DeepSight: Automated Diabetic Retinopathy Screening with Machine Learning

Abstract

Diabetic retinopathy (DR) is a leading cause of preventable blindness, requiring early detection for timely intervention. This study presents DeepSight, an automated screening system using machine learning to classify DR severity levels from retinal fundus images. Multiple modeling approaches were explored and evaluated, including end-to-end CNN models, hybrid feature extraction with traditional classifiers, and Vision Transformers (ViT). The pre-trained ViT model demonstrated the best performance, achieving 82% accuracy, 72% Kappa, and a ROC AUC of 98.9% for multi-class classification, and 96.55% accuracy with 96.06% recall for binary DR detection. To enhance interpretability, visual explanations (attention heatmaps) and natural language explanations via LLM integration (GPT-4o) were incorporated. Finally, a lightweight Flask-based web application was developed to support practical deployment and informed clinical decision-making.

The code for this project can be found on Github here: GitHub Link

1. Introduction

Diabetes is becoming an increasingly serious global health issue, with the number of people affected rising from 200 million in 1990 to 830 million in 2022 (WHO, 2023). Complications due to the disease are becoming major health issues that require effective health interventions for prevention and treatment. A specific microvascular complication of diabetes is Diabetic Retinopathy (DR). DR is a severe complication of diabetes and a leading cause of preventable blindness worldwide, affecting an estimated 103 million patients worldwide (Teo et al., 2021). Early detection and timely treatment of DR can help to significantly reduce the risk of vision impairment. However, manual screening by ophthalmologists is labor-intensive, time-consuming and subject to human error. This is especially so in regions with limited access to specialized healthcare professionals.

Convolutional Neural Networks (CNNs), a branch of deep learning, have an impressive record for applications in image analysis and interpretation, including medical imaging (Pratt et al., 2016). Based on past research, in general, the CNNs architecture is created by having many filters in one layer of Neural Network (width), deeper layers (depth), and greater resolution of input image to have better performance, such as Xception, DenseNet-201, ResNet-152, VGG-19, and NASNet-Large architecture. Among them, DenseNet stands out for its use of dense connections between layers, which promote feature reuse and alleviate the vanishing gradient problem. While scaling CNNs across these three dimensions can enhance accuracy, it often requires manual tuning, which may lead to inefficiencies. EfficientNet addresses this by employing a compound scaling method that uniformly scales depth, width, and resolution using a constant aspect ratio, leading to more balanced and efficient model performance (Tan et al., 2019).

The aim of this research is to create a classification method of diabetic retinopathy as an early detection system. This research utilizes a resized and filtered version of the dataset consisting of 3662 labelled images of which the original version of dataset is provided by the Asia Pacific Tele-Ophthalmology Society (APTOS, 2019). This research explores two different approaches. The first approach uses deep learning end-to-end, where architectures such as EfficientNet and DenseNet are employed to both extract features and directly classify images. The second approach leverages these deep learning models solely for feature extraction, followed by traditional machine learning classifiers—Support Vector Machine (SVM), Random Forest, and XGBoost—for the classification task. The third approach enhances the second by applying the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance before classification. Lastly, the fourth approach using Vision Transformer (ViT) model for end-to-end image classification. This paper details the methodologies and compares the performance of these approaches in providing a practical solution for automated retinal screening.

2. Fundamentals

2.1 EfficientNet

EfficientNet employs a principled compound scaling method to jointly expand network depth, width, and input resolution, resulting in superior accuracy-efficiency trade-offs compared to conventional convolutional architectures

(Tan & Le, 2019). Its architecture integrates mobile inverted bottleneck convolution layers and squeeze-and-excitation modules, which facilitate rich, multi-scale feature extraction while maintaining a compact parameter footprint. Although originally evaluated on general image recognition benchmarks, EfficientNet's architectural efficiency makes it well-suited for resource-constrained medical imaging tasks.

2.2 DenseNet

DenseNet introduces a hallmark dense connectivity pattern in which each layer receives inputs from all preceding layers, thereby strengthening gradient propagation, encouraging feature reuse, and reducing parameter counts (Huang et al., 2017). Its transition layers help regulate model complexity. While DenseNet was initially validated on standard benchmarks, its efficient architecture can be helpful in capturing both fine vessel structures and broader lesion patterns in retinal images.

2.3 Vision Transformers

The Vision Transformer (ViT) reimagines image classification by partitioning input images into fixed-size patches, embedding them as tokens with positional encodings, and processing these through multi-head self-attention and feed-forward layers (Dosovitskiy et al., 2021). A dedicated classification token aggregates global contextual information, which can be critical for identifying distributed features such as microaneurysms or hemorrhages. Though ViT requires substantial pretraining on large-scale datasets to generalize effectively, its fine-tuned variants have shown promise as global feature extractors in specialized domains, including retinal disease detection.

