Enhancing Transparency in Complex Machine Learning Models through Explainable AI Techniques

1. Abstract

The rapid rise of complex machine learning models, particularly deep neural networks, has revolutionized industries such as healthcare, finance, and autonomous systems. However, their opaque "black-box" nature raises critical concerns about trust, accountability, and fairness. This study explores Explainable AI (XAI) techniques to address these challenges, focusing on enhancing the interpretability of such models and identifying potential biases in their predictions. Specifically, this work evaluates methods like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), which are widely regarded for their ability to provide insights into model behavior.

As an initial step of toy problem, the Iris dataset was selected as a toy problem to investigate these techniques on a manageable yet meaningful scale. Using a deep neural network as the representative complex model, the implementation of LIME provided localized explanations, highlighting the importance of individual features in specific predictions. Building on this foundation, the toy problem transitioned to a more complex and clinically significant problem: classifying peripheral blood cell images.

This work underscores the crucial role of XAI in addressing the ethical and interpretability challenges posed by advanced machine learning models. By leveraging LIME and SHAP, the study contributes practical insights into making AI systems more transparent, equitable, and reliable across diverse applications.

2. Introduction

Context and Motivation

The rapid rise of deep neural networks and other complex machine learning models has profoundly impacted fields such as healthcare, finance, and autonomous systems. However, their "black-box" nature raises significant concerns about trust, fairness, and accountability, especially in high-stakes domains where understanding model decisions is critical. Explainable AI (XAI) techniques have emerged as a promising solution to address these challenges by enhancing model interpretability and promoting ethical AI usage.

Survey Scope and Method Choice

This study investigates two prominent XAI techniques, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME excels at providing localized insights, particularly in smaller datasets, while SHAP offers a more comprehensive view of global model behavior. Together, these techniques provide a robust framework for analyzing and interpreting complex machine learning models.

The Iris dataset was initially chosen as a toy problem to explore the application of these techniques on a manageable yet meaningful dataset. Using a deep neural network as the representative complex model, the implementation of LIME enabled an in-depth analysis of individual predictions and feature importance. Following this initial phase, the study extended to a clinically relevant problem: classifying peripheral blood cell images. The transition demonstrated the scalability and effectiveness of XAI techniques in high-stakes medical applications, reinforcing their potential for broader real-world impact.

Toy Problem/Dataset

The Iris dataset serves as an initial step to test interpretability techniques on a simple, well-defined classification problem. Comprising three classes of flowers and balanced data, it provides an excellent environment to evaluate XAI methods like LIME and SHAP in a controlled setting. Its small size and clear features facilitate rapid

experimentation, allowing for localized and global model behavior analysis with minimal computational overhead.

Building upon the results of the Iris dataset, the study aimed to extend these methods to a more complex and practically significant dataset of peripheral blood cell images [1]. This dataset, sourced from the Hospital Clinic of Barcelona and consisting of 17,092 images across eight classes, offers a rich testing ground for XAI techniques in a medical domain. The images, captured using the CellaVision DM96 analyzer, represent different blood cell types, making the dataset directly relevant for applications in hematology and automated medical diagnostics. Each image is 360x363 pixels in JPG format, posing a more challenging classification task due to higher dimensionality and visual complexity.

This phased approach ensured that XAI techniques like LIME and SHAP are first validated in a simplified context before being tested on a dataset with higher stakes and practical implications. The peripheral blood cell dataset aligns with the study's broader objective of enhancing trust and transparency in complex models, particularly in sensitive fields such as healthcare. By transitioning from a toy problem to a real-world medical dataset, the study seeks to assess the scalability and robustness of these interpretability methods in addressing intricate, domain-specific challenges

Outline

The structure of this paper is as follows:

- Background and Problem Definition: An overview of XAI concepts and challenges in explaining complex models, defining the problem of enhancing transparency through explainability techniques.
- Literature Review: Summarizes key papers on XAI methods, evaluating their strengths, weaknesses, and relevance to the implementation of LIME and SHAP.
- Methodology and Implementation: Describes the implementation of LIME and SHAP, detailing the choice of the Iris dataset, model architecture, and approach for evaluation.
- Results and Discussion: Discusses findings from LIME and SHAP implementation, including performance metrics and insights
- Challenges and Open Problems: Highlights challenges from both the literature and implementation, and suggests directions for addressing open problems in XAI research.
- Conclusion: Summarizes key findings, discusses limitations of the current work, and outlines future steps for further development and research in XAI.

