National College of Ireland

Exploiting Explainable AI tools for improving model classification explainability

MSc Research Project
Masters of Science in Artificial Intelligence

Ivan Rodriguez
Student ID: x23430133

School of Computing
National College of Ireland

Supervisor: Arundev Vamadevan

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ivan Rodriguez
Student ID:	x23430133
Programme:	Masters of Science in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Arundev Vamadevan
Submission Due Date:	11/08/2025
Project Title:	Exploiting Explainable AI tools for improving model classification explainability
Word Count:	4479
Page Count:	15

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Ivan Rodriguez
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploiting Explainable AI tools for improving model classification explainability

Ivan Rodriguez
x23430133

Abstract

As the black-box nature of complex AI models presents challenges for trust and debugging, the field of Explainable AI (XAI) seeks to provide methods for interpreting their decisions. This research investigates how prominent XAI tools can be exploited to improve the explainability of model classifications for stakeholders.

Three XAI tools, LIME, SHAP, and Grad-Cam, were applied to three different CNN models built for healthcare image classification: diabetic retinopathy, COVID19, and blood cells. The tools were evaluated and compared the basis of their performance and the quality of their visual explanations.

The findings show that no single tool is superior in all cases. However, for medical imaging, a combination of Grad-Cam and SHAP provides the most effective explanation; Grad-Cam offers a high-level visual heatmap of relevant areas, while SHAP adds valuable pixel-level precision. Although these tools require additional computational resources, they significantly enhance model transparency and can help justify diagnostic decisions.

1 Introduction

Explainable AI (sometimes abbreviated as XAI) is a relatively novel field of AI with a broad focus on being able to explain or interpret AI models, since some of them have been seen as black boxes for a long time.

The main objective of XAI is to allow stakeholders involved with an AI model to have a way to understand why the model took a certain decision or what caused the decision to move from one outcome to another. This is due to the current black box nature of AI models, where training is defined primarily by the data and the model architecture, but the weights of the decision remain hidden and cannot tell the whole story.Ali et al. (2023)

Counteracting the black-box nature of the models is of paramount importance to the different stakeholders implied in the process of understanding why a decision was made regarding those models. This is why XAI is so important; it allows the 3 different stakeholders to have a window to glimpse the reasons behind the model's decision making. The 3 different stakeholder categories are users, developers, and "external stakeholders".Ali et al. (2023)Barra et al. (2024)

The users are the people who will make use, either directly or indirectly, of the system that includes the AI model. These kinds of stakeholders are concerned with the explainability of the results, since they are directly affected by their outcomes and might need

some explanation that can give some guidance to them or even other users in the case of consultants using AI systems to produce an outcome to present to other humans. In these cases having something other than the outcome to explain the reasoning behind the decision would be of great use.Saeed and Omlin (2023)

For developers, having a way to glimpse into the black box represents a way to debug or improve the current system with more information other than the usual information of having a dataset going in, manipulating the architecture and having some results.Wang et al. (2024)Saeed and Omlin (2023)

The final kind of stakeholders are "external stakeholders" which comprise external sources with interest in having the AI models demonstrate their training is complying with current normative and laws. These kinds of stakeholders are usually concerned with the possible repercussions related to the use of AI models and usually look for ways to ensure themselves and the others the right functioning of the AI models, including holding them accountable not only for the training data used but for the decision making. Examples of these can be ensuring models do not gain biases from their trainers or their datasets.Wang et al. (2024)

Even as a novel branch of AI there exists plenty of work done upon which to build this research paper. The survey papers consulted for this research appropriately resume the state-of-the-art. First, the most relevant thing XAI is trying to develop are methods and tools that will be able to dispel the black box way of seeing AI and gradually be able to include meaningful explanations for the different stakeholders such that the decisions made by those AI models can be explained, understood and be usable by other humans and systems. In other words, build trust. Ali et al. (2023)

On the technical side of things, while there are advances in explainability, these advances are mainly focused on Classification models by using certain tools. The most famous tools are LIME, SHAP and Grad-CAM. These tools have different characteristics, but share some of them, making them comparable to each other or sometimes complementary depending on the overlap of the model being scrutinized and the characteristics applicable from the tool into the model.

In this research, the question of how XAI tools can be used to improve model classification by pointing out areas of opportunity for developers and showing meaningful responses to users will be addressed. In the following sections, the literature related to the problem is explored, the different XAI tools are discussed, and finally, the outline for the research methodology is presented.

The following is a short preview of the contents for the following sections:

Related Work: Other works related to XAI and the use of the 3 main tools is discussed here, as well as background information and essential information regarding the topics that will be discussed through this paper.

Methodology: Details regarding the investigation methodology are reviewed here. This is the theoretic approach for the study, while the following sections concern themselves with the practical approach.