2.4 Extreme Gradient Boosting Trees

Extreme Gradient Boosting (XGBoost) is an efficient, scalable tree-boosting algorithm that builds sequential trees to minimize a regularized objective function via second-order gradient descent (Chen & Guestrin, 2016). With built-in L1/L2 regularization, histogram-based split finding, and support for out-of-core computation, XGBoost efficiently handles high-dimensional deep-learning–derived features. When applied in a hybrid pipeline, XGBoost serves as a strong second-stage classifier that captures nonlinear interactions in CNN-extracted embeddings from retinal scans.

2.5 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees on bootstrap samples while injecting randomness at each split by selecting from random feature subsets (Breiman, 2001). This mechanism improves generalization and reduces overfitting. When using features extracted from EfficientNet or DenseNet, Random Forest provides not only robust classification performance but also interpretable feature importance metrics and out-of-bag error estimates, supporting practical model validation on limited ophthalmic datasets.

2.6 Support Vector Machine

Support Vector Machine (SVM) aims to find the hyperplane that maximizes the margin between classes in a high-dimensional feature space (Cortes & Vapnik, 1995). By applying kernel functions—such as the radial basis function—SVM can capture complex decision boundaries that arise in transformed CNN feature vectors. The soft-margin formulation balances margin maximization with misclassification penalties, making SVM particularly effective in small-to-moderate-sized datasets such as APTOS, where generalization and interpretability are key.

2.7 Gray Wolf Optimizer

The Gray Wolf Optimizer (GWO) is a nature-inspired metaheuristic that mimics the leadership hierarchy and hunting behavior of grey wolves to guide the search for optimal solutions (Mirjalili et al., 2014). Using the positions of α, β, and δ wolves to direct exploration, GWO balances global and local search phases through a dynamic control coefficient. Although the original paper focused on mathematical and engineering optimization tasks, GWO's simplicity and minimal tuning make it a practical choice for hyperparameter optimization in deep learning pipelines, including those used for retinal image analysis.

2.8 Cohen Kappa Score

To assess inter-rater agreement beyond chance, Cohen's Kappa Score provides a robust statistical measure that quantifies the degree of concordance between two categorical raters while correcting for expected agreement due to randomness (Cohen, 1960). In classification tasks—particularly imbalanced or ordinal problems like diabetic retinopathy severity prediction—accuracy alone can be misleading, as it does not penalize chance-level predictions. Kappa addresses this by comparing the observed accuracy to the accuracy expected by random assignment, yielding a value between –1 and 1, where 1 denotes perfect agreement, 0 indicates chance-level agreement, and negative values reflect systematic disagreement. In our pipeline, Kappa is used to evaluate model consistency with human expert labels, offering a more nuanced reflection of classification reliability than raw accuracy—especially critical in medical imaging where class imbalance and ordinal progression are common.

3. Research Experiment

3.1 Data Description

The dataset used in this study was obtained from the APTOS 2019 Blindness Detection competition, which consists of 3,662 retinal fundus images labeled according to the severity of diabetic retinopathy (APTOS, 2019). Each image is categorized into one of five classes: No DR, Mild, Moderate, Severe, or Proliferative DR. These labels represent progressive stages of the disease, where No DR indicates a healthy retina, Mild includes early microaneurysms, Moderate shows more extensive vascular damage, Severe presents widespread blood vessel blockage, and Proliferative DR—the most advanced stage—involves abnormal new vessel growth that may cause serious vision loss.

To provide visual context, Figure 1 presents representative examples of each severity level from the dataset.

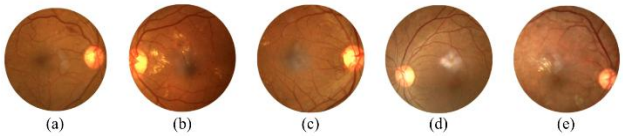

Figure 1. Sample retinal fundus images representing each diabetic retinopathy severity level: (a) No DR, (b) Mild, (c) Moderate, (d) Severe, (e) Proliferative DR.

As shown in Table 1 below, the dataset exhibits a highly imbalanced class distribution, with a significantly larger number of images labeled as No DR compared to other severity levels such as Severe and Proliferative DR. This class imbalance can hinder the model's ability to learn equally from all categories, often leading to biased predictions that favor the majority class.

