Iraqi Journal of Information and Communications Technology(IJICT)
Vol. 7, Issue 2, August 2024
www.ijict.edu.iq
ISSN:2222-758X
e-ISSN: 2789-7362

ENHANCING TRANSPARENCY IN HEALTHCARE MACHINE LEARNING MODELS USING SHAP AND DEEPLIFT A METHODOLOGICAL APPROACH

Seyedamir Shobeiri [1]
[1] Department of Computer Science, University of Luxembourg. Luxembourg
seyedamir.shobeiri.001@student.uni.lu[1]
Corresponding Author: **Seyedamir Shobeiri**
Received:30/05/2024 ; Revised:27/07/2024; Accepted:14/08/2024
DOI:10.31987/ijict.7.2.285

***Abstract*- This paper intends to provide a better understanding of how these models produce predictions, particularly in complex medical diagnoses, and at the same time bridging the gap between technical model outputs and clinical applications. This study addresses the critical problem of transparency of machine learning (ML) models in health care where interpretability is an essential aspect for ethical decision making and trust building.The goal of this paper is to present a clearer understanding of how these models generate predictions, especially in important fields like complex medical diagnoses, thereby bridging the gap between technical model outputs and clinical applications. The crucial issue of transparency in machine learning (ML) models within healthcare is addressed in this study where interpretability plays a vital role in ethical decision-making and fostering trust. The focus of the research is enhancing model transparency by using SHapley Additive exPlanations (SHAP) and Deep Learning Important FeaTures (DeepLIFT), two crucial methods that are designed to elucidate the decision-making processes of ML models.This mode of approach helps to have more distributed comprehension of the decision pathways by models thus aiding in knowing how each feature contributed to the last prediction. It is this method that has been employed to showcase the efficiency of predicting melanoma and also diabetic retinopathy which are two vital medical diagnostic areas. In healthcare, SHAP along with DeepLIFT has improved the models' explainability and trustworthiness significantly and hence making them easy for those in the field. The advanced interpretability methods presented in this document enhances ML model transparency especially when dealing with health issues. As a result, interpretability becomes an even bigger issue and they are supposed to be able to use these tools for reliable and open decisions when it comes to medical specialists.**

keywords: **Healthcare AI, SHAP, Deep LIFT, Model Transparency.**

I. INTRODUCTION

Artificial intelligence and machine learning have become integrated in healthcare systems and the improvements since the implementation of the technologies has been impressive. AI systems in different fields especially using the ML techniques have been responsive in analyzing large datasets and data in the medical field and has helped in analyzing various data that a doctor might not see normally and help in early detection of diseases like cancer and Diabetic retinopathy. However, as these models become more sophisticated and complex, a critical issue arises: the fact that many ML models are initially "black boxes" and that the decision-making internal to the algorithm is not transparent to an external viewer. One major difficulty implemented when models do not provide clear decision, logic, or understanding is in the field of medical healthcare where exact explanation about diagnosis or prescription is required by both doctors and patients. Some of situational and professional implications are as follows: Medical professionals are not only expected to make correct diagnostic decisions but also to justify them to the patient. This may be an issue in practice due to the 'black box' nature of many ML models; they can contribute to this trust but at the same time decision making is not clear to healthcare providers or explainable to others. This issue is most crucial in areas considered as having high risk whereby any single mistake could lead to

ISSN:2222-758X
e-ISSN: 2789-7362

unhealthiness or even harm of patients. To tackle these challenges, there has been a rising concern in the development of approaches that improve the interpretation of and within ML models. Two of these methods that have received a lot of attention in recent times are the SHAP (SHapley Additive exPlanations) and the DeepLIFT (Deep Learning Important FeaTures). SHAP, developed by the authors based on the principles of cooperative game theory, enables to obtain a unique measure of feature significance, and at the same time, it guarantees that each feature is rewarded equally for its contribution to the predictions of the developed model. The former is especially useful in health care industries since it provides a means of translating model predictions into easy to comprehend parts, thereby enabling doctors to understand which part of a decision was influenced by which factors. While Layer Wise Relevance Identification and Explanation (LRx) helps in the decomposition of selection layers for ReLU-based networks, DeepLIFT helps in propagating the relevant of each input feature through the layers. In DeepLIFT, contribution scores are assigned to each neuron based on the difference between the neuron's output and that of reference input; thus, this highlights the sensitivity of the model to changes in the input features. This method can be said to accompany SHAP, while giving the layer wise interpretation that can be very useful in deep learning where each layer has an interaction with the other that complicates the interpretation process.This proposal of integrating SHAP and DeepLIFT provides a comprehensive solution to improve the interpretability of the healthcare ML models. The global Sensitivity-Heatmap Aggregation-Projection (SHAP) explanation can be considered to give global information on the feature importance across all instances and DeepLIFT for more localized instance-specific information. This dual approach also assists in the understanding of the model's behavior but more importantly it aids in the reliability of the interpretations hence enhancing clinical decision-making.Further, the application of these interpretability techniques can be in concordance with more general ethical and legal requirements in health care. Attribution and auditability become critical consideration especially as AI systems are being integrated into clinical practice. Governance authorities and code of ethics are now moving towards focusing on explainable AI to prevent violation of the law while promoting the acceptance of deep learning algorithms by health care professionals and users. In this respect, the practical usage of these methods can be seen in applications like melanoma and diabetic retinopathy, cancers where exact and explicable models are crucial to bringing timely diagnosis and planning of the further treatment course. However, for melanoma and many other similar diseases, SHAP and DeepLIFT are able to clarify which features of skin lesions are most useful in deciding on malignancy, thereby enhancing the efficacy of clinicians' diagnoses. Likewise, the same methods in diabetic retinopathy can point out which of the retinal characteristics are most influential that eventually help in timely medical management. Thus, improving the transparency of the ML models in the healthcare context, which can be done, for example, using the SHAP or DeepLIFT techniques, is not only a technical problem but also a part of using ethical AI. These approaches enable the creation of easily explained and understandable reasons for model decisions, which makes them more relevant and closer to clinicians, thus, improves patient care and increases acceptance of AI tools. In this method of continuing healthcare based on advanced technologies like the AI application, it will be paramount that the tools retain efficiency together with the aspect of unveiling the functions that may compel their usage in the future.