3. Background and Problem Definition

Concepts and Theories

Explainable Artificial Intelligence (XAI) encompasses techniques aimed at making machine learning models transparent and interpretable, addressing the challenges posed by their often opaque "black-box" nature. The increasing reliance on AI in critical domains such as healthcare and finance has heightened the need for models to be understandable and trustworthy. Without interpretability, it becomes difficult to detect biases, ensure fairness, or gain confidence in the model's predictions.

Key terms and foundations include:

- Local Explanations: Methods like LIME that provide insights into a specific prediction by approximating the model with a simpler, interpretable model in a localized region of input space.
- Global Explanations: Techniques such as SHAP that provide an overall view of feature importance across an entire dataset.
- Model-Agnostic: Refers to methods or techniques that are not tied to a specific type of machine learning model. Instead, they treat the model as a "black box," analyzing its inputs and outputs without requiring access to its internal workings.
- Surrogate Model: Is a simpler, interpretable model used to approximate a more complex, often "black-box" model. The surrogate model is trained to mimic the behavior of the complex model within a specific region or neighborhood of the data

The theoretical basis of XAI combines statistical reasoning, optimization, and machine learning principles. Visualization tools and interpretable machine learning techniques further enhance the understanding of AI models, ensuring that their decisions align with human expectations and ethical standards.

By leveraging these concepts, XAI methods aim to bridge the gap between AI performance and the need for explainability in real-world applications.

Problem Statement

Explainability in artificial intelligence (AI), now referred to as Explainable AI (XAI), is a concept with historical roots extending back to the 1980s and 1990s. Early efforts to address interpretability in AI systems were exemplified by frameworks like the Explainable Expert Systems (EES), developed under the Strategic Computing Initiative by DARPA. The EES framework emphasized justifications of system actions, explanations of problem-solving strategies, and clear descriptions of terminology. For instance, the Program Enhancement Advisor (PEA), an advice system for improving Common Lisp programs, demonstrated how well-designed explanations could enhance user understanding by offering insights into the logic behind recommendations. These early systems laid the groundwork for integrating explainability into AI design, even when such systems were still niche in their application [2].

The significance of explainability grew substantially with the advent of deep learning in the 2010s. As complex machine learning models, particularly deep neural networks, achieved remarkable success across tasks like image recognition, natural language processing, and autonomous systems, the opaque "black-box" nature of these models became a pressing concern. By 2020, explainability was recognized as a key factor for the adoption of AI systems across diverse domains such as autonomous vehicles, medical diagnostics, insurance, and finance. This shift was driven not only by technical needs but also by societal and legal demands, such as the European Union's General Data Protection Regulation (GDPR), which grants consumers the right to obtain "meaningful information about the logic involved" in automated decisions [3].

Explainability is crucial not only for ensuring compliance with ethical and legal frameworks but also for addressing technical challenges. By equipping intelligent systems with explanatory capabilities, designers and developers can improve system robustness, mitigate bias, prevent discrimination, and enhance trust. These motivations underscore the importance of Explainable AI (XAI), which focuses on making complex machine learning models transparent and interpretable. This study builds on these motivations by exploring XAI techniques, particularly Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), to address the challenges of transparency and trust in AI systems, to address the challenge

of scalability from toy problems to real-world, domain-specific datasets., to identify potential biases and limitations in current XAI methods and suggesting future research directions.

Survey Overview

This work surveys the current state of XAI, focusing on methods that address the black-box nature of deep learning models. The techniques under review include:

- **LIME (Local Interpretable Model-agnostic Explanations):** A method for generating locally faithful explanations for individual predictions by approximating the complex model with an interpretable surrogate.
- **SHAP (SHapley Additive exPlanations):** A unified framework based on game theory that assigns each feature a contribution value for model predictions, offering a global view of feature importance.