Table 1. Distribution of Retinal Fundus Images Across Diabetic Retinopathy Severity Levels in the APTOS 2019 Dataset

Severity Level	Number of Images
No DR	1805
Mild	370
Moderate	999
Severe	193
Proliferate	295

3.2 Data Preprocessing

As part of the data preprocessing pipeline, all images were passed through a filtering and resizing process to standardize their quality and dimensions. Each image was read from the original dataset, and a Gaussian blur was applied and blended with the original image. This filtering technique enhances local contrast, making subtle retinal features such as microaneurysms, hemorrhages, and neovascularization more visible, which is especially valuable for improving feature extraction in later stages.

Following enhancement, each image was resized to 224×224 pixels to ensure a consistent input size suitable for convolutional neural networks, which typically require fixed input dimensions. Standardizing image size and quality helps the model learn more effectively by reducing variation due to irrelevant visual noise and differing resolutions. The preprocessed images were then saved in a new directory with filenames indicating their corresponding class labels.

To illustrate the effect of the preprocessing step, Figure 2 presents a visual comparison between the original retinal images and their corresponding enhanced versions. As shown, the Gaussian filtering significantly improves the contrast and clarity of retinal features, making lesions and vascular abnormalities more distinguishable for downstream analysis.

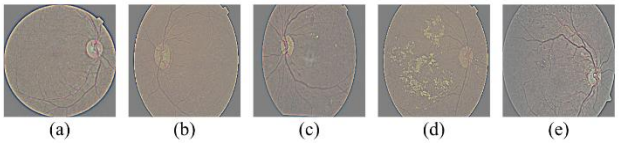

Figure 2. Preprocessed retinal fundus images representing each severity level of diabetic retinopathy: (a) No DR, (b) Mild, (c) Moderate, (d) Severe, and (e) Proliferative DR. All images have been enhanced using Gaussian filtering and resized to 224×224 pixels to improve visual clarity and standardize input dimensions.

Following preprocessing, the dataset was randomly split into training (70%), validation (15%), and testing (15%) sets to support model training, hyperparameter tuning, and unbiased performance evaluation, respectively. The preprocessed images were then saved into a new directory with filenames indicating their corresponding class labels.

3.3 Modeling

3.3.1 End-to-End Deep Learning Classification

In this approach, DenseNet121 was selected as a base architecture for transfer learning due to its efficient feature reuse and gradient flow via dense connectivity, offering a strong balance between performance and computational cost. This makes it particularly effective on small to medium sized datasets by reducing overfitting. The training was done in 2 stages.

In the initial stage, 1024-dimensional features were extracted from the frozen DenseNet121 backbone and passed through a custom classification head with three dense layers (256 → 128 → 64), using Swish activation, strong L2 regularization (1e-2), batch normalization (momentum=0.95), and dropout (0.4, 0.3, 0.4).

To address class imbalance, focal loss with class-specific alpha weights was used, supplemented by dynamic class

reweighting. Weights were further manually boosted for underrepresented yet clinically important classes (moderate ×1.6, proliferative ×1.4, severe ×1.2). Label smoothing and a confidence penalty were added to improve generalization.

RMSprop was chosen over Adam for stability in non-stationary settings, combined with cosine annealing (1e-3 → 1e-5 over 50 epochs) and early stopping.

In the second fine-tuning stage, the last 30 layers of DenseNet121 were unfrozen. A revised head (512 → 256) with ReLU activation, lighter L2 regularization (5e-5, 1e-4), batch normalization, and dropout (0.3, 0.2) was employed. Fine-tuning used a lower learning rate (1e-5) with RMSprop (centered=True) and exponential decay (0.9^epoch) for gradual convergence. With this two-stage fine-tuning approach, it prevents the model from memorizing patterns in the training set and avoids overfitting. This could be seen from the comparable validation and test set evaluation metrics in Table 2.

Table 2. Comparison of validation and test set evaluation metrics for end-to-end DenseNet121 model classification

Dataset	Acc	Kappa	F1	ROC AUC
Validation	73.0%	59%	49.3%	85.8%
Test	75.1%	62.5%	54.0%	87.3%

3.3.2 CNN Feature Extraction with Traditional Classifiers

This approach separates the feature extraction and classification stages. EfficientNetB3 is first used to extract high-dimensional feature vectors from the preprocessed retinal images. These extracted features are then fed into traditional machine learning classifiers, including Support Vector Machines (SVM), Random Forest, and XGBoost.

To address the issue of class imbalance in the training set, we applied the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples for underrepresented classes. This allows the classifiers to better learn patterns across all severity levels, potentially improving performance on minority classes. Table 3 presents the classification performance metrics before applying SMOTE, while Table 4 summarizes the results after SMOTE was applied. Both are for the test dataset. Comparing these tables allows us to assess the effectiveness of SMOTE in improving model performance, particularly for underrepresented categories such as severe and proliferative diabetic retinopathy.