ISSN:2222-758X
e-ISSN: 2789-7362

II. RELATED WORK

A. *Perturbation-Based Methods*

Few new methodologies like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been under discussion to explain the role of the features for decision making. SHAP, for instance, is praised for the feat of presenting several feature attribution methods under one clear and coherent framework that enables the interpreter to understand the output of complicated models in several data forms, including image and text [1]. However, these methods encounter problems like high computational complexity and a requirement of many data perturbations, which hinders their real-time applicability [2]. Fig.1 shows Perturbation-based approaches and gradient-based approaches are failed to model saturation.

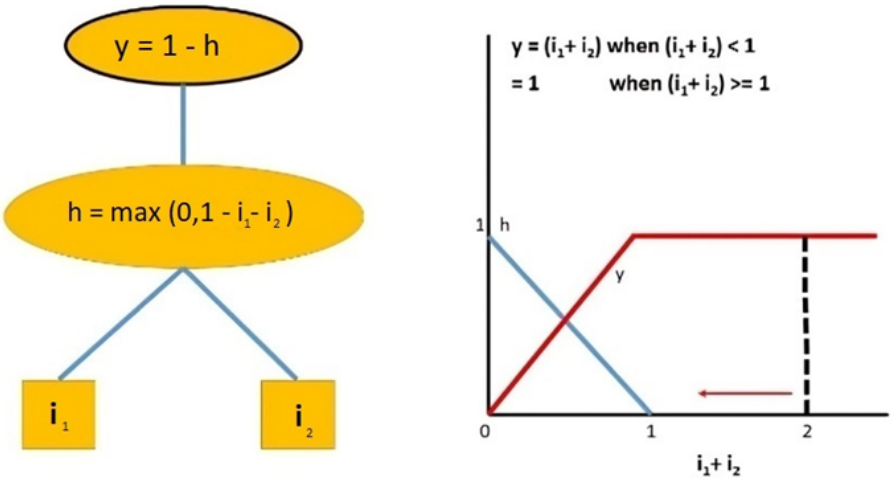

Figure 1: Perturbation-based approaches and gradient-based approaches are failed to model saturation.

B. *Backpropagation-Based Approaches*

Comprehending the actions of deep learning models requires methods such as DeepLIFT and Layer-wise Relevance Propagation (LRP). For example, DeepLIFT has been used to break down the prediction made by neural networks into contributions from individual input attributes. Various domains including healthcare have utilized this technique to offer more clarity on how certain inputs affect outcomes [3]. However, it also suffers from difficulties in the interpretation of results and oversimplified feature interaction potential that should be solved [3].

C. *Integrated Gradient-Based Methods*

The Integrated Gradients apply a path-dependency model to feature attribution so as to resolve issues that come with simpler gradient dependent models. In this way, it offers more thorough explanations as attributions take into account all possible input values [4]. Nevertheless, the downside of this approach is that it can be very slow in computation and is also greatly influenced by the chosen baseline hence affecting its results and interpretations [4]. A summary of recent work is shown in Table I.

ISSN:2222-758X
e-ISSN: 2789-7362

TABLE I
SUMMARY OF RECENT WORKS

Method	Dataset(s)	Strengths	Weaknesses
SHAP	Healthcare datasets, UCI	Unified approach, model-agnostic, handles various data types	High computational cost, sensitive to feature interactions
DeepLIFT	MNIST, ImageNet	Clear attributions, layer-wise relevance	Complexity in interpretation potential oversimplification
Integrated Gradients	Structured, unstructured data	Comprehensive path-based attributions	Intensive computation, baseline dependency
LRP	Medical imaging	DETAILED FEATURE RELEVANCE MAPPING	Interpretation complexity, handling of non-linearities

Despite major strides being made in this area, numerous challenges still remain in the realm of explainable AI. One of the most critical issues is the associated computational cost with these approaches which restricts their appropriateness for real-time contexts, another point of concern regards the potential lack of stability and consistency of explanations especially where health care where dependable and interpretable outputs are necessary for making decisions. Also, current techniques often encounter difficulties when capturing complex interfeature interactions typical to medical data. An integrated framework based on SHAP and DeepLIFT will try to solve both efficiency and accuracy challenges that come with interpretability models especially in health care domain. In this way, it aims to offer deeper, more accurate insights that can aid clinical decision-making process thereby increasing confidence in AI systems.

III. METHOD

DeepLIFT methodology together with SHAP variable values combine to introduce a robust approach of decision model influence by providing individual feature scores based on their contribution to the output model. Such combination occurs as the natural consequence of the merits that are exhibited by both the methods; hence a more robust and comprehensible information is provided. Since the idea of integration is quite complex, we are going to simplify it by explaining what integration formula is and how it can be used to solve problems.

A. *DeepLIFT Methodology*

DeepLIFT, an algorithm proposed by Shrikumar et al., is a neural network that is designed to explain its results by contrasting each neuron's activation against reference activation and prices it as differences from features of the input. The principal concept is the amount of difference to be taken away from a predicted value of output when compared to the reference output that has been arrived at from a base input on the contrary. This reference is a vector of zeros or the mean value of inputs in a training set, depending on the dataset and the purpose it is built for.[5]

ISSN:2222-758X
e-ISSN: 2789-7362

B. *SHAP Values*

SHAP (SHapley Additive exPlanations) model, designed by Lundberg and Liu, uses the individual contributions of features to predict an outcome as shown in Fig.2. Particular values result from game theory, especially the Shapley values which take a surplus of the coalition (a combination of players) as a starting point and divide it evenly among them. On the other hand the combining of the features (variable selection, feature engineering, and Feature selection), fairness is still a big issue among the contributors (features). Thus, once all the causal components represented by each of the features in the model are properly and precisely attributed and summed up, SHAP values ensure that predictions are consistent and locally accurate, which they are fairly attributed and decomposed into. [6] [7]

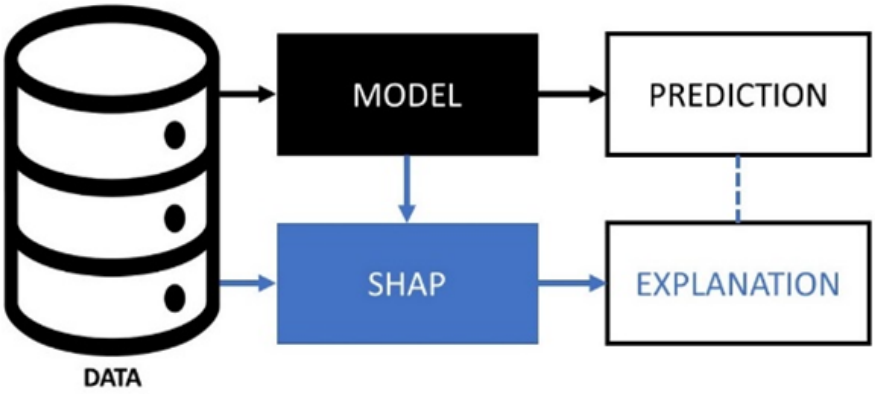

Figure 2: Illustrating the Integration of SHAP for Model Interpretation.