Design Specification: The details of the design for the experiment and the study are shown explicitly.

Implementation: The actual code and technical details are reviewed here. The foundations for the work are also discussed.

Evaluation: The results are compared, explained, and interpreted here. Possible insights are pointed out, and the research question is answered implicitly.

Conclusion and Future Work: Finally, conclusions on the whole study are presented along with possible future work.

2 Related Work

A black box is understood as a system of which the inputs and outputs can be known but nothing is known or shown about the real system.

AI models based on training and learning usually fall into this category of systems, of which the inputs (dataset, train/test split, current input) can be known, the output can be known (classification result, prediction result, etc) but how the actual system works remains a mystery.

There is somewhere a classic example where a computer vision model was trained to detect tanks in the wild but in a weird coincidence, the tank images contained clouds as opposed to the no-tank images, resulting in a model that correlated clouds to the existence of tanks in a picture.

Some tools exist that are already taking care of this problem in their own unique ways.

This research is made in the context of current Machine Learning (ML) and DeepLearning (DL) models still being considered black box(Ali et al., 2023; Zahoor, Bawany and Ghani, 2023; Islam, Assaduzzaman and Hasan, 2024; Sieradzki et al., 2024)Ali et al. (2023)Zahoor et al. (2023)Islam et al. (2024)Assaduzzaman et al. (2025)Sieradzki et al. (2024) and the real struggle of finding better ways to retrain models(Saeed and Omlin, 2023)Saeed and Omlin (2023) that show undesired aspects such as underfitting, overfitting, low accuracy scores. However only classification models will be considered for this study. The tools that will be used are LIME, SHAP, and Grad-CAM. These tools are not yet part of the standard for training classification models, so maybe this study can encourage the standardization of adding a model explainer to the training pipeline to boost confidence and trust in the stakeholders. Findings will be presented as a comparison between several models, their accuracy, the tools used, the processing overhead (as some of these tools can be resource intensive), and the perceived improvement in explainability.

2.1 XAI Definitions and Taxonomy

XAI as a novel field is still struggling with some definitions and limited consensus from the community. One of the basic struggles is to have a consensus on the precise definitions of both words that are often used interchangeably by other researchers in the field: interpretability and explainability.Barredo Arrieta et al. (2020)Liao et al. (2020)Colley et al. (2024)Ali et al. (2023)Saeed and Omlin (2023)Wang et al. (2024)

Making an actual definition for them is well beyond the scope of this paper but it is a good exercise to specify the definition for the current paper. In this paper the following definitions are used: “Interpretability enables developers to delve into the model’s

decision making process"Ali et al. (2023) "It aids in opening a door into the black-box model for users with the required knowledge and skills"Ali et al. (2023) "...explainability provides insight into the DNN's decision to the end-user in order to build trust that the AI is making correct and non-biased decisions based on facts"(Ali et al. (2023). The main distinction here being the recipient of the explanation: Interpretability for technical personnel, Explainability for non-technical personnel.

XAI also has a roughly defined taxonomy. First, the methods used by the tools to explain the AI models can be classified in 2: real time and post hoc Barredo Arrieta et al. (2020)Narkhede (2024). Real time methods just as the name suggests, are methods that are purposely built into the classification system. These systems will have access to the classification as it learns and will be able to determine the features that weigh the most for the model's decision making processBarra et al. (2024). On the other hand, post hoc methods are "late to the party", they are to be implemented to study the decisions of the fully trained model Barra et al. (2024)Narkhede (2024). In this sense, post hoc methods have a clear advantage over real-time methods since there is no need for post hoc methods to be present when the model is being trained, so they can be safely added after the training is done without much overhead Barra et al. (2024)Narkhede (2024).

There are also 2 other relevant concepts: Causality and Robustness. Causality has to do with the models actually having a clear reason for their classification. Models with high causality include the decision trees classifiers where the classes used for their classification can be easily explained. Robustness refers to the stability of the model's decisions. In other words, robustness is a metric that measures the certainty and predictability of the model, not giving seemingly random classifications for the same inputs(Shin, 2021; Ali et al., 2023). This is directly impacted by the bias and variance trade-off of the model. High-bias/low-variance models are prone to underfit in their training so the result is a simpler model that misses the mark but it is easier to explain. The high-variance/low-bias models have a tendency to overfit and have a more nuanced model while being harder to explain Barredo Arrieta et al. (2020)Lundberg et al. (2020)Ali et al. (2023)Gafarov et al. (2024). It might become apparent then, that Causality and Robustness share a connection to Bias and Variance which in turn have a relationship within model training where the objective is to optimize for both. Though not directly stated, this relationship and how well it is optimized can be the key to boost Causality and Robustness, boosting in turn explainability.