Table 3. Performance of Traditional Classifiers Before Applying SMOTE on EfficientNetB3-Extracted Features

Model	Acc	Kappa	F1	ROC AUC
SVM	70%	56%	53%	90%
XGBoost	75%	60%	48%	88%
RF	71%	53%	32%	87%

Table 4. Performance of Traditional Classifiers After Applying SMOTE on EfficientNetB3-Extracted Features

Model	Acc	Kappa	F1	ROC AUC
SVM	75%	61%	56%	90%
XGBoost	76%	63%	55%	90%
RF	73%	58%	49%	88%

Table 3 shows the performance of SVM, XGBoost, and Random Forest before applying SMOTE, while Table 4 shows this after. From the results, we can see that SMOTE does improve the performance on almost every metric, for all 3 models.

Overall, the best performing model was XGBoost after SMOTE. This model achieved 76% accuracy, and a ROC AUC of 90%. While not extremely accurate, this does mean that 76% of all patients are being correctly classified. However, it is worth noting that F1 score across the board was quite poor, with no model exceeding 60%. This is typically due to poor recall in the model.

3.3.3 CNN Features Extraction with Wrapper-Based Feature Selection

In this approach, feature extraction is performed using both EfficientNet and DenseNet architectures. The extracted features from both models are concatenated to form a combined and more informative feature representation for each image. To reduce dimensionality and retain only the most relevant features, the Grey Wolf Optimizer (GWO) is applied as a wrapper-based. The selected features are then used as input to traditional machine learning classifiers, including Support Vector Machines (SVM), Random Forest, and XGBoost. This approach aims to eliminate redundant information and enhance both the efficiency and accuracy of the classification process. In the end there are 987 features selected out of 2304 features.

Table 5. Performance of Traditional Classifiers with Gray Wolf Optimizer as Wrapper-Based Feature Selection

Model	Acc	Kappa	F1	ROC AUC
SVM	80%	69%	65%	93%
XGBoost	79%	68%	48%	90%
RF	77%	64%	50%	90%

Based on Table 5, among the three, SVM achieved the highest overall performance, with an accuracy of 80%, a

Cohen's Kappa score of 69%, an F1-score of 65%, and a ROC AUC of 93%. These results indicate that SVM not only predicted labels correctly but also maintained strong agreement with the ground truth and effectively balanced precision and recall across all classes. XGBoost, while yielding competitive results in terms of ROC AUC (90%) and Kappa (68%), demonstrated a notably lower F1-score (48%), suggesting a tendency to favor majority classes—likely due to class imbalance in the dataset. Random Forest also maintained robust AUC and Kappa values (90% and 64%, respectively), but its F1-score (50%) was lower than that of SVM, indicating less balanced predictive performance. Overall, these findings highlight SVM as the most effective model in this setup, benefiting from both the discriminative power of kernel-based learning and the compact feature subset identified by GWO.

3.3.4 Vision Transformer for End-to-End Classification

In this final approach, rather than using convolutional layers, we apply self-attention mechanisms, treating parts of the image as tokens. We used a pre-trained ViT model (ViT-Base, Patch Size 16, 224px Input, ImageNet-21K pretrained) from Google, which was available via Hugging Face. Data in the training set was first pre-processed in the same way as before, and the model was then trained over 10 epochs, with a learning rate of 0.0001, and optimizer Adam with weight decay (AdamW).

It was found relatively quickly that the ViT model seemed prone to overfitting, as we monitored accuracy over both the training and validation loss during training. While training accuracy continued to climb, validation accuracy stayed the same and even started to perform worse after 5 epochs. We hence made use of a checkpoint at epoch 5 and used the model which had been trained over 5 epochs as our final model.

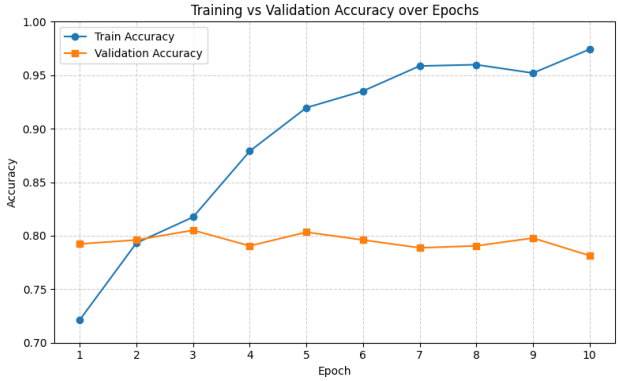

Figure 3. Training vs Validation Accuracy of ViT Model Over 10 Epochs

The model achieved an accuracy of 82%, recall of 59.5%, precision of 72.6%, F1 score of 61%, Kappa Score of 72% and ROC AUC of 98.9%.

