A Lightweight Hybrid AI Framework for Cataract Detection Using Fundus Images: Real-World Evaluation on Clinical Data

Ishaan Kunwar
STEM Program
Edison High School
Edison, United States
e-mail: ishaankunwar19@gmail.com

***Abstract*—Cataract is one of the most prevalent eye diseases affecting the elderly population. In underserved regions, a low ophthalmologist-to-patient ratio and a scarcity of specialized medical devices pose challenges for early detection. This study aims to harness recent advancements in Deep Learning (DL) to automate cataract detection. Although numerous studies have been conducted in this area, improving model accuracy and minimizing overfitting, all while maintaining a simple architecture that requires fewer computational resources, remains challenging. This research proposes a hybrid method that merges featurization achieved by a Convolutional Neural Network (CNN) with classification techniques to improve prediction accuracy. The model's predictive performance is evaluated not only on the original test dataset but also on a newly acquired image set collected independently from a hospital. Experiments are conducted across different model architectures, such as CNNs and hierarchical Vision Transformers (ViTs) in combination with classifiers, such as multi-layer perceptron (MLP), K-nearest neighbors, and RandomForest. The highest accuracy is achieved using a combination of the ConvNeXtXLarge architecture for feature extraction coupled with a MLP classifier, reaching 92.3% on the original test dataset and improving to 94% on the new hospital-based dataset.**

Keywords- cataract, convolutional neural network, vision transformer, multilayer perceptron.

I. INTRODUCTION

Cataracts are a predominant cause of visual impairment and blindness, accounting for approximately 33% of cases of impaired vision and 51% of first causes of blindness worldwide [1]. This eye ailment occurs due to the clumping of proteins in the lens, which significantly reduces its transparency. The ophthalmologists detect it by performing a manual retinal exam. They administer eye drops to dilate the pupil and then use the slit lamp, which is a specialized microscope with bright light, to clearly examine the retina for opacity.

Early detection of cataracts is crucial to preventing progressive blindness or avoiding costly surgical interventions, particularly in underserved regions where the ophthalmologist-to-patient ratio can be alarmingly low, often around 1 to 10,000.

Currently, cataract detection and diagnosis in hospitals are primarily based on clinical examinations by ophthalmologists using devices, such as slit lamps, ophthalmoscopes, and biomicroscopy. These methods involve direct visualization of the eyes' lens to assess opacity levels, and in some cases, specialized imaging techniques like ultrasound biomicroscopy and Scheimpflug imaging may also be employed. However, these procedures depend heavily on the availability of trained medical professionals and advanced equipment, often resulting in significant delays in diagnosis and treatment initiation, particularly in underserved areas.

Fundus imaging, however, presents a simpler and more accessible alternative, particularly suitable for underserved regions. Fundus cameras are relatively portable, cost-effective, and easy to operate, requiring less specialized training compared to traditional ophthalmic diagnostic methods. These characteristics enable broader deployment, even in remote or resource-constrained settings, facilitating early detection and continuous monitoring of cataracts. Thus, leveraging fundus imaging could substantially improve the scalability and reach of cataract screening programs, particularly benefiting communities with limited healthcare infrastructure.

Recent advances in Artificial Intelligence (AI), particularly Deep Learning (DL), have shown promising potential for automating medical diagnostics across various domains, including ophthalmology. Deep learning-based systems, especially Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have successfully demonstrated high performance in recognizing pathological conditions in ophthalmic imaging [2]. However, despite these successes, existing models often face critical challenges, including overfitting, excessive computational complexity and costs, and limited generalizability when exposed to datasets collected from diverse and independent clinical settings [3][4].

To address these issues, it is crucial to develop streamlined yet highly accurate models that are computationally efficient, generalizable, and robust to varied imaging conditions and demographic differences encountered in different regions. This paper proposes a hybrid approach that integrates deep feature extraction through advanced CNN architectures, specifically ConvXtnet-large, with traditional classification methods, including multi-layer perceptron (MLP), K-nearest neighbors (KNN), and RandomForest classifiers, with MLP being the chosen classification method. Evaluating the model performance not only on standard benchmarking datasets but also on independently acquired hospital datasets provides a more rigorous and realistic assessment of its generalization capabilities.

Such comprehensive validation across diverse datasets is essential for ensuring reliability and clinical applicability in underserved regions where disparities in healthcare accessibility demand robust, efficient, and accurate automated diagnostic tools. The developed approach aims to bridge gaps in oph-

thalmic care by providing a scalable, accessible, and precise cataract detection system.

The remainder of the paper is structured as follows: In Section II, the paper discusses prior work in the area of automating cataract detection using machine learning, including prior successes and limitations. Section III discusses the design of the experiment, including data collection, data preprocessing, feature extraction, architecture of classification models and of the proposed approach, and experimental procedures. In Section IV, the validation and testing results of the experiments are disclosed. In Section V, the results are discussed and explained and the limitations of the study are revealed. In Section VI, future prospects of the proposed method are detailed, and in Section VII, the paper is concluded and final thoughts are summarized.

II. Related work | Methods

Several studies have leveraged Machine Learning (ML) and Deep Learning (DL) approaches to automate cataract detection. Typically, these methodologies involve three primary stages: data preprocessing, feature extraction, and classification. Traditionally, CNNs have dominated this domain due to their strength in image-based feature extraction. However, recently ViTs have gained significant traction, demonstrating promising results in ophthalmic disease diagnosis.

