VISION-DR: Visual Insights and Saliency Integrated Overlay Neural Network for Diabetic Retinopathy

Harrison Sun
Northeastern University
Boston, Massachusetts, USA
sun.har@northeastern.edu

Scott Perryman
Northeastern University
Boston, Massachusetts, USA
perryman.s@northeastern.edu

Beyza Cavdar
Northeastern University
Boston, Massachusetts, USA
cavdar.b@northeastern.edu

Nihira Golasangi
Northeastern University
Boston, Massachusetts, USA
golasangi.n@northeastern.edu

***Abstract*—Diabetic retinopathy (DR) is a leading cause of vision impairment among diabetic patients, with early detection crucial for effective management and treatment [1]. This paper presents a computer vision approach to the diagnosis of diabetic retinopathy using a VGG-based convolutional neural network (CNN) [2], [3]. Our model is trained on a diverse dataset of fundus camera images to classify DR into various stages of severity [4]. To address the critical need for transparency and trust in medical diagnostic tools, our approach incorporates the generation of saliency maps, which highlight the specific areas within the images that influence the model's predictions [5], [6], [7]. This visualization aids in demystifying the model's decision-making process, providing healthcare professionals with valuable insights into the reasoning behind the diagnoses. The saliency overlay not only enhances the interpretability of the automated system but also serves to augment the diagnostic process by focusing attention on areas of potential concern. By presenting these findings alongside the model's classification, our tool is designed to support, rather than replace, the clinical judgment of physicians. This paper demonstrates the potential of computer vision techniques to not only automate the detection of diabetic retinopathy but also to contribute meaningful insights for improved patient care. Our results affirm the efficacy and reliability of our model, promoting its integration as a supportive tool in clinical settings.**

***Index Terms*—Computer Vision, Diabetes, Retinopathy**

I. Introduction

Diabetic retinopathy is an eye disease associated with both Type 1 and Type 2 diabetes that causes damage to the blood vessels in the retina [8]. Left untreated, diabetic retinopathy can cause permanent vision loss [9]. The Center For Disease Control (CDC) estimates that there are 9.6 million people in the United States living with Diabetic Retinopathy, of whom 1.84 million have vision-threatening diabetic retinopathy [10]. Since diabetic retinopathy is caused by swelling and leaking in the vasculature, which is associated with elevated blood glucose over long periods of time, it is imperative that the disease be caught earlier rather than later [11]. Notably, early-detection enables non-invasive treatments involving controlling blood glucose levels, blood pressure, and cholesterol levels [12]. However, late-stage diabetic retinopathy treatments are generally more invasive, including photocoagulation, steroid injections, and vitrectomy [12].

Diabetic retinopathy is typically divided into 5 stages: **0: No DR** - no hemorrhaging, microaneurysms, or abnormal vascularization occurs, **1: Mild Non-proliferative DR** - swelling in small blood vessels occurs, **2: Moderate Non-proliferative DR** - blood vessels become blocked, **3: Severe Non-proliferative DR** - ischemia occurs, blot hemorrhaging occurs, and abnormal vascularization may occur, and **4: Proliferative DR** - new, abnormal and fragile, blood vessels begin to grow in the eye [13].

Our goal is to introduce a computer vision model that aids physicians in accurately diagnosing the various stages of diabetic retinopathy. This model aims to serve as a critical tool, enhancing the precision of clinical assessments and facilitating a more streamlined diagnostic process. By integrating visual data analysis, the model provides valuable support in clinical decision-making, offering physicians insights that are essential for moni-

toring disease progression over time. Ultimately, our goal is to equip healthcare providers with technology that not only improves diagnostic accuracy but also contributes to the longitudinal study of diabetic retinopathy, paving the way for better patient outcomes.

II. Background

Diabetic retinopathy is a progressive disease that necessitates early and accurate detection to prevent severe vision loss. With the increasing prevalence of diabetes worldwide, efficient and scalable diagnostic technologies are crucial. Computer vision has emerged as a powerful tool in this context, enabling the automated detection and classification of diabetic retinopathy.

Computer vision has become increasingly prevalent in medical diagnostics [14]. Specifically for diabetic retinopathy, it typically involves the use of classifiers that categorize a single fundus image into one of five stages of the disease [15]. The VGG model, known for its robust feature extraction capabilities, is particularly favored in image recognition tasks [15], [16], [17], [18], [19], [20], [21], [22]. It effectively identifies key indicators of diabetic retinopathy such as microaneurysms, exudates, hemorrhages, and abnormal vascularization [23]. Existing applications of this model have achieved accuracies exceeding 85% in diagnosing the condition [8].

Despite these advancements, the opaque "black box" nature of deep learning models poses a significant barrier to their adoption in clinical settings [24]. Traditionally, the decision-making processes of these models are not transparent, leaving physicians without a clear understanding of why certain decisions were made [24]. To overcome this challenge, researchers have developed explainability techniques like Grad-CAM [25]. These techniques provide visual explanations by highlighting influential regions in the images at one of the final layers of the network. This not only helps bridge the gap between model decision-making and user interpretability but also supports physicians in integrating their clinical expertise with model suggestions for better-informed decision-making.

III. Problem Formulation

The goal of our project is to advance the diagnosis of diabetic retinopathy, a leading cause of blindness among adults worldwide, by implementing a dual-purpose computer vision algorithm. This algorithm is designed not only to improve diagnostic accuracy but also to enhance transparency in the diagnostic process. Current methods for detecting and assessing diabetic retinopathy depend heavily on the manual examination of retinal images by skilled clinicians. These methods are often limited by the availability of experts and can suffer from subjective variability in diagnosis. As the prevalence of diabetes increases globally, there is an urgent need for a more scalable and consistent approach to diagnosing this vision-threatening condition.

Our objectives are twofold. Firstly, we aim to accurately classify the severity of diabetic retinopathy across its spectrum, from mild to severe stages. The ability to classify with high precision is crucial as it directly informs treatment options and management strategies, potentially leading to better patient outcomes. Secondly, we seek to increase the accountability and trustworthiness of AI in medical diagnostics by generating saliency maps. These maps are intended to visually represent the critical features within the retinal images that influence the AI's decision-making process. By doing so, we aim to provide clinicians and patients with clear, visual explanations of the AI's diagnoses, promoting transparency and understanding.

In summary, our work seeks to address the significant challenge of scaling diabetic retinopathy diagnostics while maintaining the quality and reliability expected in healthcare. By achieving these goals, we anticipate that the confidence in AI-assisted diagnostics will be significantly bolstered, leading to wider acceptance and use of these technologies in clinical settings. This approach not only promises to enhance the capacity for early detection and treatment of diabetic retinopathy but also sets a precedent for the application of AI in other areas of medical imaging and diagnosis.

IV. Methodology

The dataset we use is the Kaggle Diabetic Retinopathy Detection Dataset [4]. The dataset consists of over 35,000 high resolution retina images captured with a fundus camera [4], [15]. All of the images are labeled by clinicians into 5 numbered classes corresponding to severity. The data is structured into training and testing files as well as a file delineating training and testing labels.

We have constructed a classifier employing the VGG-16 architecture, which is a convolutional neural network with 16 layers [2]. As outlined in Figure 1, our approach starts with loading the VGG-16 model pre-equipped with trained weights. Subsequently, we enable training on all layers by unfreezing their weights. The VGG architecture utilizes 3x3 convolutional filters and max-pooling to

```
BEGIN
  // Load Pretrained Weights
  vgg_model ← load_model("VGG16", pretrained=True)

  // Unfreeze all layers for training
  FOR layer IN vgg_model.parameters()
    layer.requires_grad ← True
  END FOR

  // Modify classifier structure
  vgg_model.classifier[6] ← Sequential(
    Linear(4096, 1024),
    ReLU(),
    Dropout(0.5),
    Linear(1024, 1),
    Sigmoid()
  )

  // Set device
  device ← if GPU_available() then "cuda:0" else "cpu"
  vgg_model.to(device)

  // Loss function
  criterion ← Cross_Entropy_Loss()

  // Use Adam Optimizer
  optimizer ← Adam(vgg_model.classifier.parameters(), lr=0.001)
END
```

Fig. 1: Pseudocode for adapting a pre-trained VGG-16 model for Diabetic Retinopathy classification.

minimize the model's complexity and the number of parameters needed. The classifier's configuration is altered to include a linear layer, a ReLU activation function, a dropout rate of 0.5 to prevent overfitting, another linear layer, and a sigmoid activation to produce a probability output. The model is trained using cross-entropy loss and optimized with the Adam optimizer with a learning rate of 0.001 [26], [27].

Following the training and optimization of the VGG-16 model for diabetic retinopathy detection, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) to enhance the interpretability of the model's predictions [25]. Grad-CAM is a visualization technique that highlights the regions in the input image that are important for predictions from convolutional neural networks. This method uses the gradients of the output of the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

To implement Grad-CAM, we first import the best-performing VGG-16 model trained on the diabetic retinopathy dataset. We then focus on the activations and gradients of the final convolutional layer, as this layer captures the most complex features in the image that are crucial for making the final decision. By computing the gradient of the output category with respect to the feature maps of the final convolutional layer, and then pooling these gradients over the spatial dimensions, we create a weighted map of the important features. This weighted feature map is then used to create a heatmap by performing a weighted combination of the feature maps, followed by a rectified linear transformation. Finally, this heatmap is superimposed on the original image to visually represent the areas most significant for the model's classification decision, providing insights into what the model is considering important in diagnosing

diabetic retinopathy. This method not only aids in verifying the model's focus areas but also enhances trust and understanding in its diagnostic decisions, making it a powerful tool for medical imaging analysis.

V. Results & Discussion

The goal of this project was to accurately classify the severity of diabetic retinopathy in fundus camera images and to generate a saliency map highlighting the image regions influencing the classification decision. We evaluated model performance using cross-entropy loss and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Due to significant class imbalance, we chose AUC-ROC over accuracy as a performance metric. Since AUC-ROC remains unaffected by skewed class distributions, it provides a more reliable indicator of model efficacy than accuracy. Additionally, the AUC-ROC curve is influenced by the balance between sensitivity and specificity, ensuring that the abundance of Class 0 cases does not skew the performance metric.

Figure 2 illustrates the training and validation loss trends. Notably, at around 10 epochs, both training and validation losses converge to approximately 0.15. While the training loss continues to decrease, the validation loss begins to fluctuate after a few epochs. To prevent overfitting and manage computational resources efficiently, we limited our training to 10 epochs.

Figure 3 displays the AUC-ROC for both training and validation sets, which begin to plateau at about the fifth epoch, peaking near 0.9. This indicates a 90% probability that the model accurately distinguishes between the two classes, confirming the model's robust performance in assessing the severity of diabetic retinopathy and achieving our primary objective.

Our secondary objective involves visually identifying the critical areas influencing the model's decisions. Figure 4 presents a saliency map for an image classified as Class 0 (No Diabetic Retinopathy), where the focus is primarily on three regions. The top left and bottom left corners are examined for notches indicating mirrored images. The right-hand focus assesses the blood vessels for abnormal vascularization and hemorrhaging. Conversely, Figure 5 shows a saliency map for Class 4 (Proliferative Diabetic Retinopathy), with particular attention to neovascularization driven by angiogenesis and hemorrhaging. Notably, the model disregards a large dark spot in the upper left quadrant, which we identified as a choroidal nevus — a benign, commonly occurring pigmented lesion unrelated to diabetic retinopathy [28]. This discernment demonstrates the model's capability to distinguish between relevant and unrelated features, which we consider satisfactory for our second objective. However, a comprehensive assessment of the model's performance across various features would require extensive clinical annotation.

VI. Future Work

A key area of interest is the development of an image segmentation system that can delineate detailed structures within fundus images. Current limitations stem from the precision required in labeling segmentation masks, a task that demands specialized medical expertise. Collaborating with clinicians to label these fine masks accurately would enable us to enhance our model's understanding of the spatial relationships and exact boundaries of lesions associated with diabetic retinopathy.

Moreover, integrating segmentation with our existing classification model could significantly refine its performance. Segmentation masks could serve not only to validate the areas of interest identified by the saliency maps but also to quantify aspects such as the area affected by different classes of retinopathy. This quantitative data could be invaluable in training more sophisticated machine learning models that consider the extent and specific locations of retinopathy signs as factors in their predictions.

Another promising direction involves tracking the progression of diabetic retinopathy over time. By developing algorithms that analyze sequential fundus images of the same patients, we can gain insights into the evolution of the disease. This longitudinal analysis would allow for more personalized and timely treatment decisions, potentially slowing or even reversing the progression of retinopathy.

Finally, integrating findings from existing literature into our models through the use of regularization techniques based on prior knowledge stands to significantly improve model performance and reliability. By incorporating constraints derived from the extensive literature on diabetic retinopathy and related pathologies, we can guide the learning process of our models to adhere more closely to clinically observed patterns. This regularization could help mitigate the effects of overfitting and improve generalization to new, unseen data by aligning our model's inferences with established medical knowledge.

