Early Detection of Diabetic Retinopathy with Interpretable Models

Abstract

We propose the development of interpretable computer vision models for the purpose of early detection of Diabetic Retinopathy (DR) based on retina images. Existing methods are presented as 'black box' models, thus hindering their adoption in clinical environments. This project hopes to overcome such limitations to contribute to the prevention of blindness among the global population.

1. Introduction

Diabetic Retinopathy (DR) is the leading cause of blindness globally among the working-age population. DR arises as a result of diabetes and not only threatens vision, but also signifies the onset of other dangerous systemic diseases such as kidney disorders and heart disease, thereby linking itself to a high-risk of mortality [1]. Therefore, the gravity of DR's impact on global health cannot be overstated. Despite the dangers attributed to it, the early detection and treatment of DR can prevent up to 95% of severe vision loss cases [2]. Thus, the development of effective and efficient methods for the early detection of DR is of utmost importance.

The advent of deep learning and computer vision has revolutionized the field of medical imaging, providing powerful tools for disease detection and diagnosis. In the context of DR, Convolutional Neural Networks (CNNs) and Transformer-based models have shown promising results in detecting subtle signs of DR in high-resolution retina images [3]. These models have an immense potential to augment or replace traditional, non-automated screening methods, thus enabling timely and accurate diagnosis.

However, despite their impressive performance, these models are often perceived as 'black boxes' due to their complex internal workings. This lack of interpretability is a significant barrier to their adoption in clinical practice [4]. It is evident that doctors and other healthcare professionals prefer to understand the 'how' behind a model's predictions before they can trust its decisions (the 'what' and the 'why'). This is certainly a valid concern as a misdiagnosis can have serious, even fatal, consequences.

Making an interpretable model can also provide valuable insights into the disease itself. In our case, an interpretable model could help identify biomarkers or patterns that are difficult to detect or were previously overlooked and would lead to a better understanding of the disease and help to inform the development of new treatments.

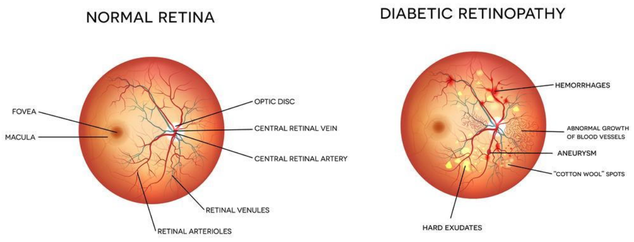

Figure 1. A normal retina versus one exhibiting signs of Diabetic Retinopathy which can be easily identified using Computer vision models but often lack interpretability.

With this project, we aim to address the issue of interpretability of deep learning models that are used for early DR detection by developing a model that is not only accurate but also understandable by leveraging state-of-the-art as well as well-established architectures such as CNNs or Transformers along with techniques for model interpretation. The goal is to provide clear and understandable reasoning that can be easily interpreted by healthcare professionals while maintaining performance, thus driving adoption in clinical settings. Techniques such as Layer-wise Relevance Propagation (LRP), Grad-CAM, and others have shown that it is possible to 'open the black box' and provide meaningful explanations for a model's predictions [5, 6]. In short, this project aims to contribute to early DR detection. By developing an interpretable model, we hope to make a difference by bridging the gap between high-performance computer vision models and their adoption in real-world clinical practice, ultimately contributing to the global fight against blindness.