Before discussing the existing tools relevant to this study, it is important to mention the terms Local explainability and Global explainability. The key difference between the 2 is that the tools using Local explainability use classification instances for its assessment while the tools using Global explainability will assess the model as a whole based on the decisions taken Barredo Arrieta et al. (2020)Ali et al. (2023)Narkhede (2024)Wang et al. (2024). This an important distinction between tools since the computational overhead for global explainers can be significantly higher than local explainers. This compounded with it being a real time method made SHAP have a significant computational overhead in one of the studies, which can become a significant limitation Lundberg et al. (2020)Barra et al. (2024)Narkhede (2024)Wang et al. (2024).

2.2 LIME Overview

As a tool for XAI LIME is focused on Local explainability, by introducing perturbations to the input data to see the output also perturbed and try to discern the most relev-

ant features for decision making while also being model-agnostic and post hoc. Alami et al. (2024)Barra et al. (2024)Nazir et al. (2024)Sieradzki et al. (2024)Srivastava et al. (2024)Assaduzzaman et al. (2025).

Deals with the uncertainty by going about local experiments and trying to figure out a simplification of the model such that for the given example tests and the obtained results from the model it can figure out roughly what could the model be seeing/processing

As the local part implies, the down sides are that the examples as well as the model simplification are not exactly what the actual model is.

The advantage is the simplification and the low cost of finding out what the model could be doing instead of having a detailed explanation that requires a comparatively enormous quantity of test cases.

2.3 SHAP Overview

SHAP is based on Game Theory and Global explainability. It also is a real time method as mentioned before. The advantage of SHAP is that it is a model agnostic tool, meaning it can be used regardless of the classification model being explained Parsa et al. (2020)Alami et al. (2024)Barra et al. (2024)Nazir et al. (2024)Srivastava et al. (2024).

Same as LIME, SHAP is a model agnostic tool that tries to make sense of the model given. Different from LIME, it does this by reading differences in the gradients of the model while training. This becomes the disadvantage for SHAP as it incurs in computational overhead more so than the other models, but gains the advantage of having an overall knowledge of the model because of the overall gradient monitoring.

2.4 GradCam Overview

This tool is mostly focused for images as it has an underlying pre-training that makes it detect the main features of an image, such as edges, shapes, colors. It is also a Local explainer and post hoc. It has already been tested for non-image classification and compared to LIME and SHAP resulting in poor results. While this assertion is obvious since it is optimized to detect patterns in images it doesn't mean that the basics for this tool can't be translated into another set of problems such as time series, however that discussion is out of the scope of this study Selvaraju et al. (2020)Lim et al. (2021)Alami et al. (2024)Barra et al. (2024)Narkhede (2024)Nazir et al. (2024)Sieradzki et al. (2024)Srivastava et al. (2024)Assaduzzaman et al. (2025).

The huge advantage of GradCam over the others is how visual it is. That is also its disadvantage.

Being a visual tool, it needs data to be in image format in order for it to work. This differentiates it from the previous tools with the disadvantage being GradCam not being model agnostic.

2.5 Gap

Although several sources have suggested using the 3 tools at the same time to interpret a single outcome and some have combined them with limited success or have suggested to combine them in their future work section, few have actually done so. Furthermore this combination also requires certain characteristics from their datasets and models, namely, the models need to be working with image related datasets. Models working with other

kind of datasets, like text, timeseries, tabular data, etc. are not suited for work with Grad-cam even if using Shap and Lime render a result.

Another gap in the literature comes from the lack of standardization regarding the metrics used to compare the results between the 3 tools. Although some attempts have been made by using loose qualitative metrics and some quantitative metrics, a holistic attempt has not been attempted yet.

2.6 Conclusion

Though the advance in the field of XAI is clear, it is also clear that it is not keeping up the pace regarding the rest of the advances in AI. And while these XAI advances remain limited to some extent and to some technologies, they require more work and more finesse to become widely used and implemented. Furthermore, it is necessary for these XAI tools to have a more specific use and guaranties that allow them to become part of the standard and pull more interest their way.

3 Methodology

The current setup for the experiments is as follows. 3 different CNNs designed for different disease recognition have been rebuilt according to their own papers. These are chest rays, retinopathy, and blood cells.

These CNNs have varying degrees of accuracy. However the output is to be evaluated by the previously mentioned tools.

Once a baseline has been achieved, the results of those tools will be evaluated and compared with each other, as well as different other experiments such as shifting CNNs around, different input training conditions, etc.

The accuracy per change on the models is to be recorded along with the evaluated XAI tools.