VII. Conclusion

This project has demonstrated the potential of computer vision models in enhancing the diagnosis of di-

Fig. 2: Loss per Epoch

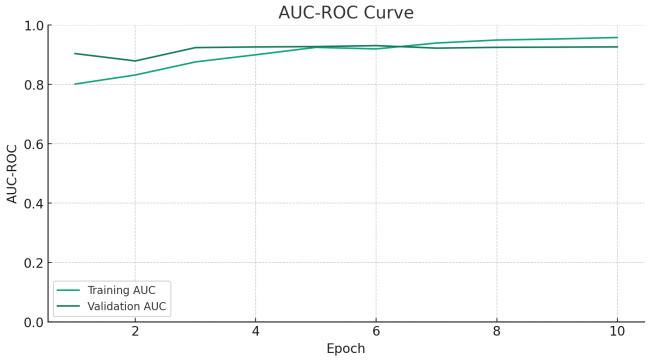

Fig. 3: AUC-ROC Score per Epoch

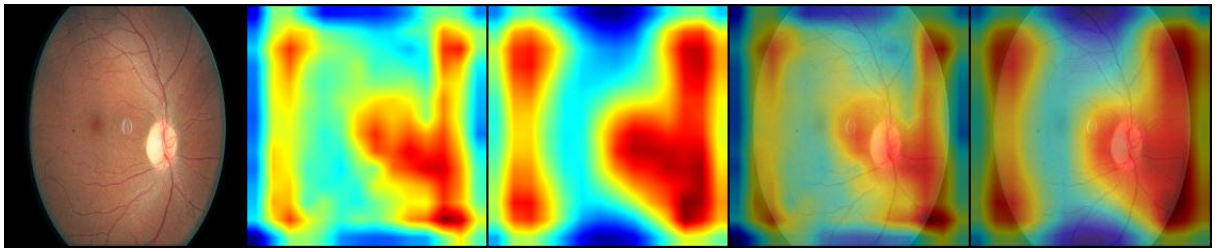

Fig. 4: Saliency Map for an Image with No Diabetic Retinopathy

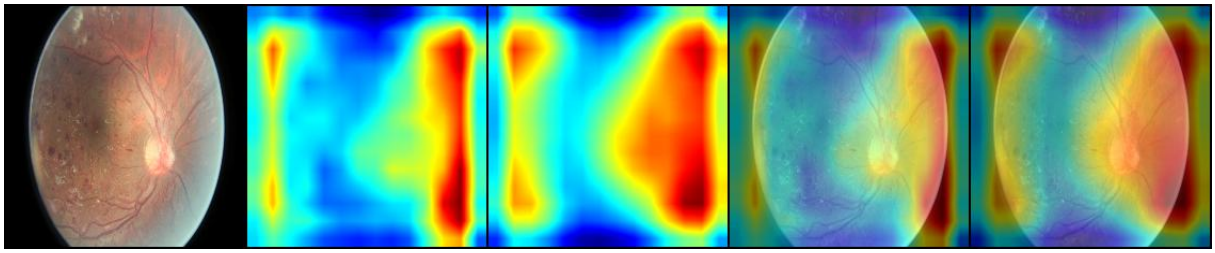

Fig. 5: Saliency Map for an Image with Proliferative Diabetic Retinopathy

abetic retinopathy through fundus camera images. By employing the AUC-ROC metric, which is particularly effective in conditions of class imbalance, the model achieved a robust ability to distinguish between different severity levels of retinopathy. Furthermore, the generation of saliency maps provided visual confirmation of the regions in the images that influenced the model's decisions, highlighting the model's capacity to identify clinically relevant features.

The model shows strong predictive performance and the insights gained from the saliency maps pave the way for more intricate explorations into image-based diagnostic processes. Future efforts could focus on advancing towards image segmentation to gain more detailed diagnostic information and on leveraging sequential image data to track the progression of retinopathy over time. Additionally, incorporating medical expertise into the labeling process and integrating established clinical knowledge into machine learning models through regularization would enhance the accuracy and reliability of diagnostic tools.

As we continue to bridge the gap between technical capabilities and clinical needs, the prospect of AI-driven tools becoming a staple in diagnostics is increasingly feasible. These tools promise not only to enhance diagnostic precision but also to enable earlier interventions, potentially altering the course of diabetic retinopathy for countless patients.

References

[1] A. N. Kollias and M. W. Ulbig, "Diabetic Retinopathy," *Deutsches Arzteblatt International*, vol. 107, no. 5, pp. 75–84, Feb. 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2828250/

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, arXiv:1409.1556 [cs]. [Online]. Available: http://arxiv.org/abs/1409.1556

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998. [Online]. Available: http://ieeexplore.ieee.org/document/726791/

[4] J. W. C. Emma Dugas, Jared, "Diabetic Retinopathy Detection," 2015. [Online]. Available: https://kaggle.com/competitions/diabetic-retinopathy-detection

[5] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, Nov. 2020. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6

[6] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," Apr. 2014, arXiv:1312.6034 [cs]. [Online]. Available: http://arxiv.org/abs/1312.6034

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization."