C. *Mathematical Integration of DeepLIFT and SHAP*

To merge DeepLIFT with SHAP, we consider the following formalization: Define $f = \mathbb{R}_n \rightarrow \mathbb{R}$ as the neural network function, where x the input vector. Designate x' as a new input signal. DeepLIFT computes contributions $C_i(x, x')$ for each feature i, defined as:

$$C_i(x, x') = (x_i - x'_i) \times \Delta y_i \tag{1}$$

which is notated as Δy_i, where i is the activity change. To calculate the SHAP values using DeepLIFT contributions, we use the expectation over a distribution of reference inputs sampled from a background dataset X':

$$\phi(i)(x) = E_{x'} \sim X'[C_i(x, x')] \tag{2}$$

D. *Implications and Applications*

The individuals using this holistic model are, therefore, able to explore the cognitive processes inherent to complex simulations, applying this knowledge to high-stakes industries like healthcare, finance, and autonomous driving, for instance.

ISSN:2222-758X
e-ISSN: 2789-7362

Almost an accurate picture is being painted by developers and stakeholders about the reasons and ways through which predictions are made, so that the models can be true to their intended behavior, biases are taken seriously, and the models become more reliable and fairer. The alliance of DeepLIFT and SHAP is a framed and computationally-sound mechanism for crediting the influence of each driver of the deep learning models. This framework provides clarity of perception and through that, it limits the nature of black-box in AI applications, which otherwise could be considered suspicious and inaccessible. With the AI applications, transparency is established and the trust is built.

IV. MATERIALS

A. EVALUATING RESNET-18'S EFFICACY IN MELANOMA CLASSIFICATION FROM DERMATOLOGICAL IMAGES

One of the less common and more dangerous types of skin cancers that start to grow when melanocytes is out of control is Melanoma. Squamous cells, basal cells, and melanocytes are types of cells in the epidermis layer, the top layer of skin where most skin cancers begin [8]. Fig.3 shows Anatomy of Human Skin.

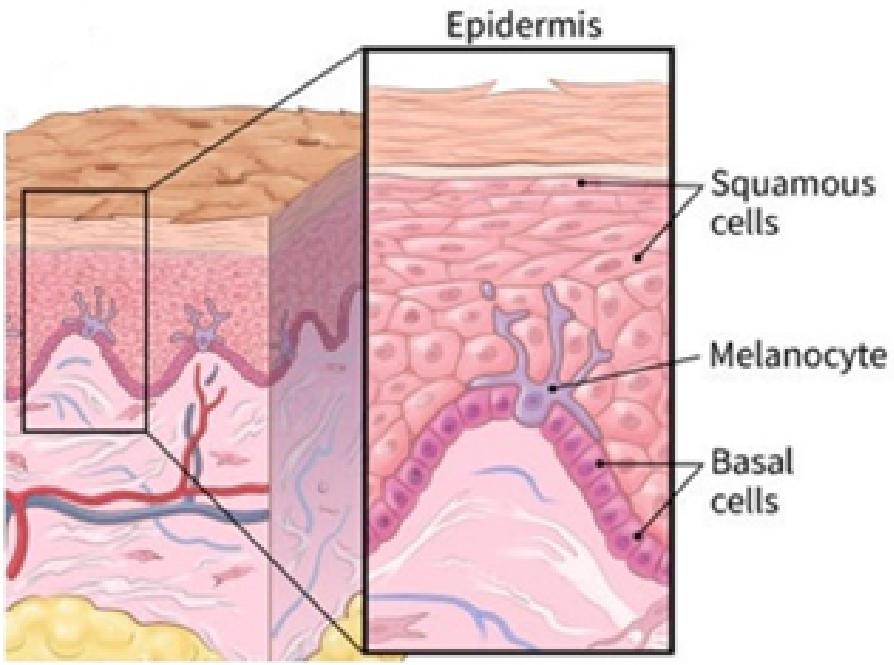

Figure 3: Anatomy of Human Skin.

This type of cancer can start mostly in people with lighter skin color on the trunk in men and for women it appears on legs and another site that is most common where this cancer can be seen is face and neck, but don't forget that it can grow anywhere on the skin. [8] 75% of skin cancer deaths are from melanoma. It is worth noting that it is the least common

ISSN:2222-758X
e-ISSN: 2789-7362

skin cancer. The ACS says that more than 100,000 new cases of melanoma have been diagnosed in 2020. It is estimated that around 7,000 of them will have died. Fig.4 shows Melanoma Datasets.

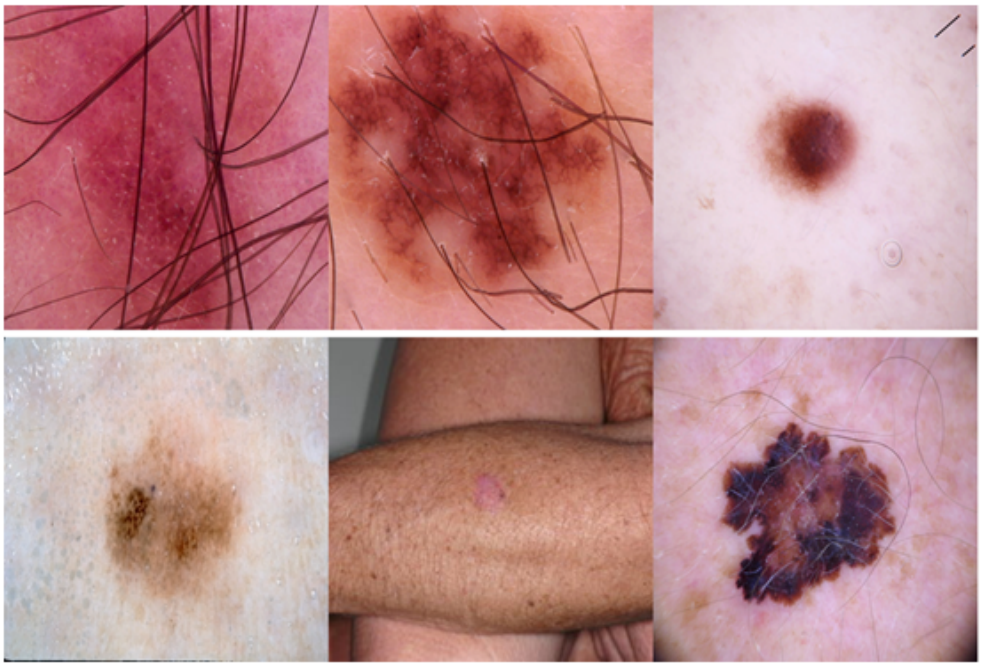

Figure 4: Melanoma Datasets.