This should show several important things:

- Changes made to the input training should reflect at the end and when compared to the baseline, the original assumption should hold but not entirely as the change in input should have made a difference.
- One of the tools (alternating) will most of the time not contribute to the overall result. This should highlight in which cases the tool strengths should be used.
- The different conditions for the experiment should not change the results in an extreme way.
- In the rare case one is completely almost obsolete compared to the other 2 or 1 a cause should be searched and found.

4 Design Specification

4.1 Global Architecture

The architecture of the system comprises a retinopathy classification model in the middle that is fed with retinal fundus images. The images are then classified, and the results

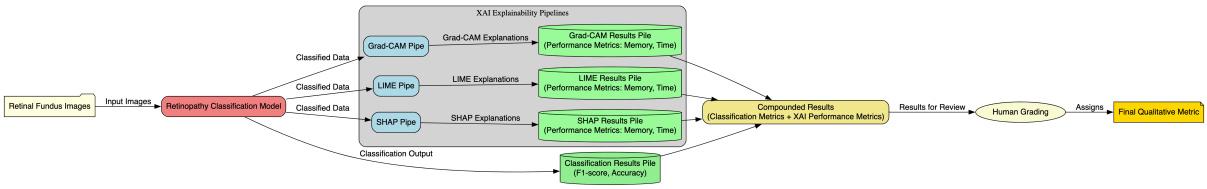

Figure 1: Achitecture diagram for retinopathy as an example of the general architecture

are entered in four different places, a result pile, the GradCam pipe, the LIME pipe, and the SHAP pipe. The results from those pipes go into their respective results piles which resemble the classification result pile, then results from the 4 piles are compounded in a single result which includes the F1-score and accuracy metrics from classification, the performance metrics (memory consumption, processing time) from each of the XAI techniques and are left for a final grading by a human to attain a final qualitative metric of the results.

4.2 Metrics for GradCam, LIME and SHAP

Researching into previous work regarding the comparisons between these methods yielded mixed results as most metrics used are of Qualitative nature instead of Quantitative Nature. This represents a problem because there is no basis for comparison other than a subjective scale. As such, other metrics others have used include Quantitative data such as processing time and memory usage. These 2 metrics are going to be the ones selected, however, since these focus only on the resource usage and not the results, there is need for a Qualitative kind of comparison. This comparison will be aided by including pixel differences and an overall performance taking into account the Quantitative metrics, pixel difference, and the Qualitative assessment.

5 Implementation

5.1 Preprocessing

Retinal data has been collected from github.com/openmedlab/Awesome-Medical-Dataset prioritizing the labeled data. Since the data images are mainly in tiff format and sometimes include black edges, the preprocessing steps include converting to PNG and cutting out the black edges. This is done using the tools included in Tensorflow, Open-CV, and Numpy.

For the Blood Cells Dataset, the only preprocessing necessary is to resize the images so that they are in a smaller size for easier training. No preprocessing was necessary for the COVID19 dataset.

Regarding the folder structure needed for classification, all these datasets were already in the proper structure, splitting their data for train, test, and validation, as well as having inner splits for their classification classes.

5.2 CNN

A basic CNN structure was used for Blood Cells and Diabetic Retinopathy classification, this structure included a Glorot Uniform initializer for the initial weights, an intial shape of 120x120x3 (image size and color channels), 2 Convolutional blocks with Max Pooling, then 2 Dropout blocks, a series of dense layers with dropouts and hyperbolic tangent activation, and a final dense layer with the classes needed for classification as output.

A similar CNN was used for COVID19 classification, skipping the dense dropout layer steps at the end.

5.3 XAI Tools

Having everything else setup and ready, the only missing part is adding the XAI tools. First, an image is extracted from the test dataset; this image will be predicted and used for explanations. Then SHAP is used to extract the gradient values; these values are plotted using the image as the background, so the plotting has meaning. Afterwards, the LIME image explainer is used to obtain an explanation mask which is overlayed on the background image, similar to SHAP. Finally, Grad-Cam is used to get a heatmap image explanation, overlayed, and saved as output.

6 Evaluation

The 3 methods for XAI (LIME, SHAP, GradCam) were added to replicated CNNs for different healthcare studies related to different fields where the common ground was the use of images for disease diagnosis. In the case of retinopathy, fundus images are used for diagnosisAli et al. (2023). In the case of COVID19, chest-ray imagesBhandari et al. (2022). In the case of blood cells (which are not actually diagnosed, rather classified) its microscope imaging Islam et al. (2024). All of these studies were replicated and then enhanced their explanations using the 3 previously described Explainable AI methods. The datasets used were as close as possible to the ones suggested in each of their hailing papers, and the individual results, as well as architecture and observations, will be explained in detail in the following sections.