The exploration of XAI techniques was greatly facilitated by the comprehensive and instructive work provided in [4]. Their study significantly eased the process of finding additional resources and related research on XAI, enabling a more thorough and grounded analysis of the field.

4. Literature Review

Overview of Methods

Two prominent XAI techniques—LIME and SHAP—form the focus of this study, given their widespread use and effectiveness in addressing the interpretability challenges of complex machine learning models.

LIME (Local Interpretable Model-agnostic Explanations) is a novel explanation technique proposed in the paper "Why Should I Trust You? Explaining the Predictions of Any Classifier." [5]. The primary objective of LIME is to provide faithful and interpretable explanations for any classifier by learning a locally interpretable model around the specific prediction being made. The method works by approximating the complex model with an interpretable surrogate model in the neighborhood of the instance being predicted. This allows LIME to explain individual predictions in a way that is both understandable and trustworthy, even for highly complex machine learning models.

LIME emphasizes two key principles for effective explanations: interpretability and local fidelity. Interpretability refers to providing an understanding of the relationship between input features and the model's output, which must be adapted to the user's level of expertise. Local fidelity ensures that the explanation aligns with the model's behavior in the specific area surrounding the instance being explained, even though it may not reflect the global model behavior. This is important because some features that are globally significant may not be as relevant for a particular prediction. LIME is designed to be model agnostic, meaning it can work with any machine learning model, offering flexibility and broad applicability in various contexts [5]

LIME aims to provide understandable explanations for complex machine learning models by focusing on local rather than global interpretability. When a model becomes too intricate, it is unrealistic to try to fit a simple, globally interpretable model that can explain every aspect of the complex classifier. Instead, LIME looks at individual predictions and works within the local neighborhood of a specific instance, *"**a**"*. By sampling nearby instances and using these to construct a simpler model, LIME can give an explanation for a particular decision. This local approach ensures that the explanation remains accurate and interpretable, even when the underlying model is highly complex.

This method works by approximating the original, complicated model with a more interpretable surrogate model that captures the behavior of the classifier in the vicinity of *"**a**"*. By focusing on local relationships, LIME can provide meaningful insights into why the model made a specific prediction for an instance, rather than attempting to explain the entire model. This technique makes it possible to explain even the most intricate models, like deep neural networks, in a way that is understandable to humans [6].

In the case of images, LIME enhances interpretability by dividing the image into "superpixels." Superpixels group neighboring pixels that share similar properties, making them more meaningful for analysis. This segmentation reduces the complexity of the image, enabling LIME to focus on the relevant parts of the image for explanation.

Superpixels are formed using the Normalized Cuts algorithm, which organizes pixels based on texture and contour similarities. This method results in segments that are coherent and relatively uniform in size, simplifying the computation for later stages of interpretation. Highlighting the superpixels that most influence the model's prediction, LIME provides clear insights into the decision-making process of the model [7].

Researchers demonstrate that the definition of the "right" neighborhood for training the surrogate model is far from trivial. The correct sampling strategy plays a significant role in ensuring that the surrogate model accurately reflects the decision boundaries of the complex model. Poorly selected neighborhoods can lead to misleading explanations, where globally important features overshadow locally significant ones, thus distorting the model's behavior [8].

In recent studies, researchers have focused on improving the interpretability of complex machine learning models by refining the balance between explanation stability and adherence. One such contribution introduces the **OptiLIME** framework, which addresses the inherent trade-off between these two factors. Stability refers to how consistent an explanation is when applied to similar instances, while adherence measures how closely the surrogate model approximates the behavior of the original machine learning model. The **OptiLIME** framework aims to maximize the stability of local explanations while ensuring a predefined level of adherence, allowing practitioners to select the most appropriate trade-off for their specific problem. It also highlights the mathematical properties of the explanations provided, offering clarity on the reliability of each explanation [9].