Multiple researchers have employed ViT-based methods with considerable success. Ali et al. [5] introduced a hyperparameter-optimized ViT model combined with Explainable AI techniques to diagnose various eye diseases from a diverse medical image dataset, achieving an accuracy of 91.40%. Similarly, Purba et al. [6] utilized a ViT architecture tailored for human eye disease classification, optimizing hyperparameters to attain an accuracy of 92.86% and recall of 85.72%. Another pertinent work by Gummadi et al. [7] implemented ViT for ocular disease classification, achieving an F1-score of 83.49%. Complementing these findings, Kumar et al. [8] conducted a comparative evaluation between traditional CNNs, specifically Visual Geometry Group-16 (VGG16) and ResNet50, and ViT on a consistent dataset, concluding that ViT demonstrated superior performance with an accuracy score of 70%.

Further advancements have been achieved through hybrid transformer models and specialized feature engineering approaches. Wang et al. [9] proposed a Transformer-based Knowledge Distillation Network (TKDNet) specifically tailored for cortical cataract grading. Their innovative methodology includes a zone decomposition strategy for extracting precise features and introduces specialized sub-scores addressing key clinical indicators, such as opacity location, area, and density. Their multi-modal mix-attention Transformer efficiently fused these sub-scores with image modalities, achieving a notable accuracy of 95.1% and recall of 81.6%.

Despite the growing popularity of ViTs, CNN-based methods remain highly relevant due to their computational efficiency and high accuracy. Khan et al. [10] successfully utilized a pre-trained VGG19 CNN model to detect cataracts from color fundus images, reaching accuracy and precision scores of 97.47%. Lai et al. [11] developed a custom CNN architecture comprising seven layers—including convolutional, max-pooling, flatten, and dense layers for cataract detection from digital camera images, achieving outstanding accuracy and recall scores of 98.5% and 97.9%, respectively. Weni et al. [12] introduced a CNN-based method incorporating dropout regularization to mitigate overfitting, obtaining an accuracy of 88%. Further, Ganokratanaa et al. [13] compared a LeNet-based CNN to a traditional Support Vector Machine (SVM) classifier, with their LeNet-CNN approach yielding an impressive 96% accuracy.

While significant progress has been made in automating cataract detection, several challenges persist. Critical areas for future research include enhancing prediction accuracy, minimizing model overfitting and computational costs, and improving generalizability by testing the model on different geographical locations.

This paper proposes a hybrid, computationally efficient approach that integrates deep feature extraction through advanced CNN architectures, specifically ConvNeXtXLarge, with traditional classification methods, including Multi-Layer Perceptron (MLP) [14], K-Nearest Neighbors (KNN) [15], and RandomForest classifier [16]. Evaluating the model performance not only on standard benchmarking datasets but also on independently acquired hospital datasets provides a more rigorous and realistic assessment of its generalization capabilities.

III. Methods and Materials

A. Dataset

This study uses the Ocular Disease Intelligent Recognition (ODIR) dataset [17] from Shanggong Medical Technology Co., Ltd. It is a structured ophthalmic database of 5,000 patients with age, color fundus photographs from left and right eyes and doctors' diagnostic keywords from doctors. This dataset represents a real life set of patient information collected by Shanggong Medical Technology Co., Ltd. from different hospitals and medical centers in China. It has images of normal eyes and images of eyes with cataract. It is then randomly divided into three parts, with 80% of the images being used for training, 10% for validation and the remaining 10% for testing the accuracy of the model. These three datasets contain nearly equivalent numbers of normal and cataract eye images. This study also utilizes additional fundus images of normal (40 patients) and cataract (10 patients) eyes obtained from GSVM Medical College, which is a public medical college in Kanpur, India. This dataset represents a real life set of fundus images of Indian patients. It is used only as a testing dataset to assess the generalizability of the model trained on the ODIR dataset. Figure 1 shows the retinal fundus images of normal and cataract eyes. Both the ODIR and GSVM datasets used in this work are publicly available and fully comply with Health Insurance Portability and Accountability Act (HIPAA) in protecting patients' health information.

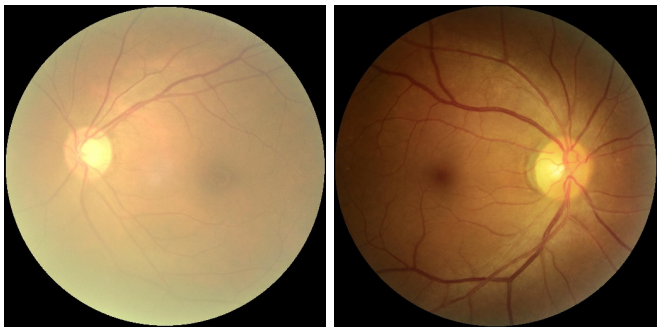


Figure 1. Eye Fundus Images. (a) Cataract. (b) Normal.

B. Data Preprocessing

A pre-processing stage is crucial in standardizing the input of fundus images obtained from multiple sources, as they may have different characteristics. The ODIR dataset is from several medical institutions in China where fundus images are captured by various cameras in the market, such as Canon, Zeiss and Kowa, resulting in varied image resolutions. The GSVM dataset of raw images contained irrelevant visual components like computer monitors and medical devices, which are cropped out to have only retinal fundus images. Each fundus image is then resized to 224x224 pixels for uniformity and then converted to RGB color space resulting in a three-dimensional (3D) array of 224x224x3. The three 2D arrays represent red, green and blue channels respectively. It is then converted from RGB to BGR and each color channel is zero-centered with respect to the ImageNet dataset, without scaling, as required by ConvNeXtXLarge.