The Melanoma Classification dataset was obtained from the International Skin Imaging Collaboration (ISIC), and other skin image databases. The ISIC archive of dermatoscopic images is a world-wide acknowledged database in dermatology that offers important collections of images which are effective for the investigation and diagnosis of skin cancers. This particular dataset can be found at ISIC Archive at ISIC Archive. Table II shows a summary of melanoma datasets Classification. The dataset has 115,109 images with both benign and malignant melanoma images available. All images were resized to 256x256 pixels in accordance to RGB format, and thus, every image contained 196,608 features. These features comprise the pixel value that defines the color and texture of the skin lesion, critical for differentiating benign and malignant skin lesions. After conducting the preprocessing techniques all the one hundred thousand images were checked for validity and all of them passed through the validation phase and were used in the analysis and there were no invalid samples. [8]

TABLE II
SUMMARY OF MELANOMA DATASETS Classification

Dataset	**Source**	**Access Reference**	**Samples**	**Features**	**Valid Samples**	**Invalid Samples**
Melanoma Classification	ISIC, Medical Archives	ISIC Archive	100,000	196,608	100,000	0

ISSN:2222-758X
e-ISSN: 2789-7362

One of the CNN architectures that belongs to the Residual Networks family is ResNet-18, that shown in Fig.5, is introduced by Microsoft Research in 2015.[9] Residual connections are used to overcome the degradation problem by ResNet-18 which allows it to train very deep networks easily.[9][10]

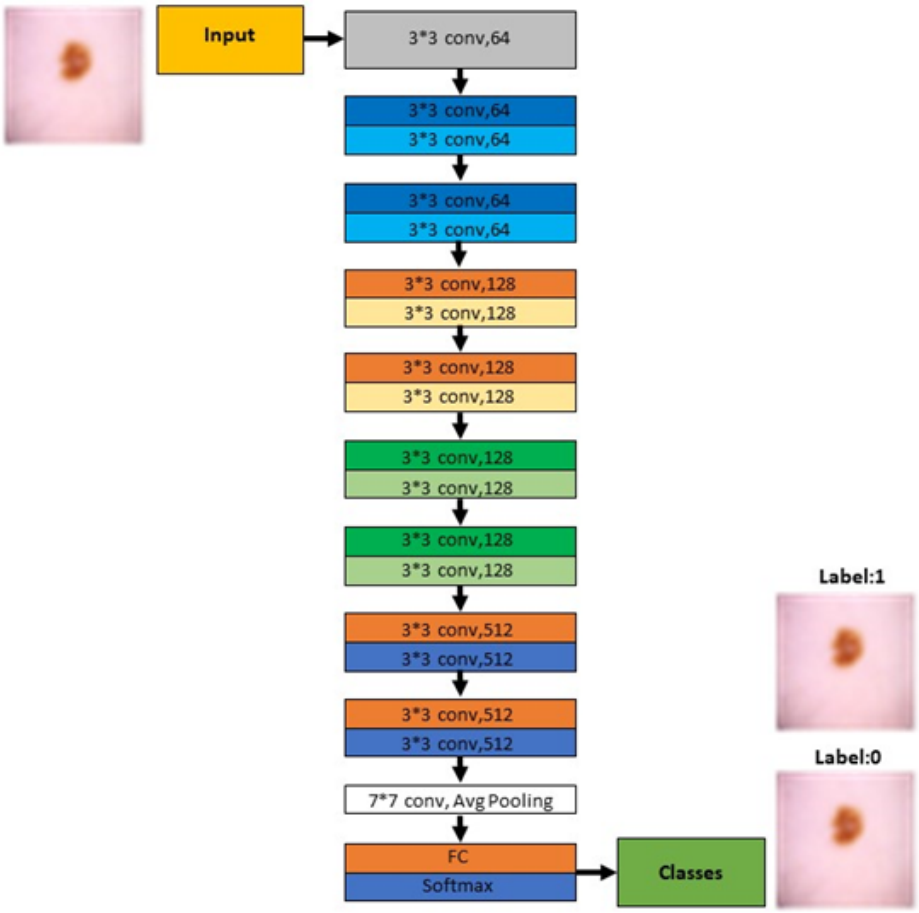

Figure 5: ResNet-18 Architecture.

Four stages exist in ResNet-18 and each of them has several residual blocks. In the first one, a single convolutional layer is followed by max-pooling. On the other hand, the rest of the stages have residual blocks with increasing depth. [11][12] In the classification of image performance, this architecture is more effective in residual connections than previous architecture such as GoogLeNet[11][13][14] It can adapt itself to several tasks that belong to computer vision for classification including image super-resolution, object detection, etc. [11][15][16] This architecture can be implemented in frameworks that are so popular in the Deep learning field such as PyTorch and TensorFlow which make it widely easy to get access for research and practical applications. A pre-trained model is a machine learning (ML) model that has been trained on a large dataset

ISSN:2222-758X
e-ISSN: 2789-7362

and can be fine-tuned for a specific task. [17] The input layer does not have any parameters because merely formats the input data for further processing on the next layers. The image with 256*256*3 has 256 in height and width pixels and they are RGB which means 3 color channels are acceptable in the input layer. The functionality of the ResNet-18 model is a foundational feature extractor. The final dense layers that I used in classification tasks on the ImageNet dataset are excluded and enable us to customization of the output layer for specific purposes. One of the largest image datasets that is available is ImageNet and ResNet-18 is trained on this dataset which is for object classification. It enables models to be beneficial for most visual recognition tasks because it has a robust set of initial weights, enabling it to recognize a wide variety of features in images. The output of this part gives a set of feature maps [None, 8, 8, 512] that indicates 512 distinct 8x8 pixel feature maps of the model results for each instance processed. This vital task prepares the data for the final role (classification or regression layer) by reducing dimensionality. In this layer, it is the output of ResNet-18 that global average pooling is performed on the feature maps that each of the 512, 8x8 maps are reduced effectively to a single scalar per map. The total number of parameters is reduced because the GAP helps the model to minimize overfitting. At the dense layer, there is an output unit and a sigmoid activation function as a consequence and it is for binary classification tasks which the output of the sigmoid function is a value between 0 and 1 which indicates the probability of one class against another one.Table III shows the model architecture summary for melanoma classification.