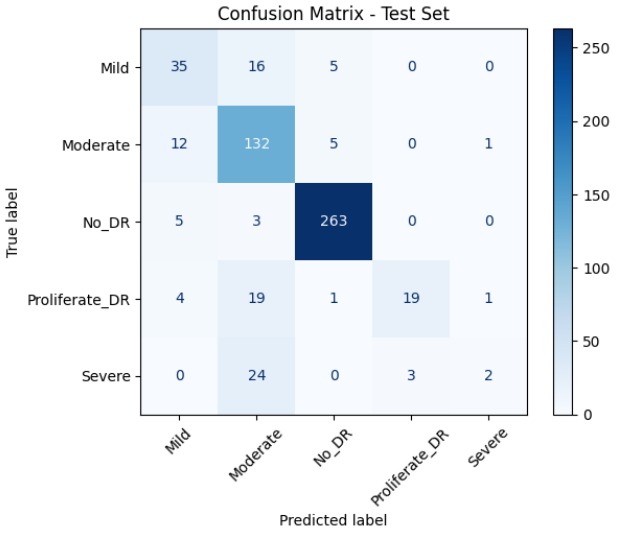

Figure 4. Confusion Matrix of ViT Model on Test Set

3.4 Evaluation

After experimenting with various modeling approaches, we selected the pre-trained ViT model as the final model. The decision was based on a comprehensive evaluation of key metrics across different modeling approaches. The following Table 6 compares the best performance results from each modeling approach across the key metrics, including accuracy, Cohen's Kappa, F1-score, and ROC AUC.

Table 6. Comparison of Best-Performing Results Across All Modeling Approaches

Model	Acc	Kappa	F1	ROC AUC
End-to-End DenseNet121	75.1 %	62.5 %	54.0 %	87.3%
CNN Feature Extraction with Traditional Classifiers (XGBoost+SMOTE)	76%	63%	55%	90%
CNN Features Extraction with Wrapper-Based Feature Selection (SVM)	80%	69%	65%	93%
End-to-End Vision Transformer	82%	72%	61%	98.9%

Based on Table 6, the pre-trained ViT model shows superior performance with highest accuracy of 82%, Cohen's Kappa of 72%, and ROC AUC of 98.9%. The results reflect that ViT model has strong agreement with the true labels and excellent ability in distinguishing between different diabetic retinopathy severity levels. Although the F1-score of ViT is slightly lower than the score from SVM with Gray Wolf Optimizer (61% to 65%), ViT's overall strong performance still provides greater confidence for real-world clinical deployment.

3.5 Discussion of Binary Classification

Given that this model is intended to be deployed in a medical context, there is concern that even slight inaccuracy could lead to catastrophic results. If a patient who does have diabetic retinopathy is told they do not have it, this could result in lack of treatment and eventual deterioration of their sight. Hence, it is worth questioning whether 82% accuracy and a 0.72 kappa score is truly sufficient for such a use case.

While we agree that 82% accuracy does leave much to be desired, we caution that this accuracy score is tied specifically to the multi-class classification, which indicates that 82% of the time, the model not only classifies whether a patient has Diabetic Retinopathy but can classify how severe their disease is.

In a real-world context, patients will likely require medical intervention at any level of DR. Hence, if the disease classifies the patient as having moderate DR when in reality they have severe DR, the consequences are not as severe, as the patient would have to seek treatment in both cases, and human doctors will have sight over the scans and be able to follow up appropriately, regardless of the model's prediction.

To this extent, we find that while multi-class classification is good to have and can improve doctors' workflow, the crucial role of the model lies in its binary classification ability (predicting whether a patient has DR or not). Hence, the suitability of the model to be deployed in a real-world use case should also be assessed on its binary classification ability. We hence reclabelled the output of the best performing model, ViT, such that all classes other than No DR were relabelled as 1, and No DR alone was labelled as 0.

The model performs well as a binary classification model, with an accuracy of 96.55%, precision of 97.1%, recall of 96.06%, F1 Score of 96.58% and ROC AUC of 98.97%. The Kappa score was also 93.09%.

In particular, we want to focus on the recall score, since recall is an important metric in a medical context. Recall measures the number of positive cases were correct identified as true positives. In other words, we are able to correctly identify 96.06% of patients who actually have DR.