C. Feature Extraction

ConvNets have always been popular for computer vision related tasks due to their inherent inductive biases like translation equivariance and sliding window strategy. Translation equivariance is important for object detection and sliding window strategy allows neighbors to share computations, which is essential for visual processing. Recently, ViTs have entered this space with better accuracy rate than traditional ConvNets and are getting increasingly dominant. The primary reason for their superiority is their global attention design, which has quadratic complexity with respect to the input image size and can quickly become unmanageable with higher resolution images. To address this limitation, hierarchical Transformers like Swin Transformer have been developed, which incorporates some of the inductive biases of ConvNets like sliding window strategy. But the resulting design is still complex, requiring significantly more computational resources than ConvNets. ConvNeXtXLarge model [18] is an enhanced traditional ConvNets, which retains the design simplicity of convolution and then incorporates features like depthwise convolution, inverted bottleneck and large kernel sizes taken from the architecture of hierarchical Transformers. It outperforms vanilla ViTs, while maintaining the simplicity and efficiency of standard ConvNets and for these reasons is used here to extract image features. The weights used are from the pretraining of this model on the ImageNet-21k dataset and then fine-tuned on the ImageNet-1k dataset. The top layer of the ConvNeXtXLarge model is replaced with a global average pooling layer to avoid overfitting. It also ensures that some spatial information is retained by averaging each feature map, which allows for higher versatility across different input variations evident in datasets of this study. It also helps in keeping the architecture simple, which leads to faster featurization and less computational resource consumption. Essentially, the idea was that intially featurizing the images and then classifying them based on these numerical features would provide better accuracies than solely applying a CNN-variant. By stripping off the classifier head of the ConvNextXlarge model, the model extracts 2048 features from the images. As opposed to the classifier head making the prediction, additional predictive models were added on the ConvNextXLarge model to improve accuracy. The set of features extracted from the dataset are randomly reshuffled to avoid subsequent classification models from learning the patterns based on the order of the images in the dataset.

In order to determine the importance of the ConvNextXLarge featurizer in the proposed pipeline, an ablation study was conducted. One of the variants of the pipeline tested in the ablation study involved substituting the ConvNextXLarge model for DenseNet-201, another featurizer model, pairing DenseNet-201 with MLP. A DenseNet-201 model, pretrained on the ImageNet-1k dataset, is a CNN that has dense connectivity, meaning that each layer takes input from all the layers that were before that particular layer and provides output to all layer subsequent to that particular layer. The model has several dense blocks, with each block containing a certain amount of layers. Each block is separated from other blocks by transition layers, which are composed of a 1 by 1 convolution layer followed by a pooling layer, with the goal being to compress the feature maps. Due to these characteristics of the DenseNet-201 model, it has several advantages, such as reducing the occurrence of vanishing gradients and cutting down on parameter redundancy. The DenseNet-201 model extracts 1920 numerical features from the fundus images and also has its top layer stripped away, replaced with a global average pooling layer for similar reasons as the ConvNextXLarge model. The reason for choosing DenseNet-201 as the substitute for ConvNextXLarge in this ablation study lies in the fact that DenseNet-201, being a classic CNN, lacks the ViT-like enhancements that ConvNextXLarge possesses, such as inverted bottleneck, depthwise convolution, and large kernels. The ablation study, in part, aims to determine the effect of removing ViT-like enhancements on the performance of the model.

D. Classification

The extracted features are then used to train MLP, KNN, and Random Forest classifiers. The MLP Classifier is a feedforward neural network having at least three layers, an input layer, one or more hidden layers, and an output layer. Each node in the input layer corresponds to a feature in the feature map.

There can be any number of hidden layers and each can have any number of nodes. They calculate a weighted sum of the inputs followed by an activation function, which adds bias and introduces nonlinearity. The output layer generates the final prediction, so in this study acts as the binary classifier having two nodes for "Normal" and "Cataract" prediction. The predicted output is compared to the actual label using a loss function and to minimize its value, weights of the nodes and activation function are adjusted during backpropagation.

The KNeighborsClassifier is a simple yet powerful lazy learning algorithm. It preserves the entire training data from the training phase and uses it to classify based on similarity measures. The class of a data point is determined by the majority or average of its K neighbors, which are found based on a distance metric.

The RandomForest Classifier is an ensemble tree learning algorithm. During the training phase, it creates a number of decision trees using random subsets of the features from the feature map. Each individual tree makes its prediction and the final prediction is determined by voting where the most frequently predicted result is chosen. These three different classifiers are paired with ConvNextXLarge and compared in an ablation study to isolate the contributions of the MLP classifier in the proposed pipeline and to assess its individual importance to the performance of the proposed pipeline.

Figure 2 demonstrates the architecture of the proposed method, including preprocessing, featurization, and the different classification models that ConvNeXtXLarge is paired with.

In order to evaluate the performance of the proposed methodology, the results have been compared with traditional CNN models like ResNet50 [19], EfficientNetb2 [20], and MobileNetv2 [21] as well as computationally costly ViTs, such as Swin transformer [22] and vanilla ViT [23].

ResNet50 is a type of Deep Convolutional Neural Network (DCNN) with 50 layers, a part of a group called Residual Networks. It uses special connections, known as residual connections, to help gradients flow effectively and overcome issues like vanishing gradients, making it reliable for tasks like image classification. EfficientNetB2 is a CNN that enhances width, depth, and resolution of images through a calculated scaling method. As the third model in the EfficientNet series, it expands upon previous models (B0 and B1) with more layers, wider channels, and higher resolution. This allows EfficientNets to achieve high accuracy with fewer parameters and less computational demand compared to older models like ResNet. ViTs break down images into small, equal sized patches, flatten these patches, and feed them into a global attention module, using positional embeddings for each patch. This approach allows ViTs to focus on broad patterns in images, often surpassing traditional CNNs, especially in large-scale datasets. SwinTransformers are a specialized version of ViTs that apply attention mechanism within local windows that are shifted across the image. By incorporating convolutional layers, they create hierarchical feature maps similar to those in CNNs, which helps them be effective in classification tasks.