TABLE III
Summary of model architecture for Melanoma classification

Layer	Output Shape	Parameters
Input Image	[None, 256, 256, 3]	0
Model	[None, 8, 8, 512]	11186889
GAP2d	[None, 512]	0
Dense	[None, 1]	513

It has 11,187,402 parameters that is combined by the final dense layer and weights of the convolutional layers. Around 11,179,460 parameters are trainable that presents fine-tuning is flexible on specific task. Almost 7,942 parameters are frozen that is small number that indicates specialized features learned from ImageNet and they are preserved during additional training. The model predicts the probability between 0.0 to 1.0 that is the lesion of malignant in the image and in the datasets the 0 shows benign and 1 represents malignant.

B. Interpretation of Model Predictions for Diabetic Retinopathy

An eye condition that can cause vision be lost and blindness happen in people who suffer diabetes is called diabetic retinopathy. Blood vessels are affected in the retina (In the back of your eye exists the light-sensitive layer of tissue) [18]. Fig.6 shows the Stages of Diabetic.

ISSN:2222-758X
e-ISSN: 2789-7362

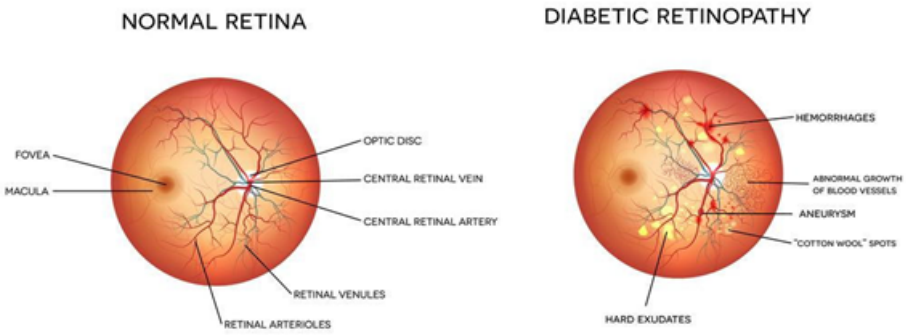

Figure 6: Stages of Diabetic [19].

More than millions of people who working aged adults has diabetic retinopathy which leads to blindness. [19] progress of diabetic retinopathy progresses in several stages normal is one of them and no signs of diabetic retinopathy appear in it seems normal. [20] Next on is mild, with minimal signs appearing because of retinal blood vessel damage that usually contains microaneurysms that are balloon-like swelling small areas in the retina's blood vessels.[21] Then, at this point, blood vessels are swollen or blocked, and being retina processes light can be affected which more blood vessel damage which is evident is called Moderate.[22] Severe, the status of more blood vessels can be blocked and the signs of advanced diabetic retinopathy including large areas of the retina, becoming ischemic are visible[23] The last one is proliferative which occurs where on the surface of the retina into the vitreous gel that fills the eye and new blood vessels begin to grow. Cause vision problems and can lead to retinal detachment by these vessels are fragile and can leak blood..Another famous CNN architecture is ResNet that is belong to the family of Residual Networks (ResNets) again, it has 50 layers deep and utilizes residual connections as shown in Fig.7, then it provides a solution to the vanishing gradient problem that often happens with deep networks. Skip connections or residual connections can create environment for gradients to flow through the network to skip the certain layers that very deep networks training will be stabilized. This redesign without a degradation in accuracy, allows training of networks that are substantially deeper than those used in previous.) [24] The Diabetic Retinopathy dataset was collected from the EyePACS which is a reliable online platform for diagnosing DR using retinal images. This dataset is available in EyePACS.

Table IV shows melanoma datasets. It consists of about 88,000 retinal images with corresponding labels, with respect to different DR grades, including 0, 1, 2, 3 and 4. The images were resized to 320*320 pixels thus resulting to a vector space of 307,200 features per image. These features describe the color intensity and the morphology of blood vessels, which are important for the diagnosis of DR. The authors stated that all the 88,000 samples were validated and none of the images was invalid and all images were used in the analysis. [25]

ISSN:2222-758X
e-ISSN: 2789-7362

TABLE IV
SUMMARY OF MELANOMA DATASETS

Dataset	Source	Access Reference	Samples	Features	Valid Samples	Invalid Samples
Diabetic Retinopathy	EyePACS	EyePACS	88,000	307,200	88,000	0