6.1 Retinopathy / Case Study 1

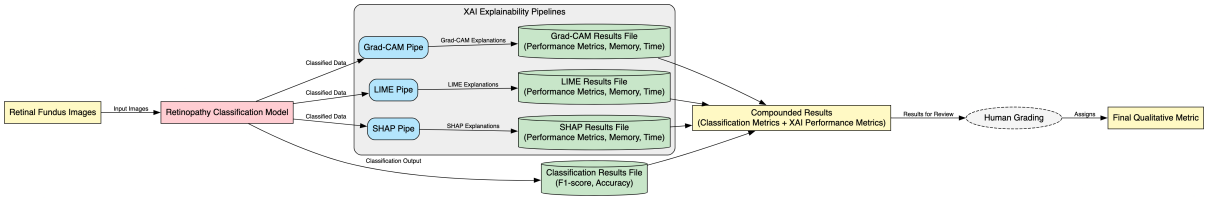

Figure 2: Architecture for XAI interpretations for Diabetic Retinopathy detection

Classification for Diabetic Retinopathy using a CNN gave excellent accuracy score results. The training and validation graphs for accuracy and loss show good climb and drop, respectively.

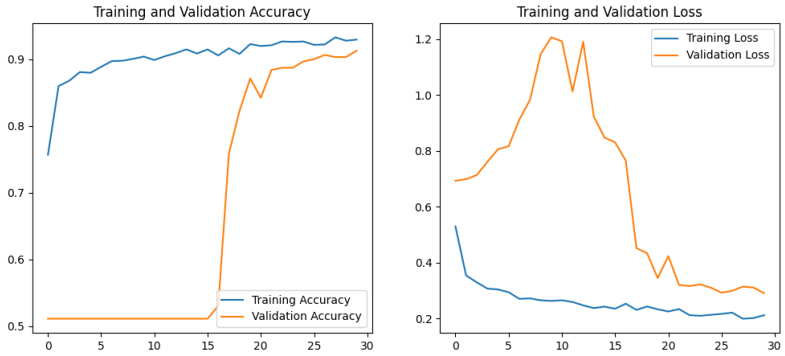

Figure 3: Accuracy and Loss for Retinopathy model training

Regarding the explanations, initially SHAP and LIME are not very specific on their own, but after Grad-cam outputs the heatmap, the evaluated region becomes apparent. Compared with SHAP and LIME, the better pairing for a human to figure out what is being classified is SHAP and Grad-cam, where Grad-cam can show the general region and SHAP pinpoints interest pixels. In this case, the LIME explanation is vague and does not match SHAP or Grad-Cam in any way.

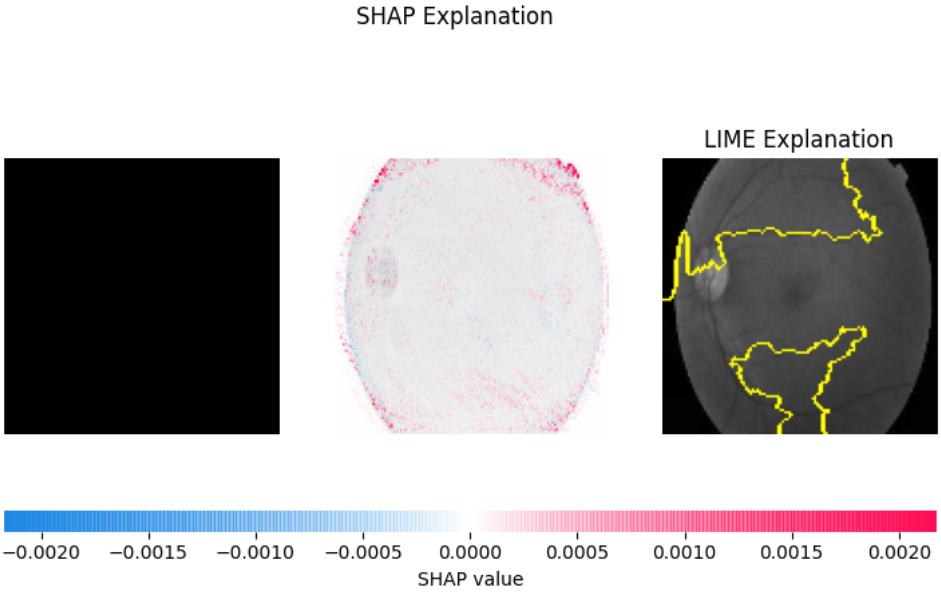

Figure 4: LIME and SHAP explanations

6.2 COVID19 / Case Study 2

The corresponding Training and Validation graphics for the classification of COVID19 images show that this CNN in particular could have benefited from the final dropout layer