E. Experimental Procedures

For training MLP, the hyperparameter learning_rate_init, which controls the step size in updating the weights during backpropagation to minimize loss function, is evaluated for values 0.01, 0.05, 0.001, 0.0001, 0.00001, 0.1 and 0.000001. For each value of learning_rate_init, the hyperparameter max_iter, which is the epoch value, is evaluated for values ranging from 10 to 110 increasing in intervals of 10.

For KNN, the hyperparameter n_neighbors, which is the number of nearest neighbors to consider in deciding the class of a data point, is evaluated for values ranging from 1 to 15.

For RandomForest, the hyperparameter max_depth, which is the depth of the decision tree, is evaluated for values in the range 1 to 7. For each value of max_depth, the hyperparameter n_estimators, which is the number of the decision trees in the forest, is evaluated for values ranging from 10 to 110.

The performance of the proposed methodology is also benchmarked against the MobileNetV2, ResNet50, EfficientNetB2, SwinTransformer, ViT model, which has a similar architecture, to assess improvements in detection accuracy, highlighting the effectiveness of the featurization over classification methods in enhancing cataract prediction performance. The hyperparameter epochs values ranging from 10 through 50 with the learning_rate hyperparameter ranging from 0.000001 to 0.05, depending on the model, are used to fine-tune the models.

Performance metrics like accuracy, precision and recall are used to identify the most effective model. Accuracy is the percentage of true prediction out of the total prediction. Precision is the percentage of true positive prediction (i.e., cataract eye) out of total positive prediction. Recall is the percentage of true positive prediction out of the total positive samples. For tasks like medical diagnosis, the cost of false negative prediction (i.e., cataract eye predicted as normal) is the highest, so a higher recall value is given the highest precedence followed by precision and then accuracy.

Finally, the computational efficiency and speed of the proposed pipeline was quantified by measuring the wall time (s) as well as CPU time (s) of the entire pipeline, including both featurization and classification, during training, validation, and testing. Wall time is the elapsed time from when the task began to when it ended, taking into account computation of the models, waiting for inputs and outputs, network delays, and several other real world factors to represent the real world time that a user must wait for the result. On the other hand, CPU time is the amount of time that the computer processing spent executing the pipeline. For the validation, testing, and collected testing dataset, two additional performance metrics were measured: inference latency (s) and throughput (samples/s). Inference latency is the total elapsed time it takes a trained ML model to take in a singular input, in this case a fundus image, and make a prediction. Inference latency is critical in this context as it measures how long a user may have to wait for a diagnosis for cataract, with speed being essential to lessen the impact of cataract. Throughput, which is mathematically the inverse of inference latency, measures the amount of predictions

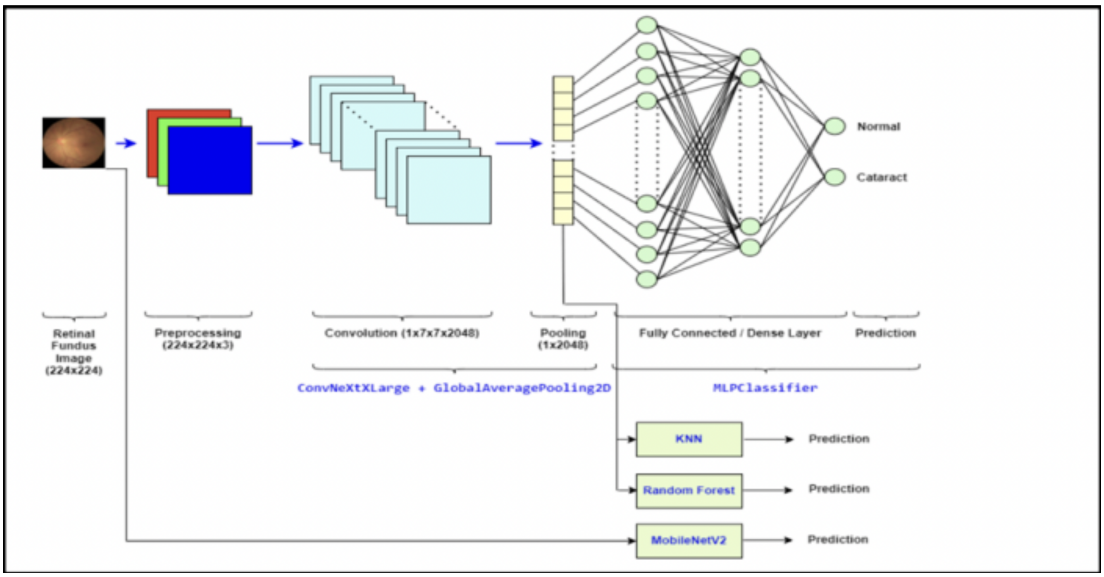


Figure 2. Architecture of proposed hybrid approach.

a trained ML model can make within a certain unit of time. Throughput is essential in this context as it measures the efficiency of the pipeline, which is essential when processing batches of patients, a common occurrence in overwhelmed and understaffed rural clinics.

IV. Results

In the case of MLP, the highest accuracy of 98.28% is achieved, with minimum number of epochs, when epochs are 40 and learning rate is 0.01. In Figure 3, the performance of MLP based on different pairs of epochs and learning rates is shown. Other combinations of epochs and learning rate result in this accuracy, including 50 epochs and 0.01 learing rate, 60 epochs and 0.01 learning rate, 80 epochs and 0.01 learning rate, and 70 epochs and 0.05 learning rate. This is the highest validation accuracy, higher than every model except RandomForest.

In the case of KNN, the highest accuracy of 96.55% is achieved with 2 neighbors and the lowest accuracy of 89.50% is achieved when the number of neighbors increases to 5, 6, 7, and 13. This performance is demonstrated by Figure 4a based on different numbers of neighbors. For RandomForest, the highest accuracy of 98.28% is achieved with 20 trees and a depth of 2. This performance is demonstrated by Figure 4b based on different pairs of trees and depths.

