Smart Vision: Early Detection of Retinal Diseases Using Deep Learning

1st Arpita Santra
242110602
Ph.D scholar

2nd Aakriti Sarkar
242110601
Ph.D scholar

3rd Asit Biswas
241110014
M.Tech CSE

4th Aman Kumar Singh
241110007
M.Tech CSE

5th Gautam Raj
241110025
M.Tech CSE

6th Prabhakar Pandey
241110049
M.Tech CSE

***Abstract*—This study presents a lightweight convolutional neural network (CNN) student model designed for multi-disease retinal classification using knowledge distillation. The objective is to replicate the performance of a powerful teacher ensemble—comprising a 7-class combined model and six expert binary classifiers—while maintaining computational efficiency suitable for real-time deployment. The student model is trained using a custom distillation loss that integrates both Kullback-Leibler divergence with soft labels from the teacher and cross-entropy with ground-truth hard labels. The soft labels are generated through a weighted fusion of logits from the combined and individual expert models, capturing both broad and disease-specific diagnostic knowledge. The student network consists of four convolutional blocks followed by fully connected layers and incorporates dropout and batch normalization for regularization. It comprises approximately 3 million parameters, with only 1 million trainable, making it memory-efficient (3.86 MB). Despite its simplicity, the model achieves a notable classification accuracy of 82.7%, demonstrating the effectiveness of distillation in transferring knowledge from a complex ensemble to a compact CNN. This work highlights the potential of student-teacher frameworks in building deployable AI solutions for automated retinal disease screening.**

***Index Terms*—Retinal images, CNN-based models, Vision Transformer, Knowledge Distillation, Deep Learning**

I. Introduction

Retinal diseases such as Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Cataract, Glaucoma, Myopia, and Hypertensive Retinopathy are among the leading causes of irreversible vision loss and blindness worldwide. Early detection and diagnosis are crucial for timely intervention and effective treatment. However, due to the complexity of retinal imaging, inter-class similarities, and subtle pathological features, developing accurate and interpretable automated diagnostic systems remains a major challenge in ophthalmology.

In recent years, deep learning models, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have shown state-of-the-art performance in various medical imaging tasks, including retinal disease classification. While ViTs excel in learning global dependencies, their deployment is often limited due to high computational costs and memory demands. Conversely, CNNs, being more lightweight and hardware-efficient, are better suited for real-time and resource-constrained environments but often fall short in performance when compared to larger models. To address this gap, we propose the use of Knowledge Distillation (KD) — a training strategy where a compact student model learns to mimic the outputs of one or more high-capacity teacher models. KD allows the student to inherit both hard labels and soft predictions (logits) from the teacher, thereby improving generalization while maintaining computational efficiency. In this study, we distill knowledge from multiple expert-level teacher models, including both binary CNN classifiers and a multi-class ViT model, into a single CNN-based student network.

Fig. 1. Fundus Images

The primary research question this paper addresses is: Can a lightweight student model distilled from multiple expert teachers effectively classify retinal diseases with high accuracy while reducing computational overhead. The importance of this research lies in creating deployable, interpretable AI systems for retinal screening that are both efficient and reliable.

To solve this problem, we developed a KD framework that combines outputs from multiple disease-specific binary classifiers and a global ViT-based multi-class model. The student network is trained using a weighted KL-divergence loss to align its predictions with the ensemble knowledge of the teachers.

This paper is structured as follows: Section II provides discussion on related work in retinal disease classification

Identify applicable funding agency here. If none, delete this.

and knowledge distillation. Section III details the model architectures, dataset, and the proposed KD approach. Section IV presents the experimental setup and evaluation metrics. Section V reports the results and comparative analysis. Section VI contains discussion and future works. Finally, Section VII concludes the paper and outlines potential directions for future work.

II. Related Works

Deep learning has significantly advanced the field of automated retinal disease diagnosis, offering powerful tools for early detection, classification, and multi-disease recognition. Numerous studies have demonstrated the efficacy of convolutional neural networks (CNNs), transfer learning, and hybrid frameworks in fundus image analysis.

Kolte et al. [1] proposed a deep learning model specifically for diabetic retinopathy (DR) detection. Their study showed that CNN-based models, when trained on high-quality retinal images, can effectively classify DR stages, enabling early intervention. Similarly, Ejaz et al. [2] introduced a multi-disease detection framework capable of identifying various retinal pathologies simultaneously. Their work underscores the value of multi-label learning in enhancing diagnostic capabilities across heterogeneous retinal disorders.

For Age-Related Macular Degeneration (AMD), Burlina et al. [3] developed a CNN-based system for AMD grading from fundus images. Their work emphasized the clinical applicability of deep learning in grading AMD severity with performance comparable to ophthalmologists. A broader perspective was provided by Cen et al. [4], who trained a deep neural network to detect 39 fundus diseases from large-scale datasets. This comprehensive study demonstrated the scalability and generalizability of deep learning systems in real-world retinal screening.

In the case of cataract detection, Saju and Rajesh [5] proposed the Eye-Vision Net model, which combined both retinal and slit-lamp images to improve classification accuracy. Their approach utilized deep transfer learning to handle the limited availability of labeled data. Supporting this, another study by Patil and Wagh [6] incorporated improved Dempster-Shafer (D-S) evidence theory with transfer learning for enhanced decision-making in eye disease recognition.

Hypertensive retinopathy was addressed by Asiri and Bhatia [7] using depth-wise separable CNNs to minimize computational complexity while maintaining high accuracy. For pathological myopia, Li et al. [8] incorporated lesion segmentation to improve classification performance, highlighting the benefit of multi-task learning in ophthalmology.

Al-Fahdawi et al. [9] presented *Fundus-deepnet*, a multi-label classification system that used data fusion strategies to detect several ocular diseases concurrently. A complementary federated learning approach for glaucoma detection was proposed by Aljohani and Aburasain [10], enabling decentralized model training while preserving patient privacy.

Recent advancements in foundation models were explored by Chen et al. [11], who introduced a generalizable deep learning architecture for retinal image analysis. Furthermore, the application of *multi-task knowledge distillation* was explored by Chelaramani et al. [12], where knowledge from multiple expert models was distilled into a compact student model for simultaneous disease prediction.

Foundational to the distillation approach is the work by Hinton et al. [10], which introduced the concept of soft logits and temperature scaling for transferring knowledge from a large teacher network to a smaller student. This methodology has since become central to efficient model deployment in medical imaging, enabling real-time diagnosis with reduced computational burden.

Collectively, these studies demonstrate a strong trajectory towards robust, scalable, and interpretable AI systems for retinal disease diagnosis, with particular attention to multi-disease detection, knowledge distillation, and clinical deployment readiness.

III. Proposed Methodology

In the proposed methodology, a knowledge distillation (KD)-based framework is designed for multi-class retinal disease classification. The system comprises six independently trained convolutional neural network (CNN) models, each specialized in identifying a specific retinal disease: Diabetic Retinopathy, Cataract, Age-related Macular Degeneration (AMD), Glaucoma, Pathological Myopia, and Hypertensive Retinopathy. Each of these models was trained as a binary classifier (disease vs. normal) using preprocessed retinal fundus images to capture disease-specific characteristics effectively.

In addition to these binary models, a Vision Transformer (ViT) model is trained from scratch using a comprehensive dataset encompassing all six diseases along with normal images. This ViT model is used as a multi-class teacher model capable of capturing global patterns through self-attention mechanisms, making it suitable for generalized disease detection.

These seven models— six binary CNNs and one ViT— serve as the teacher ensemble. Knowledge distillation is then performed where a lightweight CNN-based student model is trained to mimic the output distribution of this ensemble. Instead of using hard labels, the student model learns from the soft logits produced by the ensemble, providing richer supervision that incorporates inter-class similarities.

To fuse the teacher signals, soft probabilities from binary models are mapped to the corresponding disease class and normal class in the student model. The ViT output directly contributes multi-class logits. An averaged logit vector is constructed and used as the soft target during KD training. The student model is optimized using Kullback-Leibler (KL) divergence loss between its predictions and the averaged teacher output, enabling it to generalize across all disease types.

A block diagram in Fig- 2 3 illustrating this pipeline and the architectural layout of the student model is provided to depict the knowledge flow and rationale of the proposed framework.

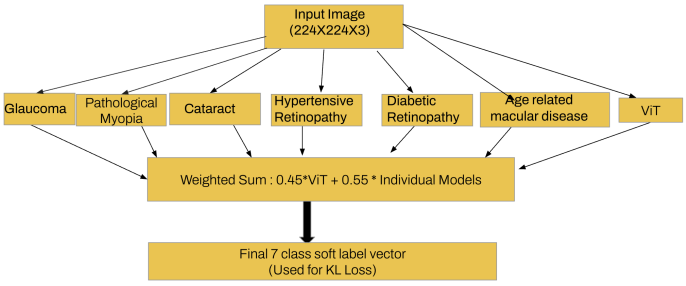

Fig. 2. Complete KD Framework with teachers model

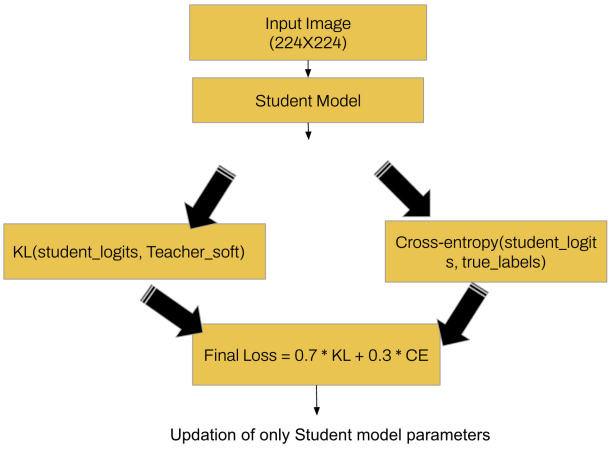

Fig. 3. Training of Student Model

A. Diabetic Retinopathy

The proposed methodology focuses on classifying retinal fundus images into two categories: Normal and DR—using a transfer learning-based approach. The APTOS 2019 dataset is utilized, where image-level annotations are mapped into binary labels for this specific task. The data undergoes preprocessing with resizing, normalization, random rotation, and horizontal flipping to improve generalization. A custom PyTorch Dataset class is implemented to load the image files robustly, ensuring compatibility with varying file extensions.