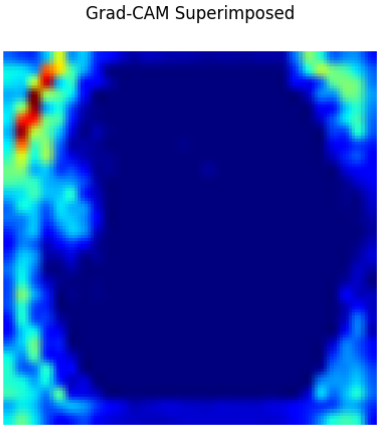

Figure 5: Grad-Cam heatmap Retinopathy detection

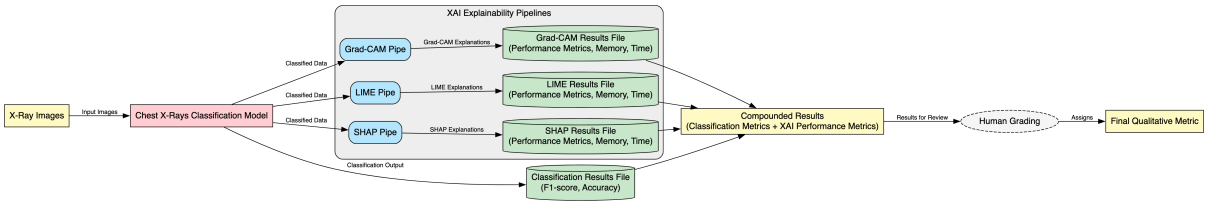

Figure 6: Architecture for XAI interpretations for COVID19 detection

steps at the end. The Accuracy graph as well as the Loss graph show hints of overfitting at the end of the training epochs.

In this case, the roles were inverted by LIME showing an area quite similar to Grad-Cam while SHAP failed to show any results.

6.3 Blood Cells / Case Study 3

Although the same CNN was used for the classification of Blood Cells as for the classification of Diabetic Retinopathy, and no jump in training shows overfitting as in COVID19, the accuracy score for this classification was quite low, probably due to increased classification classes.

Regarding the interpretations, pairing SHAP and Grad-cam seems to point to a particular region, while LIME focuses on a different region but not far from the others. In particular, SHAP values seem more scattered than LIME and Grad-cam, which might point towards lower accuracy affecting the decisiveness form SHAP compared to the previous experiments. This is not observed in the explanation given by LIME but it seems to be present to some extent on the heatmap form Grad-cam.

6.4 Discussion

Training the models presented several challenges. Regarding the architectures, some crucial missing information made the models perform with a subpar accuracy and F1-score compared to their in-paper counterparts. This deficit, although noticeable, can be

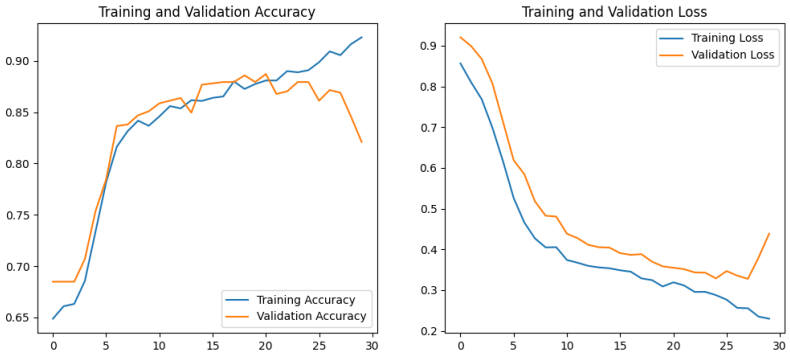

Figure 7: Accuracy and Loss for COVID19 model training

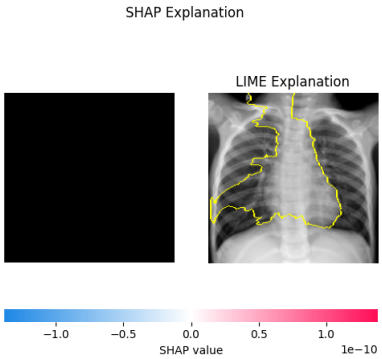

Figure 8: SHAP and LIME interpretations

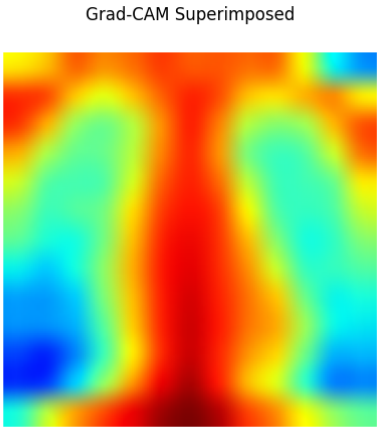

Figure 9: Grad-Cam heatmap for COVID19

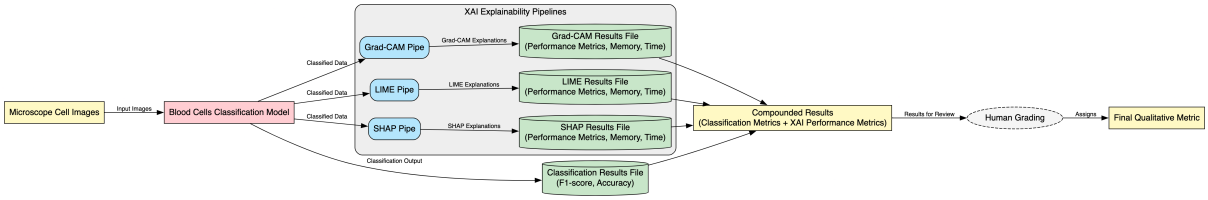

Figure 10: Architecture for XAI interpretations for Blood Cell classifications

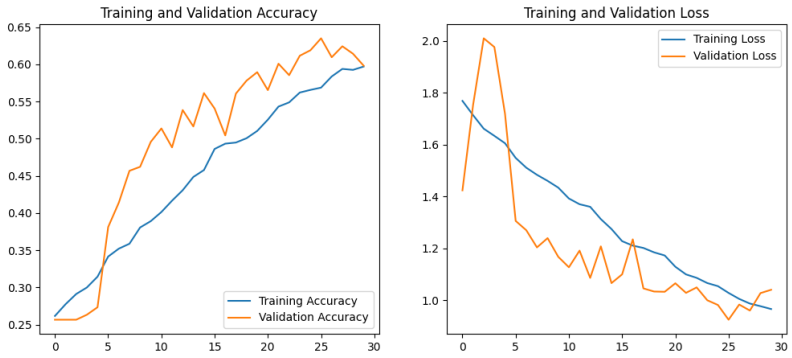

Figure 11: Accuracy and Loss for Blood Cell model training

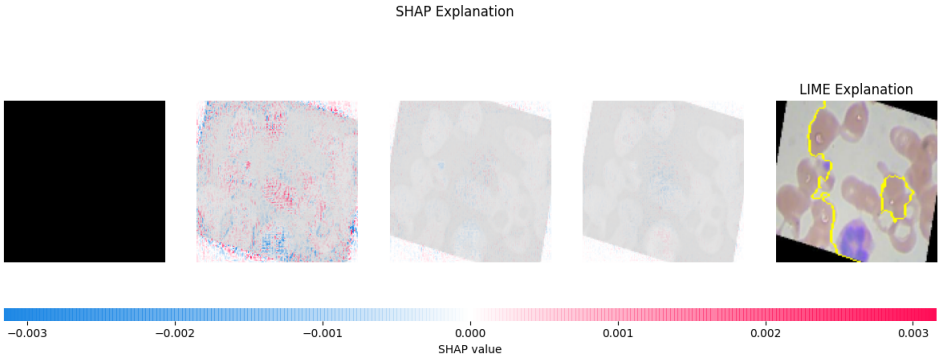

Figure 12: SHAP and LIME Blood Cells Interpretations

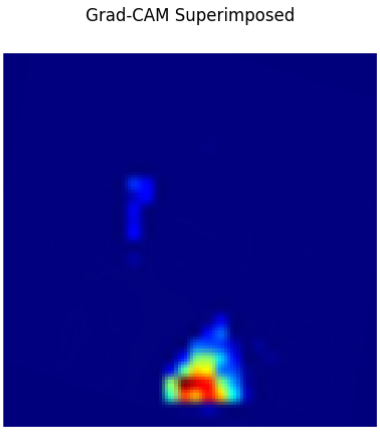

Figure 13: Grad-Cam heatmap for Blood Cell classification

negligible for the qualitative part of the study.

As discussed previously, no single XAI tool is overly determinant or unimportant. Grad-Cam works perfectly as it is its intended purpose to graphically signal the most relevant part detected for classification. LIME was the most obtuse to interpret since the points highlighted on the output image did not amount to an immediate significant evaluation until compared with either SHAP or Grad-Cam as reference, then using both it became apparent what LIME meant to highlight. This is not to signify it is not useful as Lime was the most erratic method of evaluation, working on a single local interpretation. As such, the most efficient pairing for explanations from a visual perspective would be GradCam and SHAP. From an efficiency standpoint, SHAP and LIME are better suited since they take less time to generate.

7 Conclusion and Future Work

7.1 Conclusion

Can XAI tools be used to improve model training and as justification of its decisions? 3 different graphical classification models were recreated, studied, trained, and used to generate explainable output using the 3 XAI tools LIME, SHAP, Grad-Cam. These models posed different training challenges and explainability levels. Different preferred combinations were found depending on the constraints per case. Resource constraints, time constraints.