These metrics' significant improvement over the multi-class classification metrics provides assurance that the model is sufficiently reliable. Despite only classifying 82% of patients into the exact right class, we know that the model is still able to correctly identify almost all patients who are suffering from DR, such that they can receive medical intervention in a timely manner.

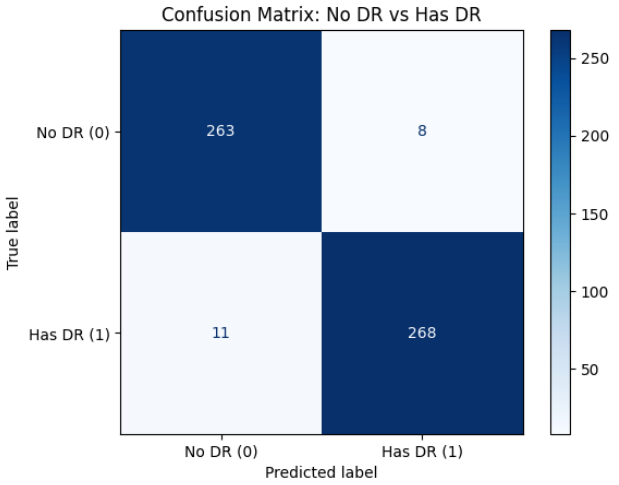

Figure 5. ViT Model Confusion Matrix for DR Detection (Binary Classification)

4. Explainability and Deployment

4.1 Grad-CAM and Attention Visualizations

Having selected our model, we move on to methods to enhance the interpretability and practical usability of our system. Deep learning models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) are inherently complex and often criticized as "black-box" models due to the difficulty in understanding how input features influence predictions. However, explainability is crucial in the medical industry, as it is important in supporting doctors' clinical judgment and decision making. Explainability further provides safeguards against misclassification and increases trust among end-users by making the model's reasoning accessible and accountable.

To address this, we employ visual explanation techniques tailored to each model class. For CNNs, we use Gradient-weighted Class Activation Mapping (Grad-CAM). This is a method that utilizes the gradients of the target class flowing into the last convolutional layer to generate a localization map. This highlights the regions of the input image that most strongly influence the output decision, offering insight into the spatial attention of the model.

On the other hand, ViTs lack convolutional layers, and are hence unable to generate Grad-CAM visualizations. Instead, we utilize attention heatmaps which visualize the

attention weights associated with the final [CLS] token, or the classification token in transformer architectures. In other words, the heatmap shows which areas were most important in generating the given classification.

4.2 Comparison Between Models

Grad-CAM and Attention Heatmaps also allow us to make understandable comparisons between our models, as it provides some intuitive understanding of why certain models perform better than others, and provides insight into why some models are more prone to overfitting that others.

In particular, we consider the best performing CNN, DenseNet121, against ViT. Below, we see how the the two different models pay attention to very different areas of the same image. DenseNet (left) saw the most activation over the left and right peripheral areas of the retina, while ViT (right) saw most attention in small portions near the bottom and middle of the retina image.

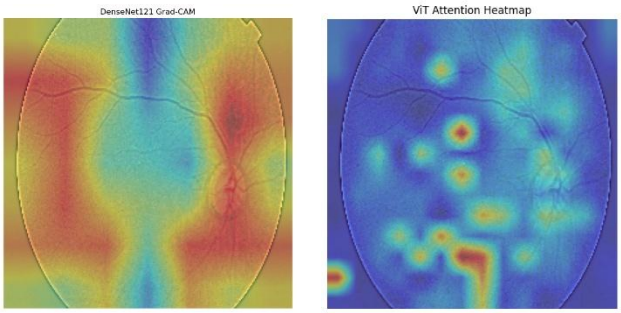

Figure 6. Visual Comparison of Model Focus Areas: DenseNet121 Grad-CAM vs ViT Attention Heatmap

These differences give us some intuition into why ViT might have performed better than DenseNet. ViT had more focused attention, as indicated by the smaller red areas, while more than half of the images saw high activation in DenseNet. Further, we obtain some insight into why ViT seemed to converge so quickly and tended to overfit the training set, since it tended to look at very small, specific areas. This might have allowed it to learn very specific details about the training set that could not be generalized to the validation or test set.

4.3 Large Language Model (LLM) Integration

While the visualization is helpful to direct attention to the most crucial parts of the retina, it may still not prove useful in cases where the user is not equipped to understand such heatmap visualizations, such as medical assistants or doctors with little experience with such deep learning models.