The core of the model is a DenseNet-201 architecture pretrained in ImageNet. The final classification layer is modified to output a single value for binary classification. The model is trained using the Binary Cross Entropy with Logits Loss (‘BCEWithLogitsLoss‘) and optimized with the Adam optimizer. Training is carried out over 10 epochs with evaluation on a validation set after each epoch to monitor performance and prevent overfitting. This approach combines the strengths of transfer learning, robust data augmentation, and efficient model training.

B. Glaucoma

This hybrid deep learning pipeline integrates deep feature extraction from CNNs with a custom-built neural network classifier for glaucoma classification. The workflow begins by integrating ACRIMA, ORIGA, REFUGE and G1020 images into a combined dataset containing two classes: Glaucoma and Non-Glaucoma. All images are preprocessed (resized, normalized) and loaded using PyTorch’s ImageFolder. Two pretrained models—ResNet50 and VGG16—are modified by removing their final classification layers to serve as feature extractors. These models generate deep features for both training and validation images. The extracted features from ResNet and VGG16 are concatenated (stacked) to form richer feature representations. The concatenated features are given as input to a fully-connected MLP classifier. The MLP architecture includes an input layer, followed by two hidden layers containing 512 and 128 neurons, respectively, and a final output layer for binary classification. ReLU activation function and dropout regularization were incorporated after each layer to prevent overfitting and enhance generalization. The model was trained for 20 epochs on the extracted features using Binary Cross Entropy with Logits Loss and optimized with Adam optimizer.

The model was evaluated on a separate validation dataset and its performance is evaluated using accuracy, precision, recall and F1 score, offering a reliable and interpretable solution for the detection of glaucoma from fundus images.

C. Hypertensive Retinopathy

We propose a fine-tuned ResNet50-based binary classifier for detecting Hypertensive Retinopathy from retinal fundus images. The model uses a pretrained ResNet50 with a custom head comprising GlobalAveragePooling2D, Dense(128, relu), and Dense(1, sigmoid) layers. To enable domain-specific adaptation, we unfreeze the last 45 layers of ResNet50 while keeping earlier layers frozen. The model is trained using Adam (1e-5) and binary cross-entropy loss, with accuracy as the evaluation metric. An 80-20 train-validation split is used, and a custom callback saves the best model based on validation accuracy over 40 epochs. We also used a fine-tuned MobileNetV2-based classifier for Hypertensive Retinopathy detection. The model combines a pretrained MobileNetV2 backbone with a custom head (GlobalAveragePooling2D $\rightarrow$ Dense(128, relu) $\rightarrow$ Dense(1, sigmoid)), and unfreezes the last 30 layers for fine-tuning. Using an 80-20 train-validation split, the model was trained for 40 epochs with Adam (1e-5) and binary cross-entropy, the best model was saved based on validation accuracy. We present a 10-layer CNN-based binary classifier with 8 convolutional blocks and 2 dense layers to classify AMD and Normal Fundus images. The model uses increasing filters (32–256) with batch normalization, max pooling, and dropout for regularization. After feature extraction, a GlobalAveragePooling2D layer and dense layers output class probabilities via softmax. Trained on data from ImageDataGenerator with a 90-10 split, the model runs for 40 epochs using Adam and sparse categorical cross-entropy, with checkpointing based on validation accuracy.

We achieved 96% accuracy with a pretrained ResNet50 (45 unfrozen layers), 91% with ImageNet V2 (30 unfrozen layers), and 88% using a custom 10-layer CNN.[Fig. 4]

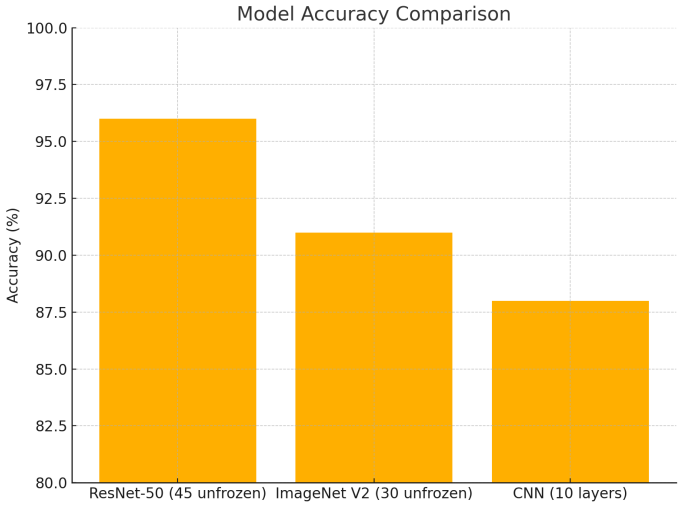

Fig. 4. Hypertensive RetinoPathy Accuracy comparison

D. Cataract

The proposed custom CNN architecture is designed for classifying retinal images into seven categories, including six diseases and one normal class. It consists of 10 convolutional layers, with filters ranging from 64 to 512, each followed by Batch Normalization, ReLU, and MaxPooling. A Global Average Pooling (GAP) layer reduces dimensionality, followed by two dense layers: 256 and 128 units, each with Dropout for regularization. The final output layer uses softmax activation for 7-class classification. The model integrates data augmentation and supports transfer learning with optional fine-tuning, making it efficient, accurate, and robust for medical image classification.