In the ablation study, which is documented in Table I, the combination of ConvNextXLarge and MLP yielded an accuracy of 98.28% in the validation dataset, 92% in the testing dataset, and 94% in the collected GSVM testing dataset. These were the highest recorded accuracies out of all combinations of components tested in the ablation study. The combination of ConvNextXLarge and KNN yielded a testing accuracy of 90.40% and a GSVM testing accuracy of 75.50%. The combination of ConvNextXLarge and RandomForest yielded a testing accuracy of 88.50% and a GSVM testing accuracy of 79.60%. The combination of DenseNet201 and MLP yielded a validation accuracy of 93.10%, a testing accuracy of 86.30%, and a GSVM testing accuracy of 91.80%.

TABLE I. Ablation Study Accuracy Results

Variant	***Val.(%)***	***Test(%)***	***GSVM(%)***
ConvNextXLarge+MLP	98	92	94
ConvNextXLarge+KNN	97	90	76
ConvNextXLarge+RF	98	89	80
DenseNet-201+MLP	93	86	92

For each classification model paired with ConvNeXtXLarge, each of the graphs depicting their validation accuracy has learning rate on the horizontal axis, validation accuracy on the vertical axis, and each line represents a different epoch number in the sequence 10, 20, 30, 40, and 50. The performance metrics of each model on the testing portion of the ODIR dataset as well as on the collected GSVM dataset are depicted in Table II and Table III, respectively.

For EfficientNetB2, according to Figure 5a, the highest validation accuracy was 96.55% at 30 epochs and 0.01 learning rate. For ResNet50, according to Figure 5b, the highest validation accuracy was 98.28% at 10 epochs and 0.001 learning rate. For MobileNetv2, according to Figure 5c, the highest validation accuracy was 94.83% at 30 epochs and 0.005 learning rate.

Moving on to ViTs, for Swin Transformer, the highest validation accuracy, according to Figure 6a was 94.83% at 30 epochs and 0.0001 learning rate. Finally, for the vanilla ViT, according to Figure 6b the highest validation accuracy was 98.04% at 50 epochs and 0.000001 learning rate.

In Table II and Table III, the performance of the classification models on ODIR and GSVM test dataset is summarized, respectively. The testing of the ODIR test dataset on the MLP model resulted in an accuracy score of 92.30%. The same

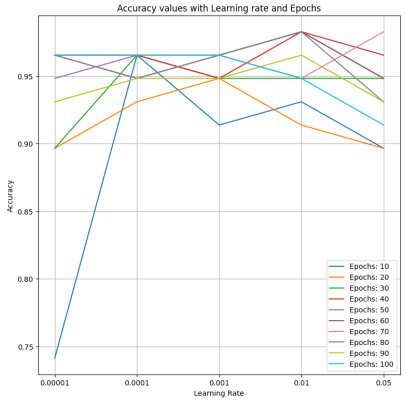
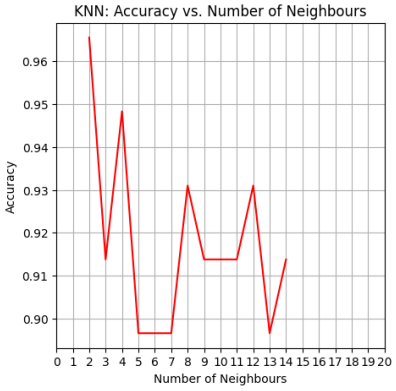
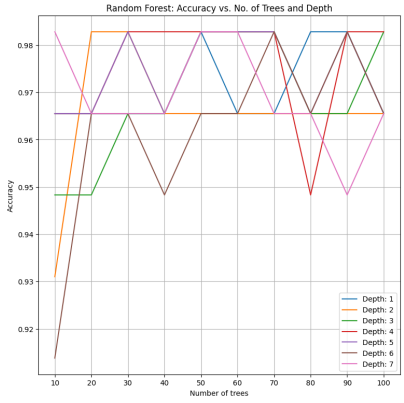


Figure 3. Hyperparameter tuning of the MLP model.

model is also tested using the GSVM test dataset resulting in an accuracy of 94%. Each of these is depicted by Figures 7a and 7b, respectively.



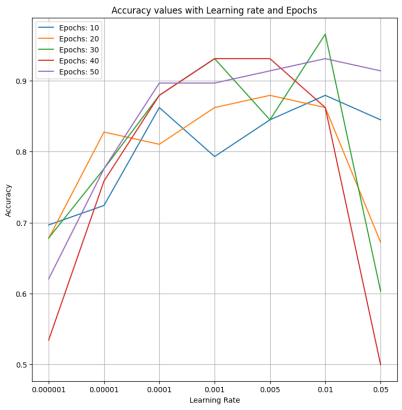
(a) K-Nearest Neighbors



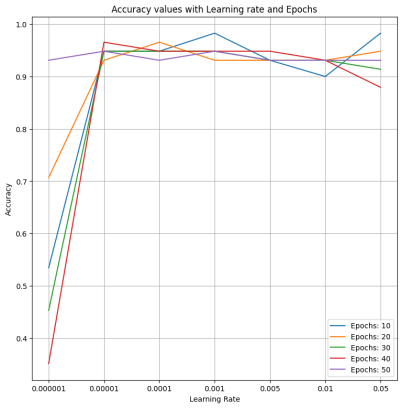
(b) RandomForest

Figure 4. Validation accuracy of ML models across their parameters.