To bridge this gap, we built a LangChain pipeline that passes the ViT attention map overlaid on the original image, to a Large Language Model (LLM). The LLM is then prompted to explain, in natural language, how specific regions of the image contributed to the classification.

For this project, we used OpenAI's GPT-4o, which supports both visual and textual inputs and produces human-readable explanations. Our pipeline passes the overlaid Grad-CAM image along with the model's numeric prediction to GPT-4o, which is prompted to generate an explanation that relates the visual evidence to the model's classification in a way that is accessible to users. An example of the LLM's output for the image in 4.2 is produced below.

```
The model predicted 'Mild' with a confidence of 63.0%. The attention heatmap overlaid on the retina image highlights specific areas that contributed most to this prediction. Here's a breakdown of the key areas of interest:

1. **Inferior Region**: The most prominent red area is located in the inferior part of the retina. This suggests significant attention was given here, possibly indicating early signs of mild abnormalities or changes that warrant further investigation.
2. **Central Retina**: Several red spots are visible in the central region. These areas may show subtle changes or features that the model associates with a mild condition.
3. **Temporal Side**: There is noticeable attention on the temporal side of the retina. This area should be examined for any early signs of mild pathology.
4. **Superior Region**: Some attention is also given to the superior part of the retina, though less intense than the inferior region. It's worth checking for any mild changes here as well.

For follow-up, focus on these highlighted areas, especially the inferior region, to assess any early signs of retinal changes or abnormalities. Further clinical evaluation and imaging may be necessary to confirm the model's prediction and ensure accurate diagnosis and management.
```

4.4 Deployment

For deployment, a lightweight web application was developed using the Python Flask framework to provide an accessible and interactive interface for diabetic retinopathy detection. The application integrates the trained Vision Transformer (ViT) model, allowing users to upload retinal fundus images and receive real-time classification results. Flask was chosen for its simplicity, flexibility, and compatibility with Python-based machine learning workflows. The deployed system processes input images

through the ViT model on the backend and returns the predicted severity level to the user via a web interface.

As shown in Figure 7, the output page presents the original retinal image alongside an attention map generated by the Vision Transformer (ViT) model. The attention map highlights the regions the model focused on when making its classification, providing visual cues that can be useful for clinical interpretation. Below the images, the predicted class label (e.g., No DR) is displayed along with the model's confidence score. Additionally, a brief region-based explanation generated by LLM is provided to guide users or medical professionals in understanding why certain areas were considered important. This feature supports more informed decision-making and bridges the gap between automated predictions and clinical insight.

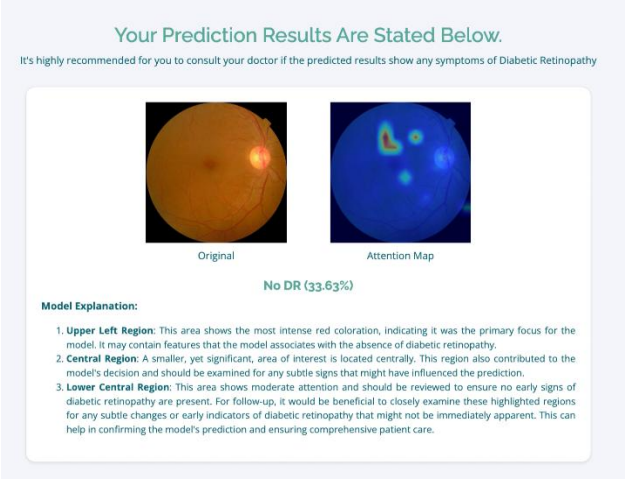

Figure 7. Sample Output Page of Web Application. It will show the original image, the attention map, and the explanation generated using OpenAI API (GPT-4o)

5. Limitations and Future Work

5.1 Limitations

Despite the promising outcomes, there are several limitations that we need to be aware of. One key limitation of this study is the dataset itself. As the dataset only has 3662 images, which is not particularly large, it might raise concerns about whether the model performance would be consistent when it is being applied to diverse clinical settings with varied patient populations. Additionally, although we applied SMOTE to address class imbalance issues, severe and proliferate classes are still underrepresented. This could potentially limit the performance of the model when it generalizes to real-world scenarios.

Another limitation relates to the model performance. Although we achieved a relatively good accuracy (82%) and a high ROC AUC (98.9%) in the final model, the F1 score (61%), especially the recall (59.5%), reveals some concerns. The low recall indicates that the model misses identifying about 40% of patients who actually have diabetic retinopathy. In real-world medical screening, recall is very important because missing a diagnosis could lead to patients not receiving timely treatment and hence a very high cost. Moreover, the Cohen's Kappa score of 72% shows that there is still a meaningful gap between perfect agreement and actual agreement. These metrics suggest that the model's performance in correctly identifying all disease cases and severity still has room for improvement.