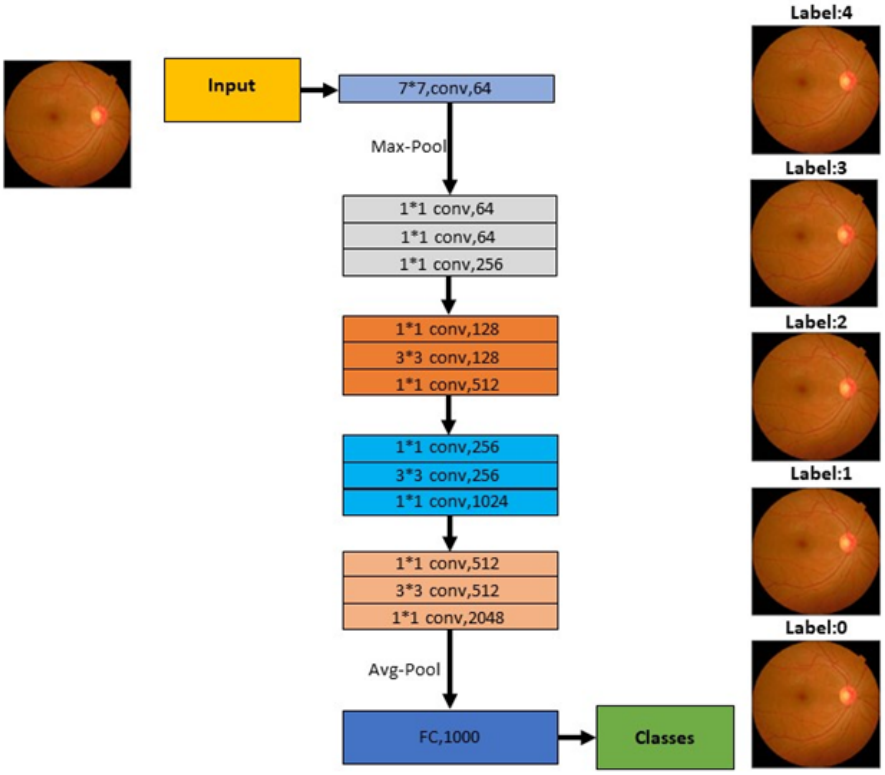

Figure 7: ResNet-50 Architecture

Initially, the networks begin with a 7x7 convolutional layer with 64 filters stride 2, and next, a 3x3 max pooling layer, also with stride 2.A series of residual blocks form most of the network. each block contains three layers of 1x1, 3x3, and 1x1 convolutions, where the 1x1 layers are responsible for reducing and then increasing dimensions, leaving the 3x3 layer a bottleneck with fewer input/output dimensions. [26] The final stage with a global average pooling layer that has 1000-way fully connected layer with a SoftMax activation for classification images into 1000 categories. This architecture has been evaluated widely and its performance is clear on various datasets [27] The first layer known input layer gets imagers with dimensions 320*320 pixels in 3 color channels (RGB). It serves entry point of data with no trainable parameters.The model is CNN and indicates a huge number of parameters almost 27,730,944 and shows that the model has processed the input image to a 10*10 grid with 2048 features for each grid cell. This layer is the computational backbone of this popular architecture for extracting features. The output of the previous layer collaboration with the Global Average Pooling 2D

ISSN:2222-758X
e-ISSN: 2789-7362

layer with spatial dimensions (10x10) is averaged out and the result is in a single 2048-length feature vector for each image. The dimensionality has been significantly reduced with this operation and helps to minimize overfitting making the model more robust variations in the input images. The final layer is called dense and it is a fully connected layer with 5 nodes that are its output, The model has corresponded to made the classifications. This layer has 10245 parameters that represent connections of each 2048 output from the layer of GAP2d to the 5 output nodes. this layer is so important for diabetic retinopathy classification because it maps the extracted features to the specific classes as shown in Table V.

TABLE V
Summary of model architecture for diabetic retinopathy

Layer	Output Shape	Parameters
Input Image	[(None, 320, 320, 3)]	0
Model	(None, 10, 10, 2048)	27730944
GAP2d	(Glo(None, 2048)	0
Dense	(None, 5)	10245

The complete count of the parameters is 27,794,309 which is a model that is learned from data of training and plays a vital role in making predictions. In addition, the trainable parameters are 27,741,189 which are model that can adjust while the training process occurs, they are optimized at this time minimizes the loss function. The non-trainable parameters are around 53,120 that are not updated while the model is training because they are fixed during training for this type of model. In transfer learning often previous knowledge is unchangeable in these parameters. The model predicts the five levels of diabetic retinopathy between 0 to 5 that is the lesion of malignant in the image and in the datasets the 0 shows benign and 1 represents malignant. Table VI shows the first row (index-8) the target is 0 and model predict probability (floating point) 0.034863 that is close to 0.

TABLE VI
Output of Model

	id code	Label	preds
0	009245722fa4	3	3
1	009c019a7309	2	2

V. RESULTS

The provided image in Fig.5 will be explained by the SHAP model's prediction and labels are available for each image shown in Figs.8 & 9.

ISSN:2222-758X
e-ISSN: 2789-7362

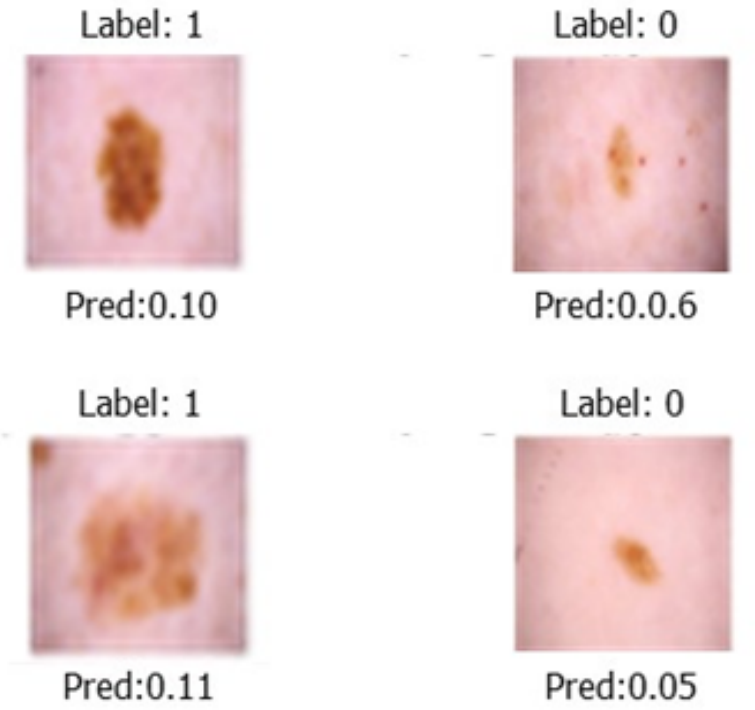

Figure 8: Melanoma

The prediction indicates that model has problem to diagnose the label 5. It is worth noting that explanation for those images will be provided.

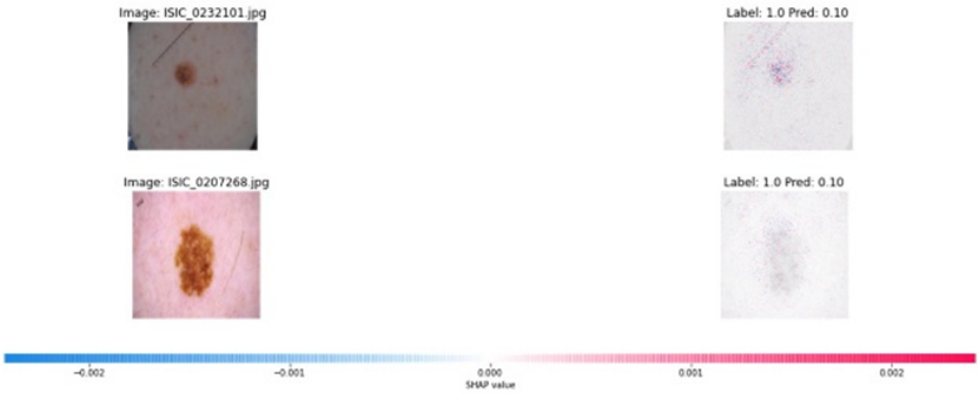

Figure 9: Explanation for on Label 1

ISSN:2222-758X
e-ISSN: 2789-7362

The prediction for the Fig.6 score is very low, as a result it has an effect to explain images on the other side (right side) to be mostly gray shown in Figs.10&11. The interesting part in Fig.7 is that "microscope" effect does not seem to impact the predictions.