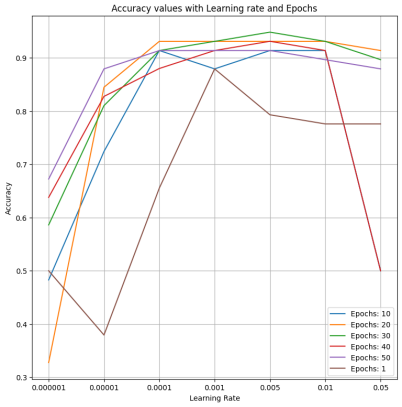
The computer system performance metrics for the proposed pipeline when applied to each dataset were measured to quantify its computational efficiency. For the training dataset, the recorded CPU time was 2 hours, 43 minutes, and 43 seconds. The recorded wall time was 2 hours, 17 minutes, and 30 seconds. The wall time of the validation dataset was 21



(a) EfficientNetB2 model

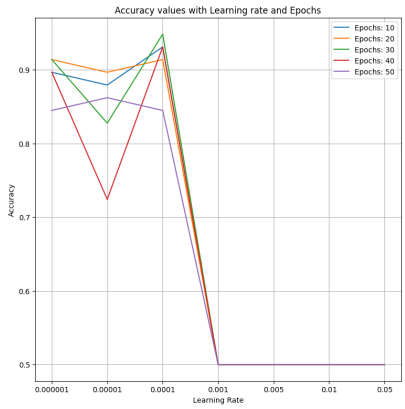


(b) ResNet50 model

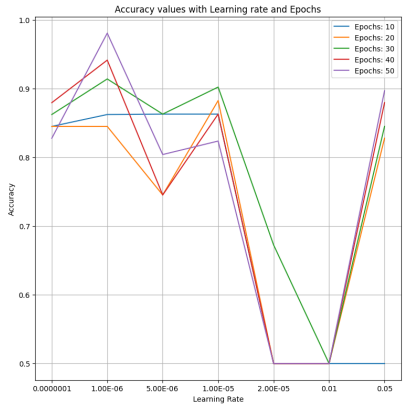


(c) MobileNetv2 model

Figure 5. Validation accuracy of CNN models across learning rates and epochs.

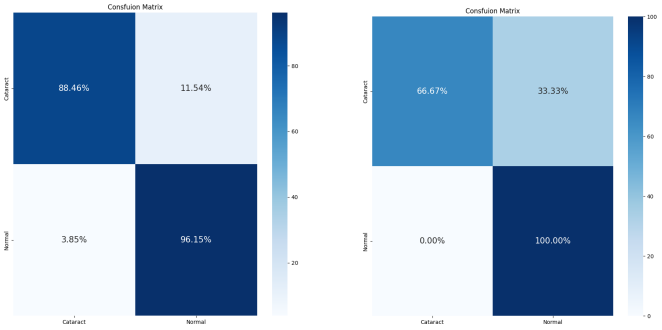


(a) Swin Transformer



(b) Vanilla ViT

Figure 6. Validation accuracy of ViT models across their parameters.



(a) ODIR Testing Portion (b) Collected GSVM

Figure 7. Performance of proposed model on testing datasets.

minutes and 5 seconds, yielding an average inference latency, for the validation dataset, of 21.8 seconds per fundus image and a throughput of 0.046 fundus images per second. For the testing dataset, the wall time was 18 minutes and 9 seconds, yielding an average inference latency of 21.4 seconds per fundus image and a throughput of 0.047 fundus images per second. For the collected GSVM testing dataset, the wall time was 15 minutes and 4 seconds, yielding an inference latency of 18.4 seconds per fundus image and a throughput of 0.054 fundus images per second. All these measurements are referenced from Table IV below.

TABLE II. PERFORMANCE OF CLASSIFICATION MODELS ON ODIR TEST DATASET

Model	***Acc.(%)***	***Prec.(%)***	***Rec.(%)***	***F1.(%)***
MLP	92	92	92	92
ResNet50	90	90	90	90
EfficientNetB2	92	92	92	92
ViT	90	92	90	90
Swin Trans.	92	92	92	92
MobileNetV2	86	86	86	86

TABLE III. PERFORMANCE OF CLASSIFICATION MODELS ON GSVM TEST DATASET

Model	***Acc.(%)***	***Prec.(%)***	***Rec.(%)***	***F1.(%)***
MLP	94	97	83	88
ResNet50	92	93	92	92
EfficientNetB2	94	95	94	94
ViT	86	93	61	64
Swin Trans.	96	93	93	93
MobileNetV2	90	94	72	78

TABLE IV. COMPUTER SYSTEM PERFORMANCE METRICS OF PROPOSED PIPELINE

Dataset	***Wall time(s)***	***CPU time(s)***	***Inference Latency(s)***	***Throughput***
Train	8250	9823	-	0.058
Val.	1079	1275	21.8	0.046
Test	850	1089	21.4	0.047
GSVM	904	1059	18.4	0.054

V. DISCUSSION | EVALUATION

The combination of ConvNeXtXLarge and GlobalAveragePooling2D for featurization with MLP Classifier as the classifier resulted in the highest validation accuracy, which was higher than similar prior studies, albeit with a different and smaller dataset. Convolution does not have high computational costs like global attention design of ViTs, which requires computationally expensive global attention modules. Additionally, just one hidden layer with 100 nodes in MLP classifier also helps in keeping the architecture simple, requiring less resources. KNN and RandomForest also had decent validation accuracies, with RandomForest actually matching MLP in validation accuracy. However, KNN and RandomForest require comparatively more resources than the proposed method.

For KNN, more neighbors after K=2 results in a drop in accuracy. This is likely due to the bias-variance tradeoff involved with involving more neighbors. In the case of RandomForest, there is not much variation in accuracy for different combinations of depth and number of trees. Every evaluated combination results in the accuracy within the range of 91.38% to 98.28%. It is also evident during tuning that computational requirements of this classifier are directly proportional to the number of trees.

