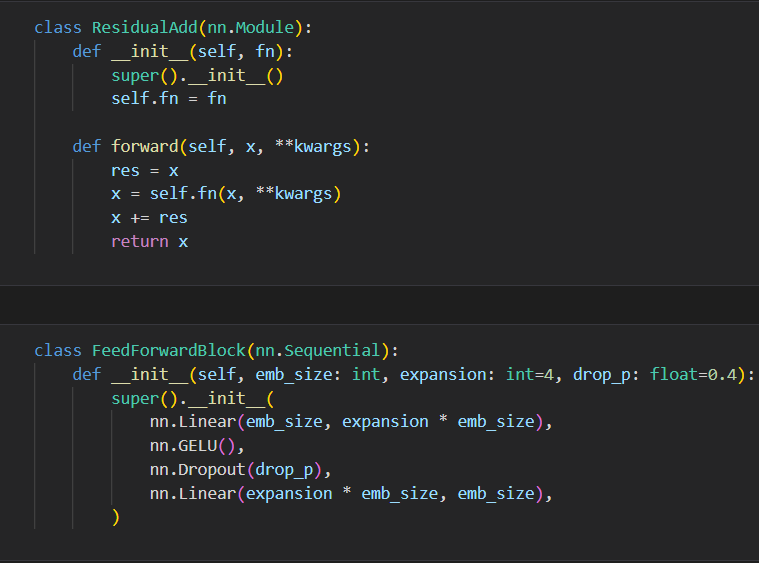
**Patch Embedding:**

****

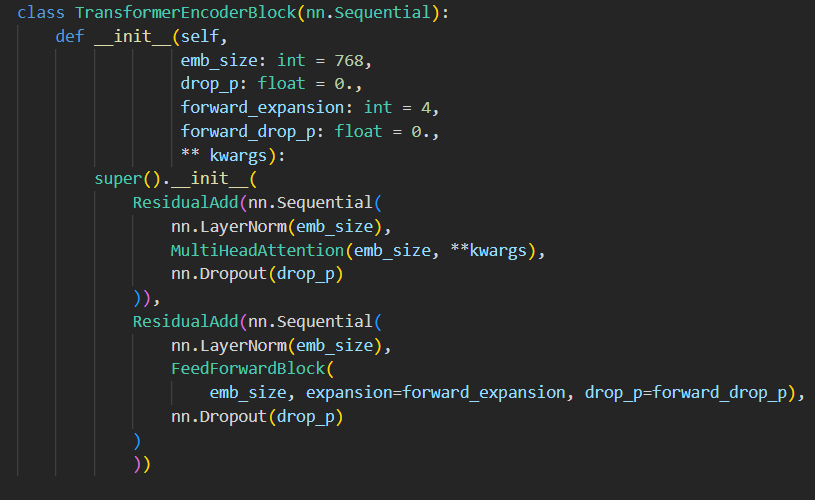
**Multi-Attention Head Code Block:**

****

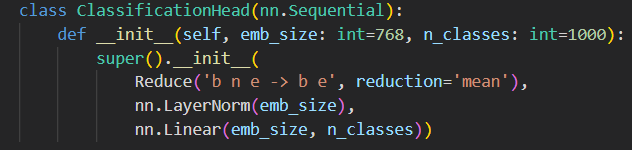
**Residual Add & Feed Forward Code Block:**

****

**Transformer Encoder Code Block:**

****

**Classification Head Code Block:**

****

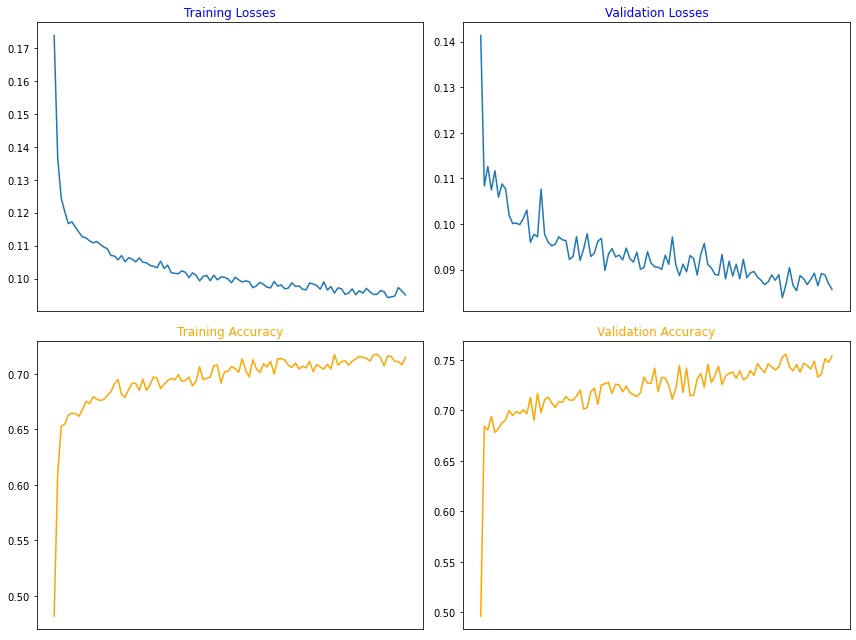
**Training the Model:**

****

**Figure 4.9 Training of ViT Model.**

The ViT model is built by including the above blocks. The patch size is 16 and the input image size is 224x224. The ViT model is trained for a total of 100 epochs.

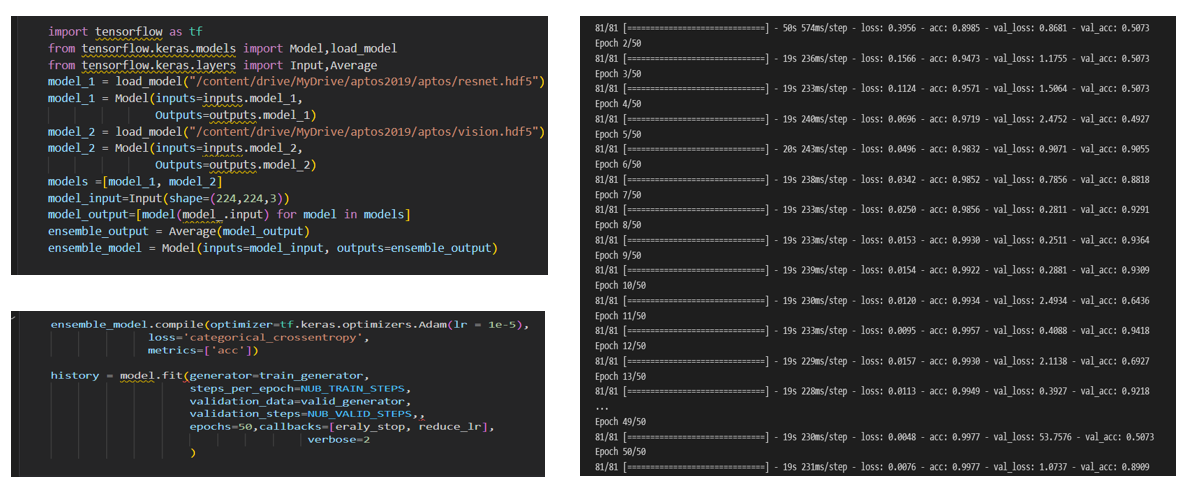
**Evaluation Metrics:**

****

**Figure 4.10 Accuracy vs Epochs & Loss vs Epochs Graph.**

The training history of the model is recorded and tabulated against the number of epochs.

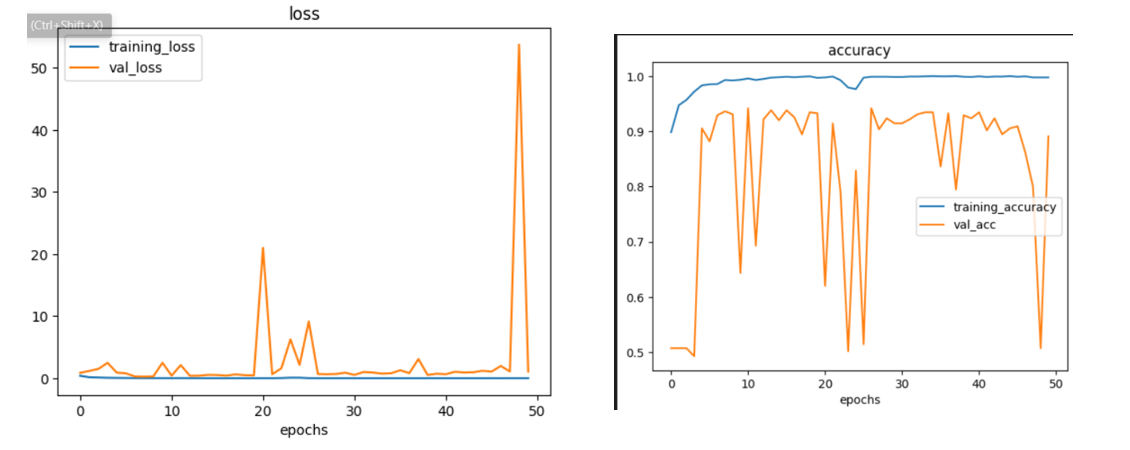
**4.5 TRAINING OF ENSEMBLE MODEL:**

****

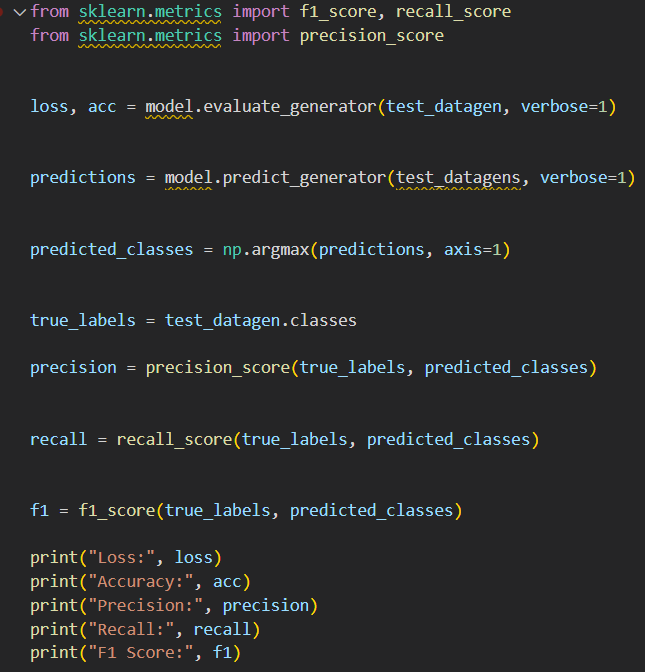
**Figure 4.11 Training of Hybrid Model**

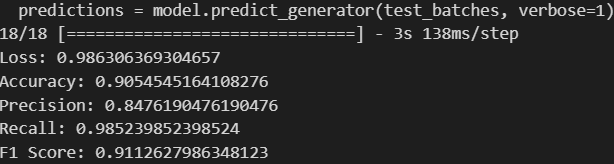
In this step, both the CNN model and the ViT model are combined using the ensemble approach. After combining both the models, the ensemble model is again compiled for a total of 50 epochs.

**Evaluation Metrics:**

****

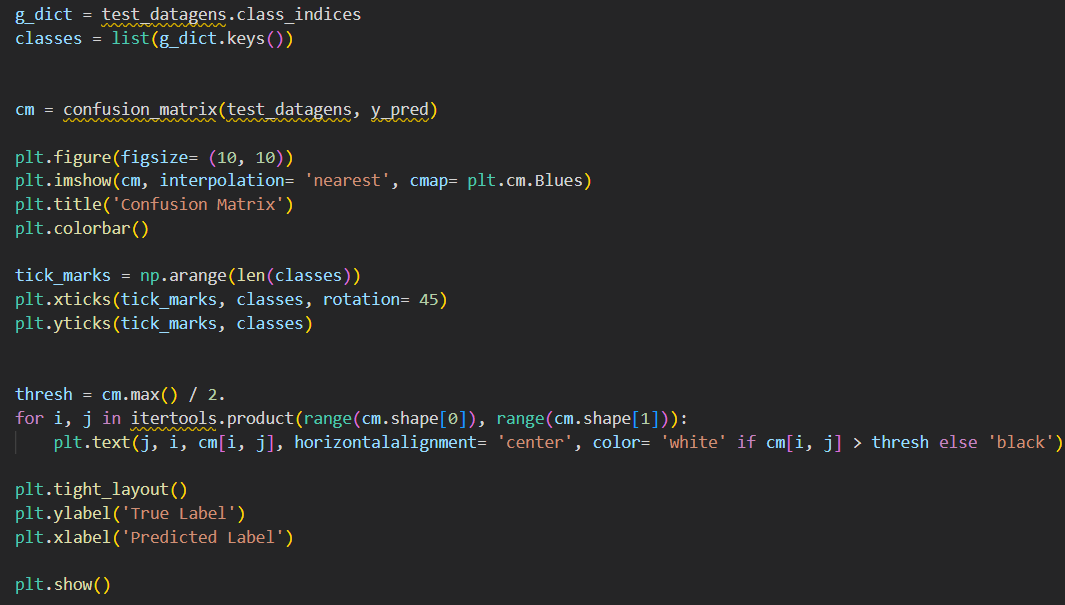
**Figure 4.12 Accuracy vs Epochs & Loss vs Epochs Graph.**

****

****

The different evaluation metrics (Accuracy, Precision, Recall, F1 Score) are evaluated and tabulated for the ensemble model.

**Confusion Matrix:**

****

****

**Figure 4.13 Confusion Matrix**

The Confusion Matrix is created for the testing data to more clearly understand the performance of the ensemble model.

**4.6 PREDICTION:**

****

**Figure 4.14 Prediction of DR.**

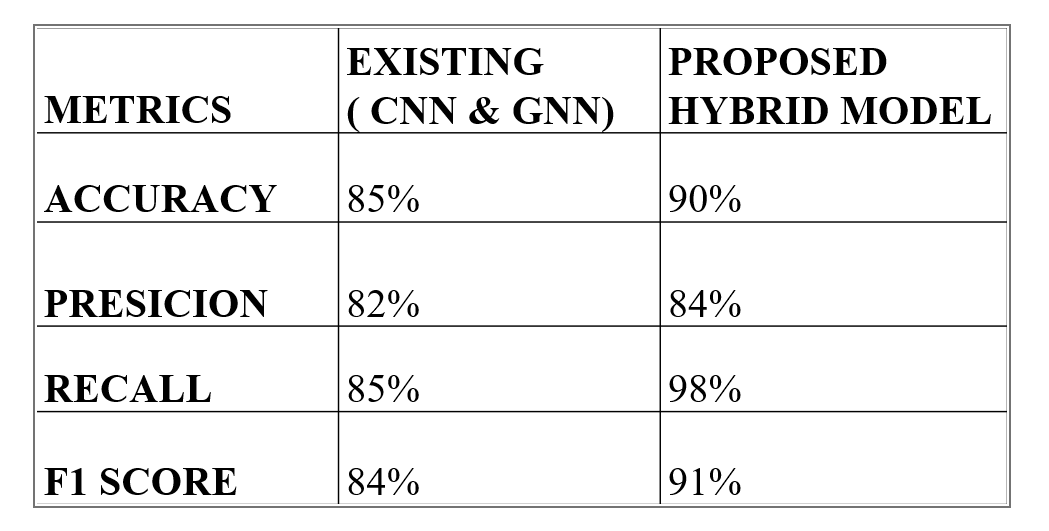
` And finally, the trained ensemble model is used to predict a new input image. The model shows good prediction accuracy.

**CHAPTER 5**

**RESULT, CONCLUSION AND FUTURE SCOPE**

**5.1 RESULT:**

* In this project, a fusion of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) alongside ensemble learning surpassed existing models in diabetic retinopathy classification. Leveraging the strengths of CNNs and ViTs, the ensemble model achieved superior accuracy.
* Notably, our approach focused on utilizing Vision Transformers for global feature extraction, a methodology not previously explored.
* This breakthrough marks a significant stride in medical image analysis, demonstrating the potential of advanced deep learning architectures and ensemble techniques for enhancing diagnostic accuracy in diabetic retinopathy and related ocular conditions.
* The individual models, ResNet and the Vision model, exhibited strong performance on their own. However, the ensemble model surpassed their individual accuracies, highlighting the effectiveness of ensemble learning in enhancing predictive capabilities.
* This result underscores the importance of leveraging diverse model architectures and ensemble techniques to achieve superior performance in complex tasks image classification.
* The Evaluation metrics are compared for the Existing and the Proposed Model.



**5.2 CONCLUSION**

In this project, Hybrid Model for Diabetic Retinopathy Classification is a groundbreaking advancement in combating vision impairment caused by DR. By combining CNN and Transformer models using ensemble learning, created a system that significantly enhances diagnostic accuracy and efficiency. This fusion leverages CNN's spatial dependency capturing and Transformer's attention mechanisms for accurate classification. Extensive validation shows superior performance metrics compared to traditional methods, ensuring reliability and effectiveness across diverse datasets. This model promises improved patient outcomes and sets the stage for more accurate diagnostic tools in ophthalmology worldwide.

**5.3 FUTURE SCOPE**

The future scope of this project encompasses several avenues for further development and impact:

* **Extension to other Ophthalmic diseases:**

The model can be further extended to classification of other ophthalmic diseases such as cataract and glaucoma from the fundus images.

* **Increasing the Robustness:**

The Robustness of the model can be further increased by training the ensemble model on other available datasets of fundus images.

* **Leveraging Multi-modal Data:**

Ensembles can be designed to incorporate not only retinal images but also additional data sources like patient demographics, lab test results, or optical coherence tomography (OCT) scans. This multimodal approach can potentially improve the accuracy and comprehensiveness of DR classification.

**REFERENCES**

[1] M. Feng, J. Wang, K. Wen and J. Sun, "*Grading of diabetic retinopathy images based on graph neural network*" in IEEE Access, vol. 11, pp. 98391-98401,September 2023.

[2] G. Ali, A. Dastgir, M. W. Iqbal, M. Anwar and M. Faheem, "*A hybrid convolutional neural network model for automatic diabetic retinopathy classification from fundus images,*" in IEEE Journal of Translational Engineering in Health and Medicine, vol. 11, pp. 341-350, June 2023

[3] S. Sundar and S. Sumathy, "*Classification of diabetic retinopathy disease levels by extracting topological features using graph neural networks*," in IEEE Access, vol. 11, pp. 51435-51444, May 2023.

[4] F. Yan, B. Yan and M. Pei, "*Dual transformer encoder model for medical image classification,*" IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 690-694, September 2023.

[5] C. -F. R. Chen, Q. Fan and R. Panda, "*Crossvit: cross-attention multi-scale vision transformer for image classification*", IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 347-356, February 2022.