[8] "Diabetic Retinopathy Detection." [Online]. Available: https://kaggle.com/competitions/diabetic-retinopathy-detection

[9] R. Lee, T. Y. Wong, and C. Sabanayagam, "Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss," *Eye and Vision*, vol. 2, no. 1, p. 17, Sep. 2015. [Online]. Available: https://doi.org/10.1186/s40662-015-0026-2

[10] "Diabetic Retinopathy Estimates | Vision and Eye Health Surveillance System | CDC," Jul. 2023. [Online]. Available: https://www.cdc.gov/visionhealth/vehss/estimates/dr-prevalence.html

[11] H.-P. Hammes, Y. Feng, F. Pfister, and M. Brownlee, "Diabetic Retinopathy: Targeting Vasoregression," *Diabetes*, vol. 60, no. 1, pp. 9–16, Jan. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3012202/

[12] A. W. Stitt, T. M. Curtis, M. Chen, R. J. Medina, G. J. McKay, A. Jenkins, T. A. Gardiner, T. J. Lyons, H.-P. Hammes, R. Simó, and N. Lois, "The progress in understanding and treatment of diabetic retinopathy," *Progress in Retinal and Eye Research*, vol. 51, pp. 156–186, Mar. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S135094621500066X

[13] W. L. Yun, U. Rajendra Acharya, Y. V. Venkatesh, C. Chee, L. C. Min, and E. Y. K. Ng, "Identification of different stages of diabetic retinopathy using retinal optical images," *Information Sciences*, vol. 178, no. 1, pp. 106–121, Jan. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025507003635

[14] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *npj Digital Medicine*, vol. 4, no. 1, pp. 1–9, Jan. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41746-020-00376-2

[15] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, "Fundus Image Classification Using VGG-19 Architecture with PCA and SVD," *Symmetry*, vol. 11, no. 1, p. 1, Jan. 2019, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2073-8994/11/1/1

[16] T. Kaur and T. K. Gandhi, "Automated Brain Image Classification Based on VGG-16 and Transfer Learning," in *2019 International Conference on Information Technology (ICIT)*, Dec. 2019, pp. 94–98. [Online]. Available: https://ieeexplore.ieee.org/document/9031952

[17] I. Ha, H. Kim, S. Park, and H. Kim, "Image retrieval using BIM and features from pretrained VGG network for indoor localization," *Building and Environment*, vol. 140, pp. 23–31, Aug. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360132318302865

[18] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Feb. 2019, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8869364

[19] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep Feature Fusion for VHR Remote Sensing Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017, conference Name: IEEE Transactions on Geoscience and Remote Sensing. [Online]. Available: https://ieeexplore.ieee.org/document/7934005

[20] X. Lu, X. Duan, X. Mao, Y. Li, and X. Zhang, "Feature Extraction and Fusion Using Deep Convolutional Neural Networks for Face Detection," *Mathematical Problems in Engineering*, vol. 2017, p. e1376726, Jan. 2017, publisher: Hindawi. [Online]. Available: https://www.hindawi.com/journals/mpe/2017/1376726/

[21] M. S. Majib, M. M. Rahman, T. M. S. Sazzad, N. I. Khan, and S. K. Dey, "VGG-SCNet: A VGG Net-Based Deep Learning Framework for Brain Tumor Detection on MRI Images," *IEEE Access*, vol. 9, pp. 116 942–116 952, 2021, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9515947

[22] S. Tammina, *Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images*, Oct. 2019, vol. 9, journal Abbreviation: International Journal of Scientific and Research Publications (IJSRP) Publication Title: International Journal of Scientific and Research Publications (IJSRP).

[23] T. S. Hwang, Y. Jia, S. S. Gao, S. T. Bailey, A. K. Lauer, C. J. Flaxel, D. J. Wilson, and D. Huang, "Optical Coherence Tomography Angiography Features of Diabetic Retinopathy," *Retina (Philadelphia, Pa.)*, vol. 35, no. 11, pp. 2371–2376, Nov. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4623938/

[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3236009

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," 2017, pp. 618–626. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

[26] Z. Zhang and M. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html

[27] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic

Optimization," Jan. 2017, arXiv:1412.6980 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6980

[28] "Distinguishing a Choroidal Nevus From a Choroidal Melanoma," Feb. 2012. [Online]. Available: https://www.aao.org/eyenet/article/distinguishing-choroidal-nevus-from-choroidal-mela

APPENDIX

Source Files