2. Planned Methodology

Regular transformers are powerful tools for machine learning tasks, but understanding how they arrive at their decisions can be challenging. This is because they rely heavily on complex attention mechanisms, where relationships be-

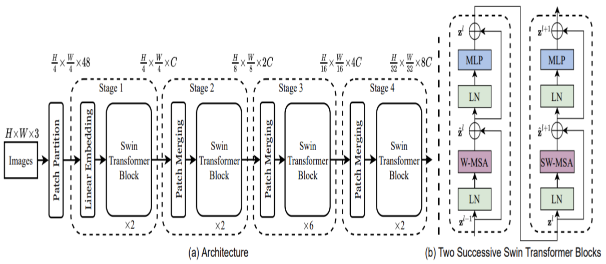

Figure 2. The Swin Transformer Architechture

tween different parts of the input data are formed in a way that is not easily untangled. Swin Transformers address this by introducing a window-based approach. By breaking the data down into smaller manageable chunks, Swin Transformers offer a clearer view into which parts of the input are most important for the final output, making the decision-making process more interpretable. We intend to utilize the dataset in [7]. The labels denote the presence of diabetic retinopathy in each image on a scale of 0 to 4 with a rank of 0 assigning a retinotopic image a label of no Diabetic Retinopathy, and a rank of 4 assigning an image a label of "Proliferative DR." The data set consists of over 35,125 training images and 53,576 test images.

2.1. Model Architecture

For the backbone of our architecture, we intend to use the Swin transformer model. The Swin Transformer introduces a hierarchical architecture that adapts the standard transformer model for vision tasks, bringing efficiency and flexibility to various scales of image representation. Its design, featuring shifted windows for self-attention, reduces computational complexity and enhances the model's ability to capture local and global dependencies in images. This structure not only improves performance on tasks like image classification, object detection, and semantic segmentation, but also aids in interpretability. By retaining spatial hierarchies and enabling attention across different regions, it offers insights into how visual features at multiple scales contribute to the model's decisions, making it easier to understand and trust the model's outputs.

For this work we intend to experiment across Swin-S, Swin-B and the Swin-B architectures.

2.2. Training

The authors of the Swin transforemer paper have presented several variations of the model in reference [8]. As a starting point we intend to start with models that are trained on large datasets (Imagenet 22k). The authors have confirmed that training with large datasets leads to better generalization and convergence.

Since the task is primarily an object classification, we intend to use a categorical cross-entropy loss function for training, along with an optimizer such as Adam.

2.3. Interpretation

We intend to evaluate our models across the following interpretability methods: Grad-CAM, LIME and SHAP.

3. Planned Experiments

The planned experiments for this project encompass two key areas: model evaluation metrics, and in-depth interpretability analysis. We will also conduct an ablation study to assess the impact of interpretability techniques on model performance.

3.1. Rating the presence of diabetic retinopathy

One of the main goals of the model would be to rate the presence of diabetic retinopathy in each image on a scale of No DR, Mild, Moderate, Severe and Proliferative DR. We will employ a set of standard metrics commonly used in image classification tasks. These metrics include accuracy, sensitivity, specificity, F1-score, and AUC-ROC to provide us with a quantitative assessment of the model's proficiency along with insights into its strengths and areas for improvement.

3.2. Interpretability Analysis

While achieving high accuracy is essential, understanding the model's rationale behind its predictions is equally important in a clinical setting. This is where interpretability techniques come into play. We aim to utilize a combination of techniques to gain better insights into the model's decision-making process.

We implement a Grad-CAM to generate heatmaps for each image classification. By comparing these heatmaps with known features of DR, we can assess whether the model focuses on areas that are in line with established medical knowledge. This analysis would help us build trust in the model's decision-making by demonstrating its focus on clinically relevant features. Furthermore, LIME (Local Interpretable Model-agnostic Explanations) is used to create simple interpretable models around specific predictions. This will provide insights into the features that influence the most while classifying individual images and in turn, offer us a localized understanding of the model's reasoning for each case. Finally, SHAP is employed to analyze feature importance across the entire dataset which assigns contribution scores to each feature to quantify the contribution to the final model prediction. This analysis provides a global view of which features play a more significant role in the model's overall DR classification decisions. By combining Grad-CAM, LIME, and SHAP analysis, we aim to achieve

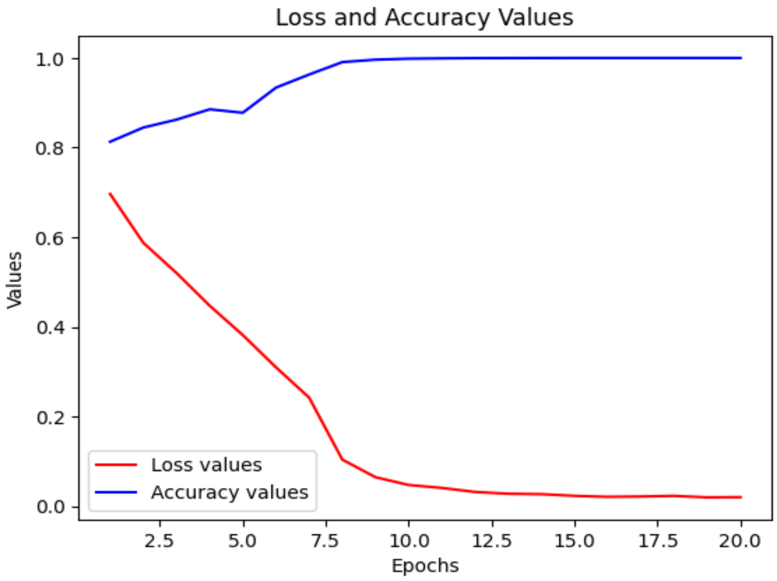

Figure 3. Loss and Accuracy Values per Epoch

a comprehensive understanding of how individual features and their interactions influence the model's DR classification decisions. This analysis of interpretability aims to enhance trust and support clinicians in making well-informed decisions, allowing them to integrate the insights provided by the model with their professional expertise.

4. Conducted Experiments

4.1. Experimental Successes

We successfully achieved part of our originally planned experiments by implementing and training a Swin Transformer on the Diabetic Retinopathy dataset as well as by introducing interpretability techniques to analyze our results. Training images were preprocessed and refitted in order to aid in model ingestion with the intention of increasing model performance while simultaneously decreasing computational and training time. Due to the complexity of the Swin Transformer coupled with the size of the training and test sets, 35,125 and 53,576 images for training and testing respectively, training was extremely resource and time intensive. Thus, we utilized NYU High Performance Computing (HPC) resources for data storage and model training purposes, which greatly expedited the process. Ultimately, we achieved the first of our planned experiments, the rating of the severity of DR, with the implementation of the Swin Transformer. After training the model to an acceptable grade of performance, we achieved the second of our planned experiments, introducing a level of interpretability analysis. After results from the Swin Transformer were achieved, we implemented Grad-CAM, LIME, and SHAP overlays to help achieve a level of interpretable results. Ultimately, they successfully granted us a look into the black box nature of detecting DR with machine learning models.

4.2. Experiment Inhibitors

With its successes, the experiments also experienced inhibitors to their progress. The largest was computing and training time. With such a large amount of data, it takes a proportional amount of time to process and train on the dataset. With a project timeline of just a few weeks and limited computing resources, processing and training time stands as the biggest blocker to the projects development. Secondly, there is the limited ability to quantify interpretability. Even for a human, it is difficult to empirically gauge how interpretable an entity is. With machine learning models it is no different. Therefore, while we try out best to visually and qualitatively measure the interpretability of our model, it is no easy task.

5. Results

5.1. Explanation of Results

The Swin Transformer was trained for 20 epochs, resulting in a minimum training loss of 0.0200 and a maximum test accuracy of 0.998. These performance results were above and beyond expectations. To have a model that can correctly classify four levels of severity for a commonly experienced and circumstantially altering condition speaks to the extreme potential for computer vision and machine learning in the world of healthcare. In reality, the model showed little performance increase after 10 epochs, but we contin-

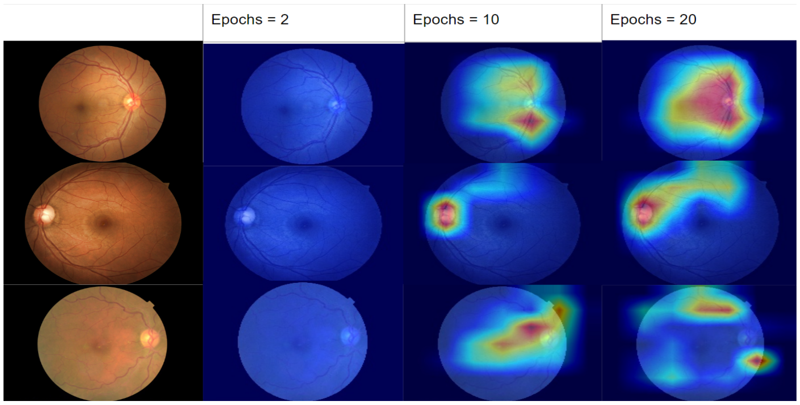

Figure 4. Grad-CAM Heatmaps

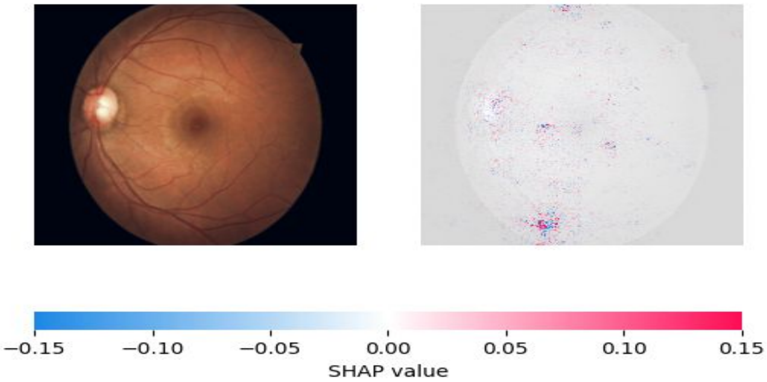

Figure 5. SHAP Overlay

ued training in order to evaluate the Grad-CAM mappings achieved with longer training time. Building on these results, we hoped to open the black box with interpretability metrics. As aforementioned, quantitative metrics of interpretability are difficult to achieve, thus the quality of our interpretability results must be observed qualitatively. From the Grad-Cam images captured in Figure 3, one can easily interpret the portions of the images that the model emphasizes its classification on. For example, as observable in Figure 4, the model learns to focus on the blood vessels present in the retinal image. This is logical, as telling signs of DR, as noted in Figure 1, are abnormal blood vessel growth as well as other occurrences such as hemorrhages and aneurysms that take place along vessels. The results of the SHAP overlay can be seen in the side-by-side comparison with the original image as shown in Figure 5. Red areas indicate areas that strongly support a diagnosis and blue areas indicate regions that negatively influence predictions, with the intensity of color correlating to the strength of influence. Although emphasis appears to be slightly put on the blood vessels running along the left side of the retina, SHAP results appear to be less than ideal in their ability to convey the interpretability of a model. Lime overlays resulted in more localized results. These results, along with the Grad-CAM and SHAP overlays, could potentially aid in a clinician's diagnosis of DR, as it could provide points of focus in a determination, but to the novice eye, it is difficult to derive meaninful results from the overlay alone.

5.2. Project Goals Achieved

The results achieved are reflective of the achievement of some of the project's initial goals. Firstly, the authors had the opportunity to gain experience with the computer vision project lifecycle. From gaining a contextual understanding, to dealing with data and modeling, and to result analysis and reporting, we as authors have been able to successfully work through a meaningful and insightful project. Next, we had the opportunity to explore and implement transformers in a computer vision context, granting us invaluable development and modeling experience. Finally, we gained experience with interpretability techniques that ultimately distinguish this project from the traditional train and test process associated with machine learning projects. In all, the experience and knowledge gained from this project has pre-

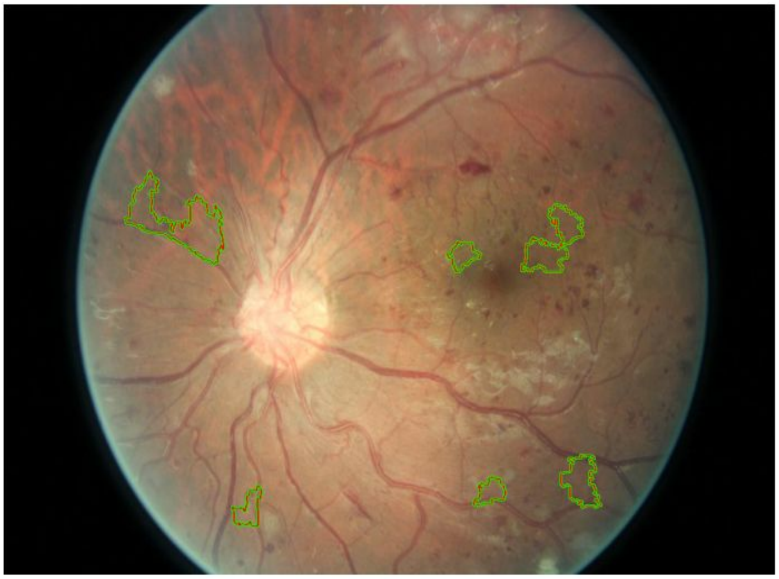

Figure 6. LIME Overlay

sented itself as invaluable.

6. Future Work

While we're proud of the work we've done, there is still much to be done. The final goal of computer vision in healthcare should be to innovate and improve upon existing processes for the sake of positive change. While this is difficult to achieve, all efforts to push the needle in the right direction accumulate eventually to a massive shift. At present, patients at risk of Diabetic Retinopathy can consult a machine learning model for accurate diagnoses. However, such conclusions yield little benefit when given without the reasonings behind them. Hence, a mixture of performance and interpretability is the key to the adoption of machine learning models in clinical settings. With a longer project timeline, one could experiment where the line between the two should be drawn by testing more models of varying performance and the usefulness of their accompanying Grad-CAM, LIME, and SHAP mappings and values. Ultimately, this could help clinicians determine what tools to adopt in aiding their analyses. Next, there is the advancement of interpretability techniques. While the techniques employed in this paper are state of the art and useful, there is much room to grow in the realm of interpretable models. For example, as aforementioned, it is difficult to quantify the interpretability of an image. Thus, advancements in this field would be fruitful for many aspects of computer vision.

References

[1] Bermejo, S., et al. The coexistence of diabetic retinopathy and diabetic nephropathy is associated with worse kidney outcomes, Clinical Kidney Journal, Volume 16, Issue 10, October 2023, Pages 1656–1663, https://doi.org/10.1093/ckj/sfad142

[2] National Eye Institute. (n.d.). People With Diabetes Can Prevent Vision Loss. Retrieved from National Eye Institute website.

[3] Bhimavarapu U, Battineni G. Deep Learning for the Detection and Classification of Diabetic Retinopathy with an Improved Activation Function. Healthcare (Basel). 2022 Dec 28;11(1):97. doi: 10.3390/healthcare11010097. PMID: 36611557; PMCID: PMC9819317.

[4] Nadeem MW, Goh HG, Hussain M, Liew SY, Andonovic I, Khan MA. Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions. Sensors (Basel). 2022 Sep 8;22(18):6780. doi: 10.3390/s22186780. PMID: 36146130; PMCID: PMC9505428.

[5] Landt-Hayen, M., Rath, W., Claus, M., Kröger, P. (2023). Fact or Artifact? Revise Layer-wise Relevance Propagation on various ANN Architectures. arXiv preprint arXiv:2302.12317.

[6] Selvaraju, R. et al. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv preprint arXiv:1610.02391.

[7] https://www.kaggle.com/c/diabetic-retinopathy-detection/data

[8] https://github.com/microsoft/Swin-Transformer

[9] Ze Liu, Yutong Lin et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv preprint arXiv:2103.14030