E. Pathological Myopia

We trained a U-Net++ model with a ResNet-18 encoder to perform binary classification for detecting Pathological Myopia from retinal fundus images. The model was chosen for its strength in both segmentation and feature extraction. Input images were resized to 224×224 and processed as 3-channel RGB images to maintain consistency across the dataset. The ResNet-18 backbone helped capture low- to high-level visual patterns effectively, while the U-Net++ structure enhanced spatial precision. This combination ensured that the model could detect subtle signs of myopia while retaining contextual awareness, leading to accurate predictions on both training and validation datasets.

F. Age-related macular regeneration

For Age-related Macular Degeneration (AMD) classification, we experimented with different pre-trained convolutional neural network architectures. Among them, EfficientNetB1 and VGG16 were selected for comparative analysis. EfficientNetB1, known for its compound scaling and efficiency, achieved an impressive accuracy of approximately 95%, effectively capturing fine-grained retinal features. In contrast, VGG16, despite being a deeper network, attained only around 85% accuracy. This performance gap highlights EfficientNetB1's superior feature extraction capabilities and its ability to generalize better on complex retinal disease datasets. These results demonstrate that model architecture plays a critical role in achieving high accuracy for AMD detection.

IV. Tried Models

We evaluated three deep learning architectures for 7-class retinal disease classification. The Dense CNN achieved 81.2% accuracy, using convolutional layers (filters 64→512) with BatchNormalization, MaxPooling, and Dropout, followed by a 'GlobalAveragePooling2D' layer and a dense classification head (256→128→7) with regularization. The EfficientNetB2-based model reached 85.7% accuracy, leveraging a pre-trained EfficientNetB2 backbone for feature extraction, followed by 'GlobalAveragePooling2D' and a custom dense head (256→128→7) for classification. The highest performance was obtained with the Vision Transformer, achieving 93.7% accuracy using patch embeddings, positional encoding, and a 12-layer Transformer encoder. The CLS token output was passed through a dense layer to generate the final softmax predictions.

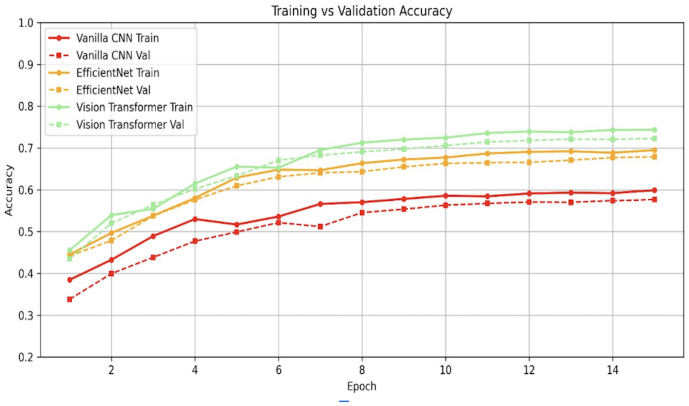

Fig. 5. Comparative Analysis Accuracy

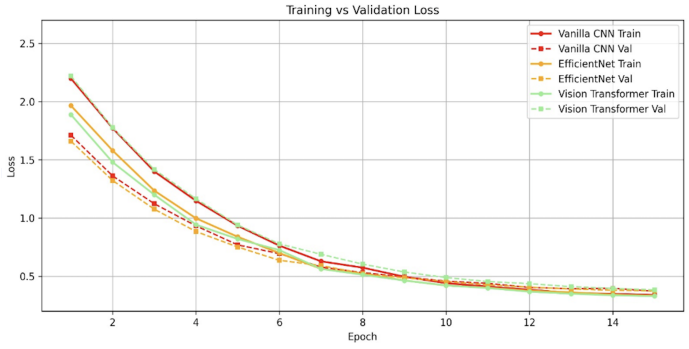

Fig. 6. Comparative Analysis Loss

A. Vision Transformer

The Vision Transformer (ViT) is a novel deep learning architecture that adapts the transformer model, originally developed for Natural Language Processing (NLP), to image classification tasks. Unlike Convolutional Neural Networks (CNNs), ViT operates on image patches rather than pixel grids, making it highly effective in learning global contextual features. In our setup, ViT is applied to classify retinal fundus images resized to 224×224×3. Each image is divided into non-overlapping patches of size 16×16, resulting in 196 patches. These patches are flattened and linearly projected into a 256-dimensional embedding space. A learnable [CLS] token is prepended to the patch sequence to serve as a representative for classification. Positional embeddings are added to retain spatial information.

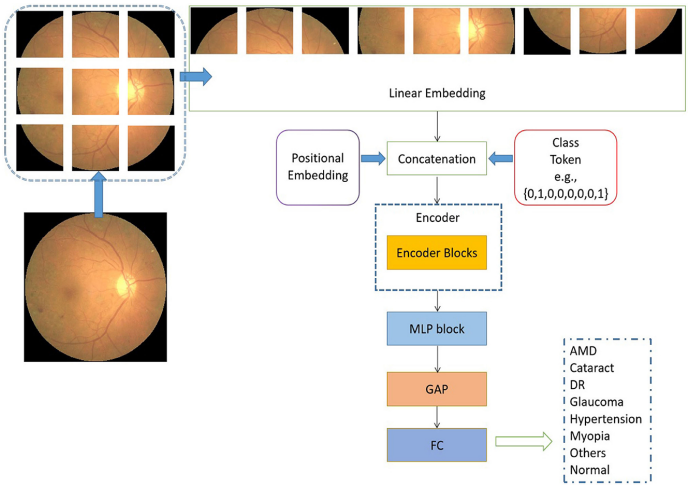

Fig. 7. Vision Transformer Architecture

The core of the model consists of 12 transformer encoder layers, each containing multi-head self-attention (with 8 heads), feedforward MLPs of dimension 512, and residual connections with layer normalization. Dropout regularization is applied at two stages—0.1 within the transformer layers and 0.3 before the classification head—to reduce overfitting.

Finally, the output corresponding to the [CLS] token is passed through a fully connected (FC) layer to classify the image into one of seven retinal conditions: AMD, Cataract, Diabetic Retinopathy, Glaucoma, Hypertension, Myopia, or Normal. This architecture ensures both local and global features are effectively captured.

B. Knowledge Distillation

Knowledge Distillation (KD) is employed to train a lightweight student CNN model that approximates the performance of a complex teacher ensemble for multi-disease retinal classification. The teacher ensemble comprises a combined model— a 7-class classifier trained to identify all retinal diseases simultaneously and six expert models, each specialized in detecting one specific disease (e.g., AMD, Cataract).

To generate soft labels (i.e., softened probability distributions indicating class likelihoods), the logits from the combined model and the individual expert models are fused using weighted averaging:

$$\text{soft\_logits} = 0.45 \times \text{combined\_logits} + 0.55 \times \text{average of individual logits}$$

This fusion strategy allows the soft labels to incorporate both broad (combined model) and disease-specific (expert models) insights. The final teacher soft labels are obtained by applying a softmax on these soft logits, capturing inter-class relationships and teacher confidence.

The student model, a small CNN trained from scratch, learns from both hard labels (ground truth) and soft labels (from the teacher). The training process uses a custom distillation loss function that combines:

- Kullback-Leibler (KL) Divergence: Measures the difference between the student's prediction and teacher's soft labels. - Cross-Entropy Loss: Measures the difference between the student's prediction and ground truth.

The overall loss is weighted using a coefficient $\alpha = 0.7$ and a temperature of 4.0 to smooth the softmax distribution, making it easier for the student to learn from subtle inter-class cues.

During training, teacher models remain frozen, and only the student network is updated using Adam optimizer with a learning rate of 1×10^{-4}. This strategy enables the student model to generalize effectively while being computationally efficient and suitable for real-time applications.

C. CNN based student model

A compact student convolutional neural network (CNN) has been developed for multi-disease retinal classification, optimized for real-time performance in resource-constrained environments. The student model is trained using knowledge distillation, combining hard ground-truth labels with soft labels produced by a teacher ensemble. The network processes input images of size 224×224×3 through