SHAP (SHapley Additive exPlanations) Lundberg and Lee's seminal paper, "A Unified Approach to Interpreting Model Predictions,"[10] introduced SHAP as a framework rooted in cooperative game theory. SHAP values are a direct adaptation of Shapley values from cooperative game theory, applied to machine learning model interpretability. Originally introduced by Lloyd Shapley in 1953 [11], Shapley values are a method for fairly distributing the total payoff of a cooperative game among players, based on their individual contributions to the collective outcome. In the context of machine learning, "players" correspond to features, the "payoff" is the model's prediction, and the "game" is the task of making a specific prediction. This method uniquely unifies six existing feature attribution approaches under a new class of additive feature importance measures. A major contribution of SHAP is its theoretical foundation, proving that there is a single solution in this class that satisfies three desirable properties: **local accuracy**, **missingness**, and **consistency**. These properties ensure that the explanations are robust, intuitive, and aligned with human understanding of feature importance. In SHAP, for model-agnostic explainers like DeepExplainer, you need to define a "background" dataset. The background is a reference or baseline dataset used to calculate the feature importances. It's critical in SHAP for explaining individual predictions, as the model's output is compared against this background to measure the impact of individual features.

Kernel SHAP, a model-agnostic approximation method within the SHAP framework, bridges the gap between classical Shapley values and modern machine learning interpretability techniques like LIME (Local Interpretable Model-agnostic Explanations). This approach assumes that the features are independent. Unlike LIME, which heuristically chooses parameters for its explanation model, Kernel SHAP derives these parameters directly from the mathematical properties of Shapley values. By using a weighted linear regression approach, Kernel SHAP efficiently approximates SHAP values, maintaining consistency and accuracy. It avoids issues such as unintuitive behavior or violations of the aforementioned properties, which can arise in other methods like LIME when parameters are not carefully optimized.

Kernel SHAP's strength lies in its ability to handle complex, non-linear models while remaining computationally efficient. It achieves this by sampling permutations of feature sets and applying a weighting scheme based on the Shapley kernel, ensuring that all possible feature combinations are considered. This allows Kernel SHAP to provide explanations that are both locally accurate and globally consistent, making it a preferred choice for interpreting black-box machine learning models across various domains.

In addition to LIME and SHAP, other XAI techniques like SmoothGrad, Integrated Gradients (IG), and DeepLIFT are commonly used for interpreting deep learning models. SmoothGrad reduces noise in saliency maps by averaging gradients over perturbed inputs, while IG attributes feature importance by integrating gradients along a path from a baseline to the input. DeepLIFT compares activations to a reference state, assigning contributions to each feature. These methods are highly effective in neural networks due to their reliance on internal model details, but their model-specific nature limits their applicability. In contrast, LIME and SHAP offer model-agnostic solutions, making them more versatile for diverse machine learning architectures.

Critical Review

Both LIME and SHAP serve as powerful tools for interpreting machine learning models, each excelling in specific scenarios but also presenting notable trade-offs. SHAP, grounded in game theory, provides mathematically consistent and precise attributions, making it a theoretically robust choice for feature explanation. This is particularly evident in methods like TreeExplainer, which significantly optimizes SHAP's performance for tree-based models, achieving computational speeds up to 100 times faster than KernelExplainer. However, SHAP can still struggle with non-tree-based models like k-nearest neighbors, where its computational demands remain a bottleneck, especially for large datasets [12].

On the other hand, LIME stands out for its simplicity and speed, making it suitable for quick, localized explanations. By making sparse linear approximations, LIME offers a straightforward mechanism to interpret predictions in specific regions of the feature space. However, its reliance on sampling can lead to instability, with explanations varying significantly based on the data perturbation technique employed. This limitation is particularly pronounced in applications involving high-dimensional or imbalanced datasets [13].

From a practical standpoint, the choice between LIME and SHAP should align with the interpretability requirements of the application and the computational constraints at hand. While SHAP excels in providing both local and global insights, its computational overhead may limit its use in real-time or resource-constrained environments. LIME, despite its limitations, remains a viable option for scenarios where speed and simplicity are prioritized. Together, these tools underscore the trade-offs practitioners must navigate when selecting interpretability methods, with algorithmic context and application-specific goals guiding their decisions.

Implementation Relevance

The selection of LIME and SHAP for this study is strongly grounded in their demonstrated effectiveness in the literature and their ability to bridge theoretical and practical aspects of XAI. LIME has been widely recognized for its ability to generate interpretable explanations for individual predictions, making it an ideal starting point for exploring localized interpretability. By applying LIME to the Iris dataset, this study seeks to validate its practical utility in a controlled environment. The dataset's simplicity facilitates a focused evaluation of LIME's ability to demystify individual decisions in smaller, structured datasets.

This study's implementation is grounded in established literature that underscores the importance of interpretability in complex machine learning models. SHAP, as introduced by Lundberg and Lee, offers a robust framework for understanding feature contributions, particularly in high-dimensional datasets. By leveraging SHAP, this study aims to analyze the decision-making process of a deep neural network applied to the peripheral blood cell dataset. This dataset's complexity presents a practical testing ground for evaluating SHAP's scalability and utility in providing transparent and actionable insights. Grounding the implementation in these theoretical foundations ensures that the study aligns with proven methodologies while addressing real-world challenges in model interpretability.

5. Methodology and Implementation (Implementation Part)

Overview of the Chosen Method:
The method I am implementing for this task involves using explainability techniques LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) to interpret the decision-making process of a model, particularly in the context of image classification. These techniques are model-agnostic, meaning they can be applied to any machine learning model regardless of the underlying structure. The goal is to generate local explanations for individual predictions, highlighting the most influential features or pixels that the model relies on for its decisions.

In the case of image classification, this approach will allow us to understand which parts of the image contribute most to the model's prediction. For LIME, the method perturbs the input image and uses a simpler interpretable model (like a logistic regression) to approximate the predictions of the complex model. SHAP, on the other hand, provides a more mathematically rigorous approach to assign a contribution score to each pixel based on Shapley values.

Toy Problem/Dataset Choice:
To illustrate and test the application of Explainable AI (XAI) techniques, this study begins with the Iris dataset, a classic and well-known dataset in machine learning. The Iris dataset consists of 150 samples evenly distributed among three distinct species of iris flowers: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Each sample includes four features -sepal length, sepal width, petal length, and petal width- measured in centimeters. The dataset is low-dimensional and easy to visualize, making it an ideal starting point for experimenting with XAI techniques. By using this dataset, we aim to demonstrate the interpretability and scalability of methods like LIME and SHAP in a controlled and simple environment. This foundation facilitates transitioning to more complex datasets.

Subsequently, the study focuses on a more intricate dataset consisting of 17,092 images of peripheral blood cells. These images were captured using the CellaVision DM96 analyzer at the Hospital Clinic of Barcelona and are categorized into eight distinct classes, each representing a specific blood cell type. Unlike the Iris dataset, this dataset introduces a significant level of complexity due to the high dimensionality of the data (360x363 pixels per image in JPG format) and the inherent variability in blood cell appearances caused by type, condition, and stage. Such complexity necessitates advanced interpretability tools to achieve accurate classification and

reliable insights into model predictions. To ensure sufficient representation and address class imbalance, the dataset was augmented to 3,500 images per class using techniques such as rotation, flipping, cropping, and blurring

By first employing the Iris dataset as a toy problem, this study establishes a baseline for explainability and computational feasibility. Then, the transition to the blood cell image dataset demonstrates the scalability of XAI techniques and highlights their potential for tackling real-world challenges in medical diagnostics, particularly in the fields of hematology and automated disease detection.

Implementation Details:
The implementation of this study was carried out in two phases, reflecting the progression from a simpler toy problem (the Iris dataset) to a more complex real-world dataset (peripheral blood cell images). Python served as the primary programming language, leveraging TensorFlow, scikit-learn, LIME, and SHAP for model training and interpretability. All implementation code in [16].

Phase 1: Iris Dataset

The Iris dataset served as the starting point to test the Explainable AI (XAI) techniques on a smaller scale. A logistic regression model was implemented using scikit-learn, given the simplicity and low-dimensional nature of the data. The dataset was split into training and testing sets (80/20), and the model was evaluated using a confusion matrix. This provided an initial baseline for understanding model behavior.

To enhance interpretability, LIME and SHAP were applied:

- **LIME**: A tabular explainer was configured to generate feature importance explanations for individual predictions. These explanations highlighted the contribution of each feature to the model's prediction, as visualized in bar plots.
- **SHAP**: Kernel SHAP was employed to provide global and local explanations of the model's behavior. The SHAP force plot for a specific instance was saved as an interactive HTML file, demonstrating the influence of each feature on the predicted outcome. You can see an example of a SHAP Trial in figure 1.

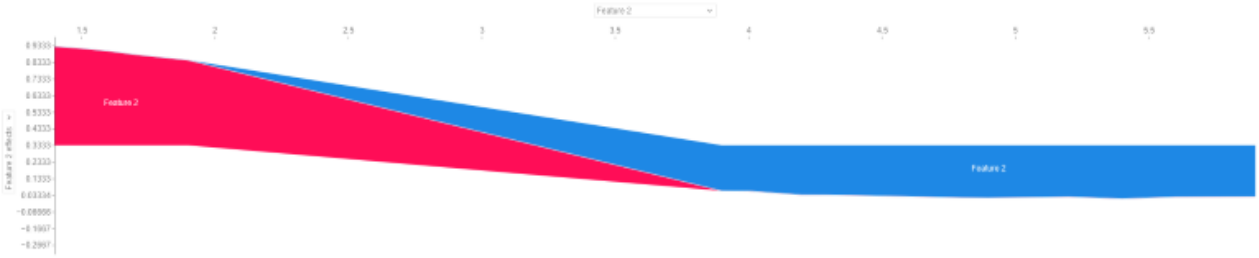

Figure 1: Example of a SHAP Trial

Phase 2: Peripheral Blood Cell Dataset

For this implementation, TensorFlow/Keras [17] was chosen due to its extensive pre-trained model repository and robust tools for transfer learning. Three different model architectures were employed: a simple CNN model with three convolutional layers, an advanced MobileNet-based model [14] utilizing depthwise separable convolutions, and InceptionV3, a deep pre-trained model initialized with ImageNet weights. The MobileNet model was specifically chosen for its efficient design, using depthwise separable convolutions that drastically reduce the number of parameters compared to standard convolutions. This makes it highly efficient for mobile and embedded systems while still providing strong performance on complex tasks.

A minimal amount of hyperparameter tuning was carried out, primarily focusing on adjusting learning rates (ranging from 0.001 to 0.0001), testing different optimizers (Adam and RMSprop), and fixing batch sizes at 32 due to memory constraints. Despite these adjustments, the models continued to exhibit poor performance on the test set. A major challenge was the computational overhead, as my computer don't has sufficient GPU memory.

For the analysis of model interpretability, both SHAP and LIME techniques were employed to explore and visualize the decisions made by the trained models, including the Simple CNN, MobileNet-based Advanced model, and the InceptionV3 model. SHAP was applied using a Partition explainer. For each model 15 images were used as background in SHAP, and the results were visualized using SHAP's image plotting functionality.

Similarly, LIME was applied using the LimeImageExplainer, which generates local explanations by approximating the model with a locally interpretable surrogate model. For each model, explanations were generated for selected test images, and the regions of the images that were most influential in the predictions were highlighted.

6. Results and Discussion

The models struggled to learn effectively, likely due to overfitting. The implementation of interpretability methods, LIME and SHAP, yielded differing levels of success. LIME performed as expected, providing localized and comprehensible explanations for individual predictions, even when the underlying model faced difficulties in learning effectively. In contrast, SHAP encountered significant challenges. Due to computational limitations, the size of the background dataset used for SHAP had to be kept low. This constraint likely contributed to SHAP failing to produce meaningful results; in all cases, it returned outputs that were essentially uninformative. The poor performance of the classification models (%12.5~ test accuracy) further exacerbated this issue. Since SHAP relies on the quality of the model's predictions to generate insights, the lack of effective learning in the model led to unsatisfactory explanations. On the other hand, LIME, being a local interpretability method, was more robust in this context, offering reasonable insights even under suboptimal conditions. To demonstrate a functional SHAP explanation, a simplified training setup was employed. This involved training the model on a much less complex problem, allowing SHAP to generate interpretable results. Thanks to this test I fixed my SHAP implementation.

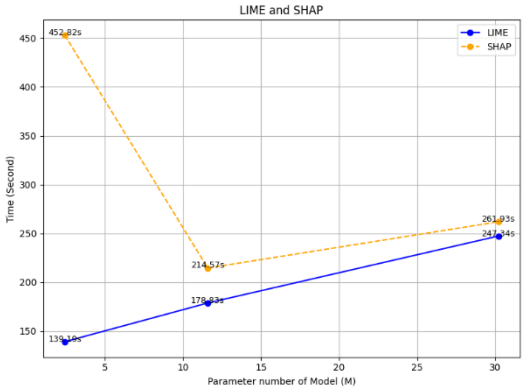

Figure 2 Time comparison of methods

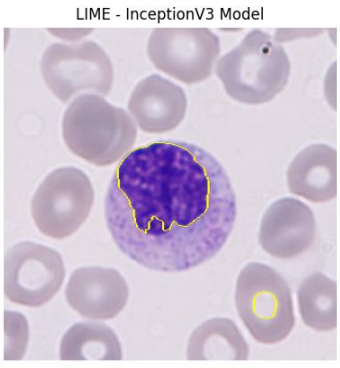

Figure 3: Example of LIME result

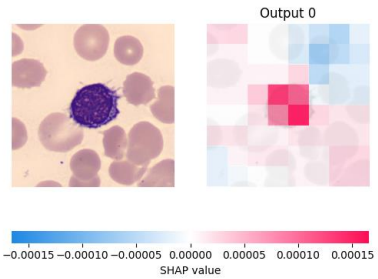

Figure 4: Example of SHAP result

The results showed that even when the model failed to learn effectively, LIME was still able to provide successful explanations. In contrast, SHAP's ability to generate meaningful interpretations was much more dependent on the model's success and required significantly higher computational resources. When considering

the time taken for these processes, it became apparent that the number of parameters in the models had a relatively small impact on the overall computation time for both LIME and SHAP.

7. Challenges and Open Problems

Challenges in Literature: The literature on XAI techniques, such as LIME and SHAP, identifies several key challenges that need to be addressed. One significant challenge is the computational load and power requirements. Achieving meaningful results with these methods often demands substantial computational resources, which can become a barrier in real-world applications, especially when working with large datasets or complex models. Another challenge is the understandability of explanations; as the complexity of models increases, the clarity and effectiveness of the explanations generated by these techniques tend to diminish. This can result in insights that are harder to interpret and act upon, ultimately reducing the usefulness of the results. A further difficulty lies in providing global explanations. While local explanations can be generated for individual predictions, offering a comprehensive understanding of the model's overall behavior remains a significant hurdle, particularly for more complex tasks where understanding the entire decision-making process can be challenging.

Challenges in Your Implementation: In the implementation of this study, several challenges arose, mainly due to the size and complexity of the dataset. The high-dimensional nature of peripheral blood cell images, combined with the diversity of cell types, posed difficulties for the models in terms of learning and generalization. Additionally, GPU memory limitations exacerbated the training process, slowing down both model development and interpretability analysis. While LIME and SHAP were effectively used to provide insights into the models' decisions, the underlying performance issues stemming from the training process diminished the meaningfulness of the results.