When working with healthcare image data classification, Grad-cam is definitely recommended because of the graphical explanations and heatmaps that help to easily pinpoint the classification sources. Additionally, SHAP can be paired with Grad-cam if resources are available, since the granularity of the gradient explanations gives the precision that Grad-cam lacks to actually point to pixels in the pictures. Lime on the other hand is recommended along Grad-cam as it is effective in matching the "big picture" Grad-cam paints but on its own Lime lacks both precision and "big picture", and can even point at the wrong regions. In general, these tools do need extra resources but can really heighten the explainability for every kind of stakeholder.

7.2 Future Work

The focus on healthcare-related images is clear in this study. A different study focusing on wider areas of image classification could confirm the precision of this study outside of the scope of healthcare.

The Grad-cam version used in this study is the most basic one, but there have been improvements since, it would certainly be interesting to use that new version to confirm or discard the notion of using Grad-cam along other tool under the assumption that newer versions of Grad-Cam could be much better than the current one.

References

Alami, A., Boumhidi, J. and Chakir, L. (2024). Explainability in CNN based deep learning models for medical image classification, *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1–6. ISSN: 2768-0754.
URL: *https://ieeexplore.ieee.org/document/10620149*

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N. and Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, **99**.

Assaduzzaman, M., Bishshash, P., Nirob, M. A. S., Marouf, A. A., Rokne, J. G. and Alhajj, R. (2025). XSE-TomatoNet: An explainable AI based tomato leaf disease classification method using EfficientNetB0 with squeeze-and-excitation blocks and multi-scale feature fusion, **14**: 103159.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S221501612500007X*

Barra, P., Della Greca, A., Amaro, I., Tortora, A. and Staffa, M. (2024). A comparative analysis of XAI techniques for medical imaging: Challenges and opportunities, *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6782–6788. ISSN: 2156-1133.
URL: *https://ieeexplore.ieee.org/document/10821983*

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, **58**: 82–115.

Bhandari, M., Shahi, T., Siku, B. and Neupane, A. (2022). Explanatory classification of CXR images into COVID-19, pneumonia and tuberculosis using deep learning and XAI, **150**.

Colley, A., Väänänen, K. and Häkkilä, J. (2024). Tangible explainable AI-an initial conceptual framework, *ACM International Conference Proceeding Series*, pp. 22–27.

Gafarov, F. M., Gafarova, V. R. and Ustin, P. N. (2024). Explainable artificial intelligence methods in text classification machine learning models for investigating human behavior in digital environments, *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, pp. 1660–1664.
URL: *https://ieeexplore.ieee.org/document/10805072*

Islam, O., Assaduzzaman, M. and Hasan, M. (2024). An explainable AI-based blood cell classification using optimized convolutional neural network, **15**.

Liao, Q., Gruen, D. and Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences, *Conference on Human Factors in Computing Systems - Proceedings*.

Lim, B., Arık, S., Loeff, N. and Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting, **37**(4): 1748–1764.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees, **2**(1): 56–67.

Narkhede, J. (2024). Comparative evaluation of post-hoc explainability methods in AI: LIME, SHAP, and grad-CAM, *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, pp. 826–830.
URL: *https://ieeexplore.ieee.org/document/10762963*

Nazir, M. I., Akter, A., Hussen Wadud, M. A. and Uddin, M. A. (2024). Utilizing customized CNN for brain tumor prediction with explainable AI, **10**(20): e38997.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S2405844024150281*

Parsa, A., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis, **136**.

Saeed, W. and Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, **263**: 110273.
URL: *https://www.sciencedirect.com/science/article/pii/S0950705123000230*

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization, **128**(2): 336–359.

Sieradzki, A., Bednarek, J., Jegorowa, A. and Kurek, J. (2024). Explainable AI (XAI) techniques for convolutional neural network-based classification of drilled holes in melamine faced chipboard, **14**(17).

Srivastava, K., Sorathiya, A., Mehta, J. and Chotaliya, V. (2024). Enhancing interpretability, reliability and trustworthiness: Applications of explainable artificial intelligence in medical imaging, financial markets, and sentiment analysis, *2024 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–9.
URL: *https://ieeexplore.ieee.org/document/10607516*

Wang, Z., Huang, C. and Yao, X. (2024). A roadmap of explainable artificial intelligence: Explain to whom, when, what and how?, **19**(4): 1–40.
URL: *https://dl.acm.org/doi/10.1145/3702004*

Zahoor, K., Bawany, N. and Ghani, U. (2023). Explainable AI for healthcare: An approach towards interpretable healthcare models, *2023 24th International Arab Conference on Information Technology, ACIT 2023*.