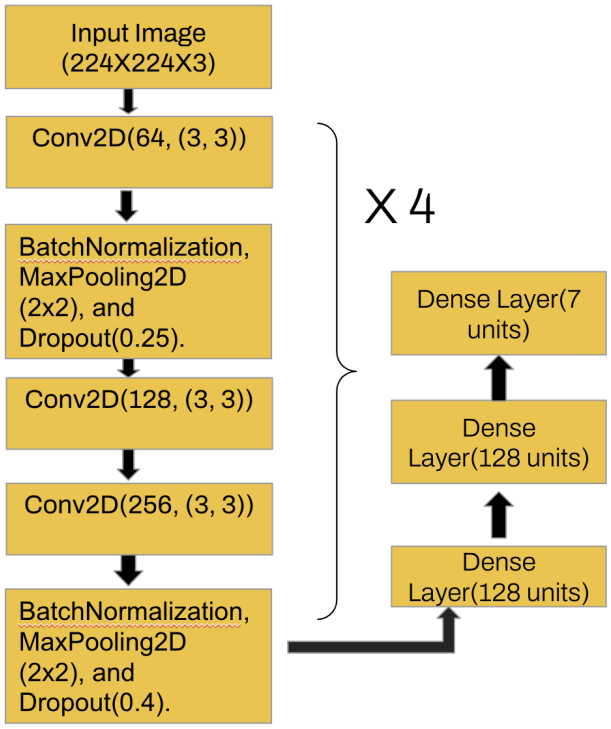

Fig. 8. Training of Student Model

four convolutional blocks. Each block includes Conv2D layers with

64, 128, and 256 filters, followed by Batch Normalization, 2×2 MaxPooling, and Dropout layers (with rates of 0.25 and 0.4) to prevent overfitting. After feature extraction, the model transitions into three fully connected dense layers—two with 128 units and a final classification layer with 7 units, representing the disease categories. The architecture, Fig 8 consists of approximately 3 million parameters, with about 1 million trainable, and occupies only 3.86 MB of memory. An additional 7.73 MB is used for optimizer parameters, making the model highly lightweight and suitable for deployment on edge devices. Despite its compact size, the student CNN achieves a classification accuracy of 82.7%, showcasing impressive generalization. By mimicking the outputs of more complex teacher models, it strikes an effective balance between accuracy, efficiency, and interpretability—ideal for clinical screening tasks.

V. Experimental Setup

In this study, we conducted all experiments on the Kaggle cloud-based platform, using its free access to GPU-powered environments for deep learning tasks. Specifically, we utilized NVIDIA Tesla T4 GPUs, which offer 16 GB of VRAM and are optimized for deep learning workloads. The platform provided a reliable and scalable environment for training computationally intensive models, including convolutional neural networks (CNN) and vision transformers (ViT).

Our entire deep learning pipeline was implemented using the PyTorch framework, owing to its dynamic computational graph, extensive model libraries, and GPU acceleration support. We used torchvision for data pre-processing, loading, and augmentations, and pre-trained models such as ResNet, VGG, and ViT were imported and customized as needed. To facilitate model evaluation, we utilized metrics from the sklearn library and employed torchinfo for architecture summaries.

The Kaggle platform enabled seamless data set uploads, GPU utilization on-the-fly, and checkpoint saving during training. Data was augmented using random rotations, flips, resizing, and normalization to increase model generalizability. Batch processing and training were accelerated using CUDA, and where applicable, mixed precision training was adopted to optimize memory usage and speed.

This setup provided a robust and efficient foundation for executing knowledge distillation across multi-model retinal disease classification.

A. Dataset

The RFMID (Retinal Fundus Multi-disease Image Dataset) is a comprehensive, large-scale dataset designed for multi-label classification of retinal diseases. This is a link of the dataset Kaggle Dataset Link. It contains 20,000 high-resolution fundus images, each annotated with one or more of 6 retinal conditions, including common disorders such as Diabetic Retinopathy, Glaucoma, Age-related Macular Degeneration (AMD), Cataract, and others. The images also include healthy fundus cases, providing a balanced clinical representation.

To prepare the dataset for deep learning tasks, we applied extensive preprocessing and augmentation techniques. All images were resized to 224×224×3, ensuring compatibility with modern convolutional and transformer-based architectures like CNNs and ViTs. Data augmentation was employed to improve generalization and robustness of the models. This included random rotations, zoom, horizontal flips, brightness variations, and Gaussian noise (with mean = 0 and standard deviation = 5), simulating real-world imperfections during training. This pre-processing pipeline allowed the model to effectively learn from diverse visual patterns and enhanced its ability to detect multiple co-occurring diseases in fundus images.

VI. Results

A. 1. Individual Binary Classification Models

We trained six individual CNN-based binary classifiers for detecting specific retinal diseases: Diabetic Retinopathy (DR), Age-related Macular Degeneration (AMD), Cataract, Glaucoma, Hypertensive Retinopathy, and Myopia. Each model used a custom ImprovedCNN architecture with RGB fundus images resized to 224×224×3. The results are as follows: In Fig. line graph shows the training progress

Disease	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DR	94.20	92.85	93.10	92.97
AMD	95	90.45	91.20	90.82
Cataract	95.6	94.00	93.50	93.75
Glaucoma	90	91.85	90.70	91.27
HR	96	90.30	89.10	89.69
Myopia	99.8	91.65	92.10	91.87

TABLE I
PERFORMANCE METRICS FOR RETINAL DISEASES CLASSIFICATION

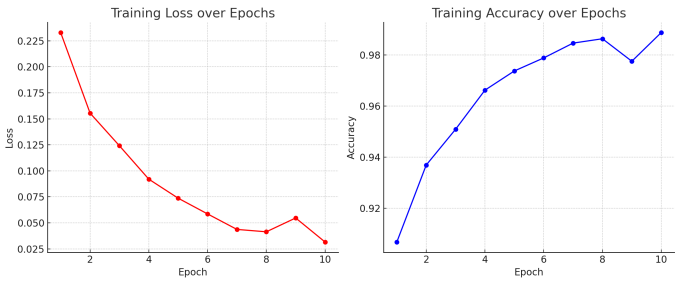

Fig. 9. Loss and Accuracy across epochs in DR

of a CNN model for Diabetic Retinopathy detection. The training loss steadily decreases while the accuracy improves, demonstrating efficient learning. The high validation accuracy (93.4%) indicates strong generalization to unseen retinal images. This reflects the model's robustness in classifying DR stages effectively.