8. Conclusion

Summary of Findings: From both the literature review and the implementation, it became clear that LIME and SHAP are valuable methods for improving model transparency. However, their effectiveness diminishes when applied to more complex and high-dimensional datasets, such as peripheral blood cell images. While these techniques show promise in simpler tasks, their application to real-world datasets revealed significant challenges. The explanations generated by these methods, although understandable, were not sufficiently actionable in this case. This was not due to limitations of LIME and SHAP but rather because the models did not provide meaningful or accurate predictions due to issues with the training process. But as a result of research of papers for this survey I can say that the need for further optimization and refinement of XAI techniques for real-world applications becomes evident, especially when dealing with complex, high-dimensional data.

Link to Future Work: Future research should focus on developing more efficient and scalable XAI techniques that can handle the challenges posed by high-dimensional and complex datasets. Improving the robustness of methods like LIME and SHAP for real-world applications, particularly in fields like healthcare and image classification, would be valuable. Additionally, exploring new approaches for model interpretability, such as combining multiple XAI techniques or developing novel methods tailored to complex data, could help bridge the gap between theoretical models and practical applications. Addressing computational limitations through more efficient algorithms and hardware optimizations would also be critical in making these techniques more accessible and effective for large-scale use. And also there is currently no standard for assessing these methods. Standardization issues need to be resolved for this field to progress.

References

[1] Acevedo, A., Alférez, S., Merino, A., Puigví, L., & Rodellar, J. (2019). Recognition of peripheral blood cell images using convolutional neural networks. *Computer Methods and Programs in Biomedicine*, *180*, 105020. https://doi.org/10.1016/j.cmpb.2019.105020

[2] Swartout, W., Paris, C., & Moore, J. (1991). Explanations in knowledge systems: design for explainable expert systems. *IEEE Expert*, *6*(3), 58–64. https://doi.org/10.1109/64.87686

[3] Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2020). A historical perspective of explainable Artificial Intelligence. WIREs Data Mining and Knowledge Discovery, 11(1). https://doi.org/10.1002/widm.1391

[4] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Ser, J. D., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99(101805), 101805. sciencedirect. https://doi.org/10.1016/j.inffus.2023.101805

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv.org. https://arxiv.org/abs/1602.04938

[6] Garreau, D., & Von Luxburg, U. (n.d.). *Explaining the Explainer: A First Theoretical Analysis of LIME*. http://proceedings.mlr.press/v108/garreau20a/garreau20a.pdf

[7] Ren, & Malik. (2003). Learning a classification model for segmentation. *International Conference on Computer Vision*. https://doi.org/10.1109/iccv.2003.1238308

[8] Laugel, T., Renard, X., Lesot, M.-J., Marsala, C., & Detyniecki, M. (2018). *Defining Locality for Surrogates in Post-hoc Interpretablity*. ArXiv.org. https://arxiv.org/abs/1806.07498

[9] Visani, G., Bagli, E., & Chesani, F. (2020, June 10). *OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms*. ArXiv.org. https://arxiv.org/abs/2006.05714

[10] Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*. https://arxiv.org/abs/1705.07874

[11] Shapley, L. (1953) A Value for n-Person Games. In Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317. - References - Scientific Research Publishing. (2017). Scirp.org. https://www.scirp.org/reference/referencespapers?referenceid=2126587

[12] Md. Mahmudul Hasan. (2024). Understanding Model Predictions: A Comparative Analysis of SHAP and LIME on Various ML Algorithms. *Journal of Scientific and Technological Research*, *5*(1), 17–26. https://doi.org/10.59738/jstr.v5i1.23(17-26).eaqr5800

[13] Chen, Y. Z., Calabrese, R., & Belen Martin-Barragan. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research*, *312*(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

[14] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Andreetto, M. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. https://arxiv.org/pdf/1704.04861

[15] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *Information Fusion*, *106*, 102301. https://doi.org/10.1016/j.inffus.2024.102301

[16] b21945815. (2024). *GitHub - b21945815/Explainable-AI-Techniques: ODTU 562 Machine Learning Course Project*. GitHub. https://github.com/b21945815/Explainable-AI-Techniques

[17] *Tutorials | TensorFlow Core*. (n.d.). TensorFlow. https://www.tensorflow.org/tutorials