In the case of MLP, the highest accuracy of 98.28% is achieved when epochs are 40 and learning rate is 0.01. The accuracy increases with the increase in epochs for every single value of learning rate. After an optimal value of epochs is reached, the accuracy plateaus for each learning rate. The lowest learning rate of 0.00001 has the maximum deviation in accuracy ranging from around 74% to 97% as epochs increase. The highest learning rate of 0.05 has second highest deviation in accuracy ranging from around 89% to 98% with increase in epochs. It demonstrates that too low or too high learning rate during backpropagation can adversely impact accuracy. For higher learning rates, such as 0.05, weights are updated too quickly, resulting in oscillations around convergence and thus making number of epochs cause large variation in validation accuracies. However, since 0.05 as a learning rate is not too large, the deviation in accuracy based on epochs is not too drastic and maximum accuracy of 98.28% is still achievable with this learning rate when paired with enough epochs, in this case 70.

In the ablation study, given that the combination of ConvNextXLarge as the featurization model and MLP as the classifier model yielded the highest accuracies in classifying the fundus images among all three of the datasets (validation, testing, and GSVM testing), clearly, both of these models contribute heavily to the performance of the pipeline. Likely, ConvNextXLarge and MLP outperformed the pairing of ConvNextXLarge and KNN because when there is a large number of features involved, such as 2048 features, the Euclidean distances that KNN calculates between points to classify a point tend to concentrate as close neighbors and far away neighbors appear to be roughly equidistant from the datapoint currently being classified, thus making KNN's prediction inaccurate. MLP likely trumped RandomForest's performance as well since RandomForest uses the aggregate prediction of axis-aligned decision trees, which perform optimally on tabular data but often stumble when trying to classify an image based on the smooth, high-dimensional feature maps produced by CNNs such as ConvNextXLarge. The ConvNextXLarge model likely outperformed the DenseNet-201 model due to its ViT-like enhancements, such as depthwise convolutions and inverted bottleneck, which boosted its accuracy without requiring computationally expensive global attention modules.

The evaluation of various models, including MobileNetv2, ResNet50, EfficientNetB2, ViT, and SWINTransformer, across different learning rates and epochs shows the impact of hyperparameter tuning on model performance. Lower learning rates frequently demonstrate slower convergence, leading to suboptimal accuracy, as observed with MobileNetv2 and similarly noted in other models like EfficientNetB2 and ViT. In contrast, moderate learning rates, particularly around 0.0001 to 0.005, consistently yield higher stability and precision, resulting in efficient convergence without the risk of overshooting the global minimum, a pattern evident across both ResNet50 and MobileNetv2. High learning rates, such as 0.05, introduce significant volatility, destabilizing the training process as illustrated by diminished results in MobileNetv2, ViT, and EfficientNetB2.

For the ODIR testing performance of each model, the proposed hybrid approach of pairing ConvNeXtXLarge to featurize fundus images and using MLP to classify each image based on these features got the highest accuracy of 92% and highest recall of 92%. These performance metrics were matched by Swin Transformer as well as EfficientNetB2. The proposed approach likely performed one of the best because ConvNeXtXLarge leverages the strengths of both the CNNs (ConvNet / CNN) and hierarchical ViTs for featurization. The inherent inductive biases of CNN, like translation equivariance and sliding window strategy, work together with the depthwise convolution and inverted bottleneck of ViTs to extract image features. Thus, strong spatial representations are fed into MLP, which allows MLP to make accurate predictions.

EfficientNetB2 also shared the same high performance metrics due to its compound scaling of fundus images, which effectively scales the width, depth, and input resolution of inputted images using a user-specified scaling coefficient. This scaling allows it to capture finer details and improve representation of images. For Swin Transformer, it first splits the images into patches that it then flattens into feature vectors. By applying self-attention to small local windows that are then shifted across the image to ensure cross-window communication, the model is capable of paying attention to local features as well as maintain global awareness, thus allowing it to notice small features in the fundus images and generalize better on new datasets.

On the GSVM dataset, the differences between each of these three respective model were more. While the hybrid proposed approach, EfficientNetB2, and Swin transformer had similar accuracies despite the proposed approach's simplified architecture, Swin transformer and EfficientNetB2 had higher recall values than the proposed approach did. This is likely because the MLP head, with only 100 nodes in one hidden layer, was unable to properly detect all positives, hence its lower recall. The model likely requires more training on noisy real world hospital data to be able to properly generalize to real world datasets and their quality issues. The hierachial window-based self-attention of Swin Transformer and the compound scaling of EfficientNetB2 likely allowed each model to notice small details in fundus images of the training dataset, thus allowing them to generalize to hospital data even with its flaws.

The vanilla ViT likely performed much worse than the others because ViTs, due to their global attention modules, require large training datasets to properly generalize to other datasets and often miss small localized details. MobileNetv2 is a lightweight model that does not translate well to datasets that have a lot of noise. While the proposed approach has a lower recall than Swin Transformer and EfficientNetB2 on real-world hospital data due to the relatively low quality of the data, the proposed approach still performs on par with, and sometimes better than (in the case of validation dataset) state of the art models on quality datasets that resemble its training dataset, and further training on more real world hospital data will likely allow it to generalize to imperfect hospital data.

Additionally, since MLP has only one layer with 100 nodes and ConvNeXtXLarge lacks computationally expensive attention-based modules, the proposed approach is computationally more efficient than other models, including vanilla ViTs with their global attention modules. Finally, the proposed approach, due to the replacement of ConvNeXtXLarge's FTC with a Global Average Pooling Layer, is able to reduce overfitting of the model on the training dataset.

Considering the computer system performance metrics for the proposed pipeline, the inference latency times, computed for each of the three non-training datasets, are roughly similar to each other. Taking into account the relative number of images in each of the three datasets, the average inference latency is 20.6 seconds, which is much faster than the standard 30-60 minutes a standard cataract examination may take involving a specialist and advanced equipment. For each dataset, the CPU time exceeded the wall time, regardless of the other delays that the wall time considers. This is because of the use of parallel processing when running the pipeline, using multiple cores at the same time as opposed to a single-threaded process.