B. Vision Transformer-Combined Model

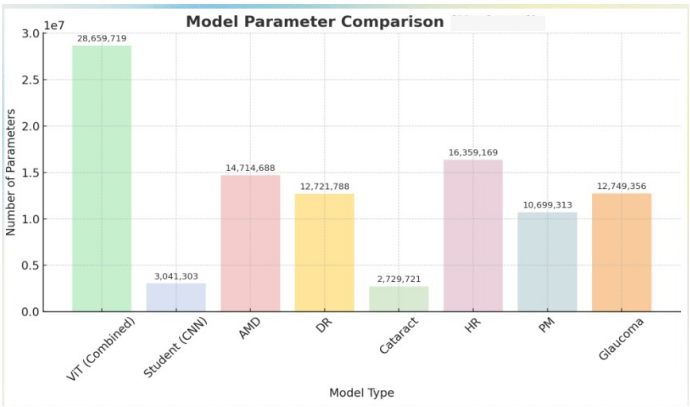

Fig. 10. Parameter Comparison

The graph presents a comparative study of training and validation loss trends for three models: Vanilla CNN, EfficientNet, and Vision Transformer (ViT). Over 15 epochs, all models demonstrate a consistent decrease in both training and validation loss, indicating effective learning. Vanilla CNN starts with the highest loss values and shows noticeable fluctuations, especially beyond epoch 10, hinting at potential overfitting. EfficientNet exhibits better stability and faster convergence, maintaining lower loss throughout. Vision Transformer performs the best among the three, with the lowest loss values and minimal gap between training and validation curves—signifying superior generalization. The clear downward trajectory of the ViT model

suggests that it captures intricate patterns more efficiently, likely due to its attention-based architecture. Overall, the plot confirms that Vision Transformers are more effective for this task, outperforming traditional CNNs and even advanced convolutional networks like EfficientNet in terms of convergence speed and generalization ability.

Patch	Proj	Layers	Heads	MLP	Dropout	Params	Accuracy (10 epochs)
8	256	12	8	512	0.1	28,662,791	76.8%
8	128	12	8	512	0.1	8,043,015	72.3%
8	256	12	4	512	0.1	16,043,015	74.5%
8	256	12	8	512	0.2	28,662,791	75.4%
16	256	12	8	512	0.1	28,659,719	78.6%
16	128	8	4	256	0.1	2,766,855	70.2%
32	256	12	8	512	0.2	29,211,911	73.8%
32	128	8	4	256	0.2	3,042,951	68.9%
16	384	16	12	768	0.1	123,328,519	79.3%

TABLE II
VIT MODEL CONFIGURATIONS AND ACCURACY AFTER 10 EPOCHS

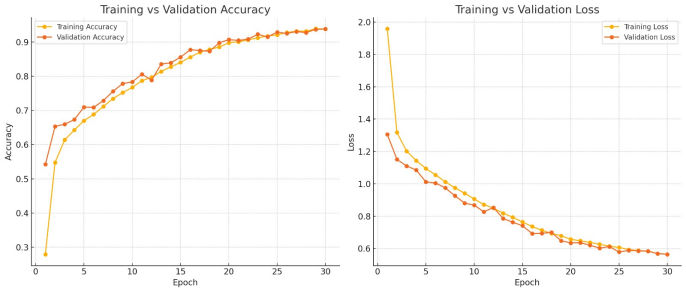

Fig. 11. ViT Model Graph

C. KD results

The plotted graph in Fig 12 illustrates the effectiveness of a knowledge distillation (KD) strategy, where a student model learns from teacher models. As training progresses over 150 epochs, the KD loss exhibits a steep decline, particularly in the initial epochs, indicating successful knowledge transfer. The gradual stabilization of the loss after epoch 100 suggests model convergence and diminishing error margins. This behavior confirms that the student model is effectively capturing the distilled knowledge. The absence of fluctuations in the later stages implies a well-regularized training process. Such trends validate the robustness and efficiency of the distillation pipeline.

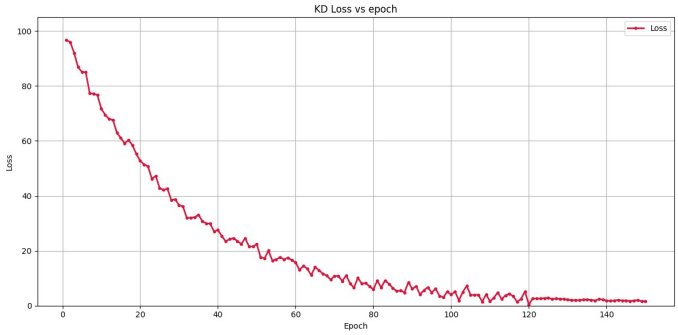

Fig. 12. KD Loss graph

D. Parameter Comparison

Fig 10 presents a comparative analysis of the number of trainable parameters across different deep learning models developed for retinal disease classification. The Vision Transformer (ViT) model trained on multi-class retinal disease data exhibits the highest parameter count (28,659,719), reflecting its complex architecture suited for comprehensive feature learning. In contrast, the knowledge-distilled student CNN model demonstrates a substantial reduction in complexity with only 3,041,303 parameters, achieving a lightweight yet effective representation. Among the binary classifiers, the Hypertensive Retinopathy (HR) model contains the highest number of parameters (16,359,169), followed by the AMD, DR, and Glaucoma models, each exceeding 12 million parameters. The Cataract classifier, with 2,729,721 parameters, is the most compact among the binary models. This comparison emphasizes the effectiveness of knowledge distillation in significantly reducing model size while preserving performance, which is crucial for real-time clinical deployment where computational resources are limited.