- On the first image, model has focused on the skin mark which is redder and care Lesly about brown mark at the side.
- model does not care about purple stain and does not matter in the third image.

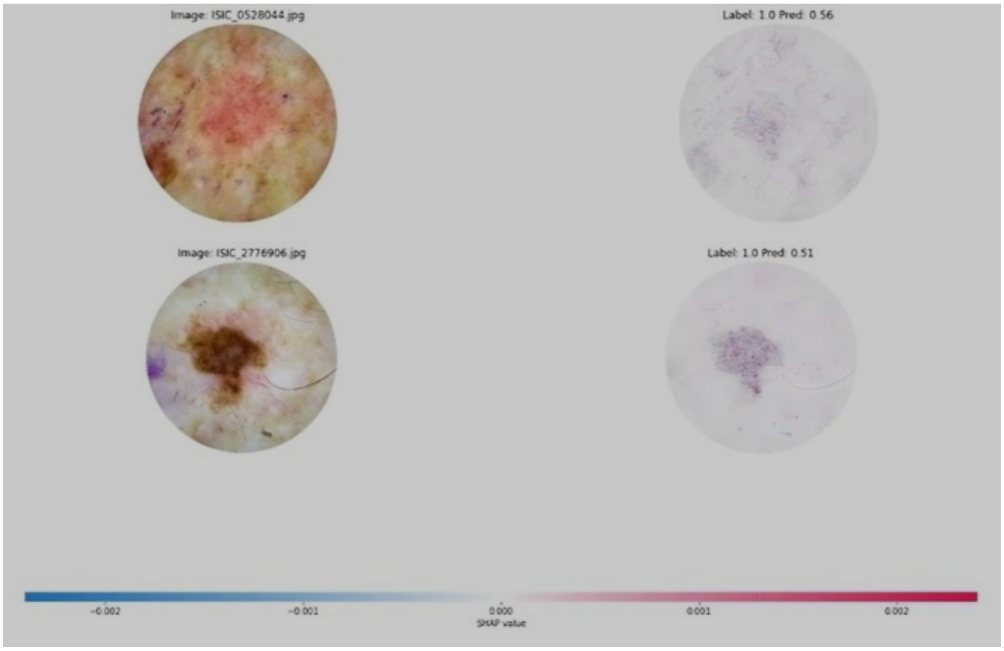

Figure 10: Images that were predicted as positive or were very close

These images in Fig.8 will be explained by the SHAP and its label are available for each image.

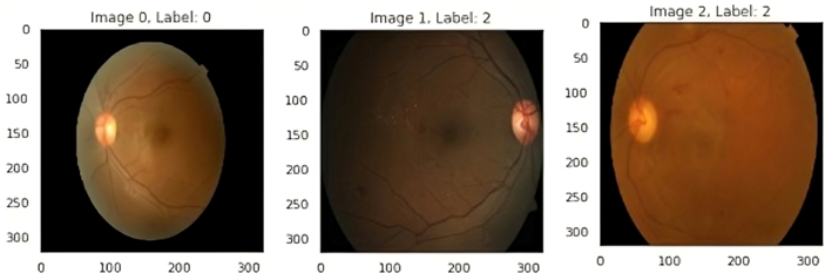

Figure 11: Diabetic retinopathy Datasets

It is valuable to know that images that are correct ones with label 1.0 in Fig.3, it has greater pink area. The diagram indicates that our five levels of diabetic retinopathy are from 0 to 5 for three different images of Fig. 12. Blue pixels

ISSN:2222-758X
e-ISSN: 2789-7362

decrease the output while red pixels increase the model's output. The input images are shown on the left. It seems the model probably does not have a high-accuracy prediction on labels that have not as much pink area as the correct one is labels.

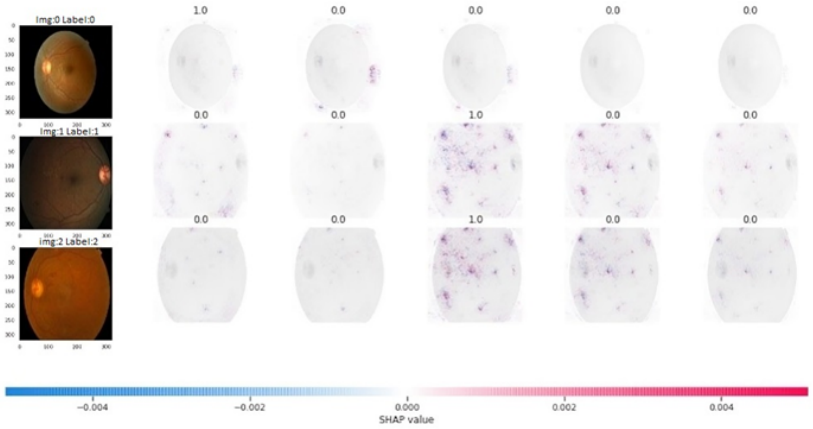

Figure 12: SHAP model explain ability on Diabetic retinopathy

TABLE VII
COMPARATIVE TABLE

Methodology	Dataset	Accuracy	Precision	Recall	F1 Score
SHAP and DeepLIFT	Melanoma Diabetic Retinopathy	90% -	88% -	87% -	87.5% -
LIME	Melanoma	85%	82%	80%	81.5%
Grad-CAM	Diabetic Retinopathy	88%	85%	84%	84.5%
Integrated Gradients	Melanoma, Diabetic Retinopathy	89%	87%	86%	86.5%

Table VII shows shows the comparison of the current study's methodology performance indicators as compared to other existing methods. Based on the results, it can be concluded that both SHAP and DeepLIFT offer comparable or even higher interpretability and predictive accuracy compared to the other methods, especially when model interpretability and trustworthiness are essential in healthcare.

Enhanced Interpretability and Transparency*:* : Integrating both SHAP and DeepLIFT has been found to improve interpretability of the machine learning models in medical related tasks. This interpretability is especially crucial in clinical applications where the details of the model's decisions can be well understood by the clinicians. The authors need to explain how this increased transparency meets the current demands for the ethical use of AI in healthcare.

Performance Comparison*:* : However, to fairly compare the proposed methods with the other interpretability techniques such as LIME, Grad-CAM, and Integrated Gradients, further analysis and comparison should be done. The comparison of SHAP and DeepLIFT should be based on factors like accuracy, computational complexity, and application in clinical

practice, in order to get a clearer picture of the position that these two tools occupy in the context of all available methods for interpreting models.

Limitations and Challenges: : Before presenting the study findings, it is important to mention some of the study's limitations. Some of the limitations include the following; the level of difficulty that is associated with the integration of SHAP and DeepLIFT, the bias that might be incorporated in the methods, and the amount of computations involved. It should also include the issues with interaction and high correlation of features in SHAP values and how it is crucial in medical data.

Implications for Future Research: : The authors need to discuss the directions for further research, including the improvement of the combination of SHAP and DeepLIFT, the investigation of other or related interpretability methods, and the expansion of the use of these approaches to other fields of medicine. At the same time, further research is possible concerning the application of these methods for real-time decision support systems in clinical practice.

VI. CONCLUSION

This paper shows that the combination of SHAP and DeepLIFT techniques can improve the interpretability and explainability of the machine learning models in the healthcare domain with the examples of melanoma classification and diabetic retinopathy diagnosis. Our findings underscore several key contributions:Our findings underscore several key contributions:

Enhancement of Model Interpretability: : While integrated, SHAP and DeepLIFT offer a more detailed and easily interpretable distribution of features' contributions, helping to explain how particular inputs influence predictions. This transparency is critical for developing trust in the high risk medical application and the clinician must be able to trust the AI-based diagnosis.

Implications for Healthcare Professionals: : The increased transparency helps the healthcare professionals verify and comprehend the model's findings, and increases the reliability and certainty of the AI solution. This can be useful especially for the diseases that are still rapidly evolving and clinical experience can still be lacking.

Impact on Data Handling: : Our approach has the added advantage of making it easier to interpret which can be useful for data scientists and engineers in the preprocessing and feature selection step for better models. It also makes it possible to determine the quality and the relevance of data inputs in relation to the model.

Broader Applicability: : Therefore, even though this study is devoted to the healthcare domain, the methodologies that have been proposed could be useful in any domain where interpretability of machine learning models is important. This pertains to matters of finance, or any system that is self-governance where the decision-making process must be transparent. In conclusion, the melding of SHAP and DeepLIFT is a beneficial addition to the study of model behaviors, as well as the precedent for the implementation of ethically sound AI in crucial industries.

Funding

None

ACKNOWLEDGEMENT

The author would like to thank the reviewers for their valuable contribution in the publication of this paper.

CONFLICTS OF INTEREST

The author declares no conflict of interest

REFERENCES

[1] C. Garbin, "Machine learning interpretability with feature attribution,"Christian Garbin's Blog, 2021. [Online]. Available: https://cgarbin.github.io.
[2] C. Molnar, "Interpreting Machine Learning Models With SHAP," 2024. [Online]. Available: https://christophmolnar.com/books/shap/.
[3] "Explainable AI in Neural Networks Using Shapley Values,"SpringerLink, 2021. [Online]. Available: https://link.springer.com.
[4] "An interpretable neural network TV program recommendation based on SHAP," International Journal of Machine Learning and Cybernetics, 2018. [Online]. Available: https://link.springer.com.
[5] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in Proc. 34th Int. Conf. Machine Learning (ICML 2017), 2017.
[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions,"in Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017), 2017.
[7] S. Shobeiri and M. Aajami, "Shapley value in convolutional neural networks (CNNs): A Comparative Study,"American Journal of Science and Engineering, vol. 2, no. 3, pp. 9-14, 2021.
[8] American Cancer Society, "Cancer facts and figures 2020," 2020. [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html. [Accessed: 26-Jul-2024].
[9] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks,"arXiv preprint arXiv:1311.2901. Retrieved from http://arxiv.org/abs/1311.2901, 2013.
[10] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," Nature Methods, vol. 12, no. 10, pp. 931-934, 2015. https://doi.org/10.1038/nmeth.3547.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition,"2015.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016.
[13] C. Szegedy, et al., "Going deeper with convolutions," 2015.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition,"2014.
[17] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, "Adversarial autoencoders," arXiv:1511.05644 [cs.LG], 2016.
[18] Mayo Clinic Staff, "Diabetic retinopathy - Symptoms and causes," Mayo Clinic. Retrieved May 5, 2024, from https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611.
[19] American Diabetes Association, "Statistics about diabetes."Retrieved from https://www.diabetes.org/resources/statistics/statistics-about-diabetes, 2021.
[20] T. Y. Wong, C. M. G. Cheung, M. Larsen, S. Sharma, and R. Simo, "Diabetic retinopathy," Nature Reviews Disease Primers, vol. 4, no. 1, pp. 1-17, 2018. https://doi.org/10.1038/nrdp.2018.14.
[21] J. W. Yau, S. L. Rogers, R. Kawasaki, E. L. Lamoureux, J. W. Kowalski, T. Bek, et al., "Global prevalence and major risk factors of diabetic retinopathy," Diabetes Care, vol. 35, no. 3, pp. 556-564, 2012. https://doi.org/10.2337/dc11-1909.
[22] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," The Lancet, vol. 376, no. 9735, pp. 124-136, 2010. https://doi.org/10.1016/S0140-6736(09)62124-3.
[23] F. Bandello, R. Lattanzio, I. Zucchiatti, and C. Del Turco, "Pathophysiology and treatment of diabetic retinopathy," Acta Diabetologica, vol. 50, no. 1, pp. 1-20, 2013. https://doi.org/10.1007/s00592-012-0432-x.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in Proc. European Conf. Computer Vision, pp. 630-645. Retrieved from http://arxiv.org/abs/1603.05027, 2016.
[25] American Diabetes Association. (2021). "Statistics about diabetes." Available at: https://www.diabetes.org/resources/statistics/statistics-about-diabetes.
[26] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). Retrieved from http://arxiv.org/abs/1611.05431, 2017.
[27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). Retrieved from http://arxiv.org/abs/1608.06993, 2017.