The limitations of this study include a lack of large datasets to train classifier models and the local nature of the datasets. The study uses ODIR dataset for training, which has just a few hundred fundus images for cataract as opposed to thousands of images typically required to train models efficiently and avoid overfitting. Both the ODIR and GSVM fundus image datasets are of patients from south Asia. It is not clear if the accuracy of the model will be the same if tested on fundus image datasets from other parts of the world. Further study and more diverse sources of datasets are required to address these aspects.

VI. Conclusion and Future Work

This study establishes a robust framework for cataract detection using deep learning and traditional classifiers, showcasing strong performance on both benchmark and hospital-based datasets. The few seconds that it takes the web app to predict the presence of cataract from a fundus image is much faster than skilled medical personnel, using advanced detection equipment, would be able to without even considering the fact that these personnel can only visit a clinic once every few weeks or even months due to understaffing. Nonetheless, there remain several promising avenues for future research. Expanding the dataset size with diverse fundus images from various geographic locations will help improve the generalizability and robustness of the model across different populations. Moreover, incorporating techniques like transfer learning from larger ophthalmologic datasets or integrating advanced data augmentation methods could further mitigate overfitting and improve performance.

Additionally, incorporating explainability methods and visualization tools to interpret model predictions could provide clinicians greater confidence in AI-assisted diagnoses, promoting better clinical acceptance and decision-making.

Integration of multimodal data, combining fundus imaging with other diagnostic modalities, such as Optical Coherence Tomography (OCT) or clinical patient histories, could further enhance diagnostic accuracy and reliability. Furthermore, longitudinal studies assessing the real-world clinical impact and economic feasibility of deploying this AI-based cataract detection system will be crucial to translating research advancements into practical healthcare improvements.

With regards to the use of this web app by clinicians, a possible improvement to the web app could be "Clinician-in-the-Loop Testing," where a clinician could participate in the predictions made by the web app by having the web app identify certain fundus images that it is unsure of and thus passing them off to the clinician for a more detailed review. The rate at which clinicians accept or reject the predictions of the web app could also be recorded as another metric for performance. Finally, efficiency benchmarks can be used to demonstrate the efficiency and speed of the model on different hardware. For instance, a possible benchmark is throughput, which is the number of fundus images that can be classified within a certain amount of time. Other measures, such as CPU usage, memory consumption, and power draw of the web app when predicting can also be measured. These benchmarks are heavily affected by the quality of hardware used. Since rural clinics will often be limited to hardware with limited computing power, these metrics help further quantify the computing resources that the pipeline may require to ensure that it does not exceed computational limits.

References

[1] D. Pascolini and S. Mariotti, "Global estimates of visual impairment: 2010", *The British journal of ophthalmology*, vol. 96, pp. 614–8, Dec. 2011. DOI: 10.1136/bjophthalmol-2011-300539.

[2] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs", *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[3] Y. Bao *et al.*, "Self-adaptive transfer learning for multicenter glaucoma classification in fundus retina images", in *Ophthalmic Medical Image Analysis: 8th International Workshop, OMIA 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 8*, Springer, 2021, pp. 129–138.

[4] S. Lu *et al.*, "Deep learning-driven approach for cataract management: Towards precise identification and predictive analytics", *Frontiers in Cell and Developmental Biology*, vol. 13, p. 1 611 216, 2025.

[5] M. S. Ali and M. Islam, "A hyper-tuned vision transformer model with explainable ai for eye disease detection and classification from medical images", *BS thesis, Faculty of Engineering and Technology Islamic University*, 2023.

[6] S. O. Purba *et al.*, "Classification of eye diseases in humans using vision transformer architecture model", in *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, 2024, pp. 71–75.

[7] S. D. Gummadi and A. Ghosh, "Classification of ocular diseases: A vision transformer-based approach", in *International Conference on Innovations in Computational Intelligence and Computer Vision*, Springer, 2022, pp. 325–337.

[8] D. Kumar, B. Bakariya, C. Verma, and Z. Illes, "Cataract disease identification using transformer and convolution neural network: A novel framework", in *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, IEEE, 2023, pp. 1230–1235.

[9] J. Wang *et al.*, "A transformer-based knowledge distillation network for cortical cataract grading", *IEEE Transactions on Medical Imaging*, vol. 43, no. 3, pp. 1089–1101, 2023.

[10] M. S. M. Khan, M. Ahmed, R. Z. Rasel, and M. M. Khan, "Cataract detection using convolutional neural network with vgg-19 model", in *2021 IEEE World AI IoT Congress (AIIoT)*, IEEE, 2021, pp. 0209–0212.

[11] C.-J. Lai *et al.*, "The use of convolutional neural networks and digital camera images in cataract detection", *Electronics*, vol. 11, no. 6, p. 887, 2022.

[12] I. Weni, P. E. P. Utomo, B. F. Hutabarat, and M. Alfalah, "Detection of cataract based on image features using convolutional neural networks", *Indonesian Journal of Computing and Cybernetics Systems*, vol. 15, no. 1, pp. 75–86, 2021.

[13] T. Ganokratanaa, M. Ketcham, and P. Pramkeaw, "Advancements in cataract detection: The systematic development of lenet-convolutional neural network models", *Journal of Imaging*, vol. 9, no. 10, p. 197, 2023.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[16] L. Breiman, "Random forests", *Machine learning*, vol. 45, pp. 5–32, 2001.

[17] G. Challenge, *Peking university international competition on ocular disease intelligent recognition (odir-2019)*, 2019.

[18] Z. Liu *et al.*, "A convnet for the 2020s", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

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

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

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[22] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

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