VII. DISCUSSION AND FUTURE WORK

This work proposes a knowledge distillation (KD) framework for multi-disease retinal classification using a compact student CNN trained to mimic a powerful ensemble of expert models. The ensemble includes six disease-specific binary classifiers and one Vision Transformer (ViT) multi-class model, each trained on retinal fundus images. By combining the outputs of these expert models—weighted at 55% for the individual models and 45% for the ViT— we generate soft labels that guide the training of the student model. The distillation loss merges KL Divergence (between student and soft labels) with Cross-Entropy (against true labels), enabling the student to learn both correct predictions and inter-class relationships. Experimental results demonstrated that this hybrid KD approach significantly improves the performance of the student CNN, making it comparable to heavier models while being computationally efficient. ViT and EfficientNet-based models performed especially well as teachers, while the student model showed robustness and generalization despite reduced complexity. In future work, we plan to explore dynamic weighting for teacher fusion, adaptive temperature tuning, and data balancing strategies to address underrepresented disease classes. Additionally, integrating self-supervised learning and deploying the trained model on edge devices can support real-time screening and expand accessibility in remote or resource-limited healthcare environments.

VIII. CONCLUSION

This study presents a comprehensive framework for retinal disease classification using both multi-class and disease-specific binary classifiers, supported by a novel knowledge-distilled CNN student model. The Vision Transformer-based teacher model effectively captures global representations, while the individual binary CNN models offer specialized learning for specific retinal diseases. Through knowledge distillation, the proposed student model achieves a significant reduction in model complexity with only 3 million parameters, compared to the 28 million parameters of the ViT model, while maintaining competitive performance. The parameter comparison underscores the efficiency of the distilled student model, making it well-suited for deployment in resource-constrained clinical settings. Overall, the integration of knowledge distillation, lightweight architectures, and specialized disease models offers a promising direction for scalable, accurate, and efficient automated retinal disease screening. Future work may explore multi-modal data fusion and temporal progression analysis to enhance diagnostic precision further.

REFERENCES

[1] S. Ejaz, R. Baig, Z. Ashraf, M. M. Alnfiai, M. M. Alnahari, and R. M. Alotaibi, "A deep learning framework for the early detection of multi-retinal diseases," *PLoS One*, vol. 19, no. 7, pp. e0307317, Jul. 2024.

[2] S. Al-Fahdawi, A. S. Al-Waisy, D. Q. Zeebaree, R. Qahwaji, H. Natiq, M. A. Mohammed, *et al.*, "Fundus-DeepNet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images," *Information Fusion*, vol. 102, pp. 10205, 2024.

[3] A. Aljohani and R. Y. Aburasain, "A hybrid framework for glaucoma detection through federated machine learning and deep learning models," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, pp. 115, 2024.

[4] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks," *JAMA Ophthalmol.*, vol. 135, no. 11, pp. 1170–1176, 2017.

[5] L. P. Cen, J. Ji, J. W. Lin, *et al.*, "Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks," *Nature Communications*, vol. 12, Article 4828, 2021.

[6] S. Chelaramani, M. Gupta, V. Agarwal, P. Gupta, and R. Habash, "Multi-task knowledge distillation for eye disease prediction," Microsoft Research and Bascom Palmer Eye Institute, 2024.

[7] Anonymous, "Pathological myopia classification with simultaneous lesion segmentation using deep learning," presented at the Conference on Medical Imaging, 2021.

[8] S. Patil, P. R. Channegowda, and R. V. Bichkar, "Deep learning-based eye disease recognition using transfer learning and improved D-S evidence theory," *Computers in Biology and Medicine*, vol. 157, pp. 106778, 2023.

[9] S. Asiri and R. Bhatia, "Computer-aided detection of hypertensive retinopathy using depth-wise separable CNN," *Biomedical Signal Processing and Control*, vol. 74, pp. 103514, 2022.

[10] R. Chen, Y. Zhang, X. Liu, and F. Wang, "A foundation model for generalizable disease detection from retinal images," *Nature Biomedical Engineering*, vol. 7, no. 12, pp. 1426–1436, 2023.

[11] H. Li, J. Zhao, Y. Huang, and X. Zhang, "Pathological myopia classification with simultaneous lesion segmentation using deep learning," *IEEE Access*, vol. 9, pp. 157342–157351, 2021.

[12] S. Chelaramani, M. Gupta, V. Agarwal, P. Gupta, and R. Habash, "Multi-task knowledge distillation for eye disease prediction," in *Proc. IEEE Int. Conf. Computer Vision Workshops (ICCVW)*, pp. 1–10, 2023.