Lastly, our final model ViT has a tendency to quickly overfit because it tends to memorize very small, overly specific details from the training data. Without robust regularization, this issue could reduce the effectiveness and ability to generalize on new unseen data in the real-world clinical scenarios.

5.2 Future Work

To overcome these limitations, future work should first focus on the dataset itself. It is important to collect more retinal fundus images, particularly those showing Mild, Moderate, Severe, and Proliferate DR, as it would enhance the model's performance and generalizability to real-world clinical uses. Additionally, exploring more advanced data augmentation techniques like Generative Adversarial Networks (GANs) could potentially balance the class distribution better, solving the imbalance issue and making the model more robust.

For the modeling part, future work could explore more hybrid approaches, such as combining convolutional neural networks and vision transformers. Since CNN performs quite well in detecting detailed, localized features from images, and ViT is good at capturing global contexts across the entire images, combining these two approaches might significantly boost the model classification performance. In addition, to reduce the overfitting in ViT, advanced regularization strategies or techniques specifically designed to deal with overfitting can be explored and applied.

Furthermore, adding professional clinician feedback or relevant knowledge databases into the web application can refine the interpretability and explainability of the results by aligning visual explanations more closely with real-world clinical reasoning, effectively improving the practical utility of the application.

Lastly, future research could incorporate patient data beyond just the retinal images. For example, including biomarkers such as glucose level could potentially lead to more comprehensive predictions. This approach would not only help validate the model accuracy but also improve preventive healthcare interventions.

6. Conclusion

This project successfully developed an automated diabetic retinopathy screening system, integrating powerful deep

learning architectures with traditional classifiers and explainability methods. Our experiments showed that while CNNs and hybrid pipelines delivered decent baseline results, the ViT outperformed other models in prediction accuracy, agreement with ground truth labels, as well as separability. Although the recall was moderate in multi-class classification, binary classification performance was excellent, meeting the critical requirements of real-world medical screening contexts where missing positive cases must be minimized.

By integrating visual interpretability techniques such as Attention Heatmap and natural language explanations generated by OpenAI's GPT-4o, we bridged the gap between complex black-box model outputs and practical clinical understanding, promoting more informed decision-making. Furthermore, we deployed a lightweight web application using the Flask framework and integrated the ViT model alongside attention heatmap and LLM-based explanations, allowing users to upload retinal fundus images and receive interpretable classification results. This application enhances practical accessibility for both end users and healthcare professionals.

However, important limitations still remain, particularly regarding dataset size and class imbalance issues. Future work should focus on expanding the dataset, particularly for underrepresented classes. Directions such as hybrid CNN and ViT models, clinician feedback incorporation, and multimodal patient data integration could also be explored for improving model robustness and real-world utility. Overall, DeepSight demonstrates a significant step toward creating a more accurate, interpretable, and accessible AI-driven diabetic retinopathy screening system.

Citation and References

Asia Pacific Tele-Ophthalmology Society. (2019).

"Blindness Detection". Retrieved from: https://www.kaggle.com/c/aptos2019-blindness-detection/overview.

Breiman, L. (2001). *Random Forests*. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: *A scalable tree boosting system*. arXiv. https://arxiv.org/abs/1603.02754

Cohen, J. (1960). *A Coefficient of Agreement for Nominal Scales.* Educational and Psychological Measurement, 20(1), 37-46. https://doi.org/10.1177/001316446002000104

Cortes, C., Vapnik, V. (1995). *Support-vector networks*. Machine Learning 20, 273–297. https://doi.org/10.1007/BF00994018

Dosovitskiy, et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv. https://arxiv.org/abs/2010.11929

Gill et al., (2004). *Accuracy of screening for diabetic retinopathy by family physicians.* Annals of Family Medicine, May 2004, 2(3), 218–220. https://doi.org/10.1370/afm.67

Huang et al. (2017). "*Densely Connected Convolutional Networks,*" 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). *Grey Wolf Optimizer*. Advances in Engineering Software, 69, 46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007

Pratt et al., (2016). *Convolutional neural networks for diabetic retinopathy*. In Procedia Computer Science, 90, 200-205.

Savoy, *Dx-DR for Diabetic Retinopathy Screening.* American Family Physician 101, no. 5 (2020): 307–308. https://www.aafp.org/pubs/afp/issues/2020/0301/p307.html

Tan, M., & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* In Proceedings of the 36th International Conference on Machine Learning on (Vol. 97, pp. 6105-6114). PMLR.

World Health Organization. (2023). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes