Enhancing Interpretability of Convolutional Neural Networks: A Decision Tree Approach Using Grad-CAM and Wavelet Transforms for Multi-Class Classification

Name: Jeevan Umesha

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May-December 2024

APPROVED:

Associate Professor Russel Pears

Teaching Professor Renee Bryce

Assistant Professor Yunhe Feng

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Russel Pears, for their invaluable guidance, support, and encouragement throughout the entire process of my research. Their expertise, insight, and constructive feedback have been instrumental in shaping the direction of this thesis, and I am deeply grateful for the time and effort they dedicated to helping me succeed.

I would also like to thank the members of my thesis committee, Professor Renee Bryce, and Professor Yunhe Feng, for their insightful suggestions and recommendations. Their diverse perspectives and expertise have significantly enhanced the quality of my work.

A special thanks to department of Computer Science and Engineering (CSE) for providing the resources and facilities necessary to conduct my research. The access to advanced computational tools and research materials was crucial to the success of this project, and I am grateful for the opportunities the institution has provided me.

I am also grateful to my fellow researchers and colleagues in the Department of CSE for their collaborative spirit and for sharing their ideas and experiences. The academic environment at University of North Texas (UNT) has been a source of motivation and inspiration, and I appreciate the knowledge exchange that has occurred throughout the course of my studies.

I would like to acknowledge the support of my family and friends, especially my parents and a special friend, for their unwavering encouragement, understanding, and patience. Their love and support have been a constant source of strength during both the challenges and successes of my academic journey.

Lastly, I extend my thanks to all the individuals and researchers whose work has contributed to the foundation of this thesis. The knowledge and resources provided by the academic community have been invaluable, and I am truly appreciative of their contributions to my research.

This thesis would not have been possible without the support and guidance of all those mentioned above. I am deeply thankful to each one of you for helping me reach this point in my academic career.

Jeevan Umesha

11/15/2024

Table of Contents

Introduction.....7

1.1Background and Motivation.....7

1.2 Research Overview and Objectives.....8

Literature Review.....10

2.1 Introduction to Explainable AI (XAI).....10

2.2 Seminal Contributions to XAI.....12

2.2.1 Rule Extraction from Neural Networks.....12

2.2.2 Deconvolutional Networks and Visualization.....14

2.3. Modern Advancements in CNN Explainability.....16

2.3.1 Gradient-based Methods.....16

2.3.2 Layer-wise Relevance Propagation (LRP).....18

2.3.3 Concept-based Explanations.....19

2.4. Hybrid Methods and Symbolic Interpretability.....19

2.5. Applications of Explainability in CNNs.....22

2.5.1 Medical Imaging.....22

2.5.2 Autonomous Systems.....22

2.5.3 Multi-modal Applications.....23

2.6. Gaps and Challenges.....23

2.7. Future Directions.....25

Methodology.....27

1. Data Preparation.....27

2. Choice of Model: VGG16.....27

3. Dataset Selection.....28

4. Feature Map Extraction and Grad-CAM Heatmap Generation.....29

5. Decision Tree Construction.....30

6. Future Work.....31

Results.....33

Discussion.....54

Conclusion.....64

References.....69

LIST OF FIGURES

Page

Figure 1: Examples of the dataset……**27**

Figure 2: Examples of Feature Maps Generated**29**

Figure 3: Examples of HeatMaps Generated using Grad-CAM**30**

Figure 4: Feature Map Extraction: Contour Detection and Calculation**33**

Figure 5: Decision Tree-1……**34**

Figure 6: Decision Tree-2……**38**

Figure 7: Decision Tree-3……**43**

Figure 8: Graph depicting training accuracy, validation accuracy**48**

Figure 9: Decision Tree with Grad-Cam features of CelebA……**50**

Figure 10: Decision Tree with Feature-Map features of CelebA**50**

Figure 11: Examples of denoising images with the Wavelet transform**51**

Figure 12: Training/ Validation accuracy after Wavelet transform**52**

LIST OF ABBREVIATIONS

AI	**Artificial Intelligence**
ML	**Machine Learning**
CNN	**Convolutional Neural Networks**
CRP	**Concept Relevance Propagation**
FERNN	**Fast Extraction of Rules from Neural Networks**
LRP	**Layer-wise Relevance Propagation**
MLP	**Multilayer Perceptron**
OHE	**One Hot Encoding**
RICE	**Rotation Invariant Contour Extraction**
VGG	**Visual Geometry Group**
XAI	**Explainable Artificial Intelligence**
ILSVRC	**ImageNet Large Scale Visual Recognition Challenge**
Grad-CAM	**Gradient-weighted Class Activation Mapping**
SHAP	**SHapley Additive exPlanations**
LIME	**Local Interpretable Model-agnostic Explanations**
RxDNN	**Rule Extraction for Deep Neural Networks**
TCAV	**Testing with Concept Activation Vectors**

Abstract

This research extends previous work on interpreting Convolutional Neural Networks (CNNs) to apply improved techniques to understand the model's ability to make decisions in multi-class classification tasks. Classification by CNNs was explained in the original study using feature map extraction and decision trees. In this work, a technique known as Grad-CAM (Gradient-weighted Class Activation Mapping) is used to obtain heat maps that illustrate regions in the image responsible for the CNN class distinction. To increase the readability of these heatmaps, wavelet transforms are used to denoise the cropped images prior to the use of Grad-CAM. It is then followed for construction of a decision tree by extracting features; area, perimeter, and aspect ratio from heatmaps. It lies in this approach of offering insightful understanding of the CNN's decision and its rationale for categorizing image data into six different classes. Here I attempt to provide some additions to the current body of model interpretability research by focusing on the behavior of CNNs and by laying groundwork for more practical and effective methods of explaining intricate machine learning.

Chapter 1

Introduction

1.1Background and Motivation

Artificial Intelligence (AI) has undergone rapid growth in the current years thus leading to massive changes in several fields of the economy. Known as a fanciful idea in 1950s by Russell and Norvig (2009), AI has become an established technology evident in almost every domain of human endeavour, including health, finance, entertainment, among countless other areas of application. Indisputably one of the most striking improvements has been the emergence of the so-called large language models that have recently been renowned for their capacity to write humanlike texts, translate between languages, and provide the summary of the article. These models have gone a long way in improving industries, they are of aid to professionals and the general public alike, as argued by Vaswani et al. (2017).

The usage of AI to solve natural language processing problems is already relatively well understood, but it has had nearly as much of a revolutionizing effect on image processing activities. Current state-of-art algorithms like Convolutional Neural Networks (CNNs) design solutions for many problems, including image recognition, object detection, segmentation, and tracking (LeCun et al., 1998). CNNs have become widely used due to the availability of powerful, purpose-built hardware such as GPUs, and image processing has achieved great heights compared to more conventional methods of machine learning. Several architectures of CNN including AlexNet (Krizhevsky et al., 2012), VGG (Simonyan & Zisserman, 2015) and ResNet (He et al., 2016) have achieved impressionistic performances on a number of tasks like ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and the errors rates have been reduced to an extent that the CNN has become the standard method in this In the last decade the deep neural architecture has grown more complex as they incorporated deeper networks and adopted better weight initialization and regularization methods. These improvements have made it possible for the CNNs to improve their levels of accuracy thereby solving complex challenges in different image related functions (He et al., 2016; Szegedy et al., 2015). However, as these models have evolved, they have become black box models and their decision making processes that lead to specific predictions are not fully clear to the end user (Ribeiro et al., 2016).

These models have become opaque, and researchers and organizations have now started raising questions on whether the outcome of the models is actually interpretable or not, especially in industries like healthcare, security, finance and justice. Although CNNs might yield superior accuracy, a major problem with them is that they are not easily understandable, which is why user trust and consequently, adoption are affected. For instance, while the deep learning model could deliver satisfactory results such as a diagnosis, the need to know why the given deep learning model came up with the particular recommendation is equally important to guarantee that such suggestions are acceptable and reliable to the health professionals (Doshi-Velez & Kim, 2017). This is especially important in the cases where the implemented AI systems are to be applied in the fields where people's lives, security and

monetary value are involved. To mitigate these challenges, there has risen what is known as Explainable Artificial Intelligence (XAI). XAI defines itself as that which aims to increase the trust and credibility of AI through providing ways of understanding the rationale behind execution of intricate models by users, as stated by Gilpin et al., 2018). However, with simple architectures like MLPs, the requirement for model interpretability was raised early and efforts have only recently been aimed at CNNs for image classification tasks (Bau et al., 2017). Making techniques by which specific decision-making of CNN is interpreted has emerged as a major research area, and approaches such as Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), and Saliency Maps (Simonyan et al., 2013) are adopted for developing visualizations. It is utilized to highlight which part of an image has the most impact to the CNN in its classification decision. However, the vast majority of them still employ feature map extraction and hence they can be considerably time consuming; besides, their decision making process is often not very clear. Newer techniques like the Gradient-weighted Class Activation Mapping (Grad-CAM) have worked on this front, to make it more geometrically feasible by coming up with heatmaps that illustrate the part of the image that contributes to the CNN decision making. It has been referred to as Grad-CAM because the technique relies on gradients of the class score with respect to the feature maps of the final convolutional layer Within a short time, Grad-CAM became prominent primarily because of its capacity to offer human interpretable explanations of CNN decisions.

Nonetheless, one weakness of the current approaches can be summarised in that Grad-CAM often yields heatmaps of low quality and such resolution may not represent what the model 'sees' clearly. Attempts to overcome this have led to the development of smoothing techniques including the wavelet transforms to improve the quality of such heatmaps in eliminating noise in order to improve interpretability (Gonzalez and Woods, 2002). Hence, this research seeks to enhance the Grad-CAM technique by applying wavelet transforms so that the heatmaps developed can be comprehensible to the users of the model. The rationale for this study stems from increased demand for ways to better explain results from 'black box' classifiers like the CNNs. Although CNNs have shown great success in image classification, they are fatalistically opaque, and this is a strong barrier in gaining trust and ubiquitous use, especially in safety-sensitive activities. In this research, the use of Grad-CAM and wavelet transforms is aimed to propose further improvements in CNN explanation, as well as to advance Explainable AI as a field. Furthermore, this work strives to further improve interpretability in multi-class image classification problems, and contribute to opening the black box of CNNs and other high-performance AI algorithms used in today's world.

1.2 Research Overview and Objectives

Research Overview

It is a fact that nowadays AI has grown very fast and, in many fields, especially in areas where they require a high volume of data such as image processing. Specifically, Convolutional Neural Networks (CNNs) play a critical role in applications such as object detection, object recognition and segmentation, and the degree of accuracy is generally very high. Nonetheless, CNNss, while they are widely recognised as highly effective, they are accused of being 'black box' models. Due to this problem, the emergence of the field of Explainable AI (XAI) that focuses

on the gap between high efficiency and understandability to explain how the models make their predictions.

This work intends to shed light on the primary classification methods used in CNN filters through visualization of model saliency using Grad-CAM. This approach also increases interpretability through the generation of heatmaps that draw attention to parts of an image that influence CNN predictions. Further, wavelet transforms are used on these images before heatmap is produced to get better and smoother heatmap presenting more meaningful results. For analysis based on Grad-CAM heatmaps, the area, perimeter, and aspect ratio of the bounding box that encapsulates the heatmaps are selected and incorporated to build decision trees for classes that reflect the flow of combined CNN's decision. This research also investigates how these interpretive techniques help improve the understanding and efficiency of the CNN classification, creating a framework that may prove useful to fields which heavily rely on the trustworthiness of AI predictions.

Research Objectives

1. **To analyze and understand the decision-making process of Convolutional Neural Networks (CNNs):** This objective is to understand CNN classification behaviour through the process of Grad-CAM visualisation in an attempt to interpret model outputs.

2. **To enhance interpretability in CNN-based classification using Grad-CAM and wavelet transforms:** This work aims at ensuring that feature representations generated by Grad-CAM heatmaps are visually clearer and interpretable by applying wavelet transforms to images before the heatmaps' generation.

3. **To extract specific interpretative features from Grad-CAM heatmaps and construct decision trees:** This objective consists of analysing Grad-CAM heatmaps to estimate numerical features like area, perimeter and aspect ratio and then construct decision trees based on these numerical features in order to understand how CNN classifications are made.

4. **To assess the applicability of Grad-CAM and decision tree methods for interpretability in complex classification tasks across multiple classes:** The research will also measure and compare the two methods for multi-class problems, and offer some idea about their applicability and generality.

5. **To contribute to the broader field of Explainable AI (XAI) by offering a methodology that enhances trust and transparency in AI systems:** This is to achieve the objective of XAI by showing how interpretative methods can make deep learning models better explainable especially in the areas where interpretability is essential.

This research seeks to contribute an invaluable framework for practice that goes beyond simple adherence to technical precision, to engender public trust in AI systems by revealing the thought processes behind the resultant forecasts.

Chapter – 2

Literature Review

2.1 Introduction to Explainable AI (XAI)

XAI is a novel and relevant subdiscipline of machine learning and is aimed at responding to the problem of black box models, which is characteristic of deep learning. Despite the recent advances in the effectiveness of these models in several application domains, they are usually highly sophisticated, and consequently, their mechanisms are often referred to as 'black boxes'. This lack of interpretability hinders their deployment in critical real-world environments including the health sector where diagnostic choices need to be comprehensible; the financial sector, where control and standards compliance is important; and in self-governing systems where the ability to understand system behaviour is critical in ensuring safety and reliability. The goal of XAI, therefore, is to help demystify these models while maintaining their credibility, ensuring end-users, regulators, and stakeholders can understand, verify, and effectively utilize the results of these systems. To achieve this, XAI techniques are broadly divided into two main categories: intrinsic and extrinsic methods. Intrinsic methods focus on designing models that are inherently explainable, including decision trees, linear models, or rule-based systems, simplifying interpretation and fostering trust. These models are more rudimentary and less obscure in exchange for a lack of efficiency in certain complicated applications. Conversely, post hoc methods are confined to creating explanation techniques once the models are built and are highly complex. These techniques are ideally suited to the case of neural networks and ensemble models since they enable the discovery of new insights without the need to change the topology. Another way to categorise XAI methods relies on the division into local and global explanations. Local techniques are

concerned with reporting out a model's forecast and give specific information why a specific choice was made for a single instance. Some of the local techniques under consideration include SHAP and LIME which provide feature importance scores to explain individual predictions. On the other hand, global methods seek to give a global picture of how a model operates, motifs, features and decisions irrespective of a specific, localized data region. These heuristics are beneficial when a generalizable approach to its decision-making process is of interest or when there may be a certain consideration, which is suspicion to bias in a potential solution.

This literature review identifies the major contributions that form the basis of the XAI space with specific consideration to the progress made in explaining neural networks especially the Convolutional Neural Networks (CNNs). CNNs, standard in computer vision tasks, are inherently difficult to interpret because of their layered feature maps and feature vectors of potentially large dimensionality. It touches on areas such as Grad-CAM and saliency maps that connects high level abstract features and human interpretable visualization. Moreover, some advances have been made on the integration of both visual and logically symbolic interpretability, so the explanations generated are at once visually interpretative and logically structured and thus both types of users are equally served, technical and non-technical. Due to these considerations, XAI seeks to overcome the primary drawback of model interpretability making machine learning systems more usable in real world scenarios. Going forward, the methods have gradually shifted from ones that provide high-performance but are opaque, to ones that can be at least partially understood and measured but are as effective as their opaque counterparts while not compromising the core decision making. In this review, an endeavour of integrating state-of-art concepts and recent developments in XAI, and

providing a broad vision is attempted; focusing on making neural networks, or specifically CNNs, more interpretable and interactive for broader use.

2.2 Seminal Contributions to XAI

2.2.1 Rule Extraction from Neural Networks

XAI as a research discipline that first emerged as researchers sought for rule extraction techniques to be applied mainly in shallow neural networks. These methods were employed for the translation of the so-called “black box” of neural networks into forms that are more comprehensible, for example, decision trees or sets of if-then rules that would help to explain how the decision or prediction was made. In implementing these ideas, one of the key seminal works in this area was the Fast Extraction of Rules from Neural Network . Originally, FERNN was developed to derive decision trees and investigate the activations of the hidden neurons as the principles for decision making in the model. By combining information gain and cross entropy loss function with a penalty function FERNN was able to make sure that only relevant neurons contributed to rule generation. By doing so, the approach optimally removed irrelevant features and boiled the extracted rules down to the minimum number in order to enhance readability. An impressive performance by FERNN was noted when used in tabular datasets where the features are discrete and there existing logical and/or simple, discrete relationship between the variables in the form of decision trees. However, the method encountered some challenges that reduce the suitability for wider application. In this respect, one restriction was the ability to work effectively with continuous features which are a common occurrence in most real-world datasets. The FERNN model cannot incorporate continuous features without first being discretized which, despite the use of refined discretization techniques, reduces the quantity and quality of features that the model can

effectively employ. Moreover, FERNN was not easily scalable to other complex tasks such as image classification or natural language processing due to the high dimensionality of the features as well as complex interdependencies between features that made the rules extraction process less efficient. However, follow-up research introduced superior methods to overcome these issues. One such method was the pedagogical rule extraction method called TREPAN that was created by Craven and Shavlik in 1996. TREPAN helped the progress of the field by reproducing hierarchical sets of rules that were very close in decision space to the neural network in question. While TREPAN used a querying and sampling strategy systematically to develop rule-based models which remained as faithful as possible to the initial model's patterns, FERNN did not have such regulating norms. As the result, this method filled the gap between transparency and accuracy and subsequently should be considered a landmark method in realizing XAI.

Even though TREPAN was developed to overcome shortcomings associated with earlier methods, it, too, had its own difficulties, the most prominent of them relating to scalability and computational complexity. Although it could query and sample through space and excel in small to medium data sized environments, querying and sampling made a considerable bottleneck in high dimensional or big data. Querying is the use of synthetic data points to map the decision boundaries of the model, something often costly and a logistical nightmare especially in deep learning models with millions of parameters. However, unlike some other methodologies, rules generated by TREPAN are hierarchical and interpretable, and while the method is claimed to asymptotically compress models it suffers from a drawback of being unable to mine highly dense and complex data such as image or video. These issues outline the progress made in the development rule extraction methods as well as future work required in the field of XAI. While the first methods like FERNN are conceptually connected to

interpretable neural networks they are limited by scalability and only support training on low complexity data. Techniques such as TREPAN generalized these ideas for more complex environments, yet at the same time, such methods are beset by high computational demands. Closely related to the scale of the current and future problems, the demand for exact, fast, and understandable methods continues to be relevant, inducing search for new synergistic approaches based on the rules of explanation and the possibilities of the latest deep learning methodologies. These developments suggest that progress towards functionally efficient AI can be made while at the same time meeting a basic requirement of any high risk, high impact application; the ability to explain its actions.

2.2.2 Deconvolutional Networks and Visualization

Matthew D. Zeiler and Rob Fergus, in their work published in 2014, they proposed what they called deconvnets, as an intuitive approach for decoding the internal processes of Convolutional Neural Nets (CNNs). Deconvnets in fact involve the process of mapping activations back to the original pixel space by convolving the layer's feature maps with the corresponding weights and adding them through element wise nonlinearities. Here, deconvnets allowed to identify how these CNNs extracted and located key features in the images as well as offer a glimpse into a tiered feature learning process that is characteristic of these models and which are essential in applications such as image classification and object detection. This was further enhanced by systematic occlusion experiments where sections of an image were covered to investigate how the output of the network was impacted, leading to identification of regions that had the most effect on the model. Deconvnets were among the first attempts to address the problems of feature visualization and served as groundwork for methods intended for interpreting CNNs. They showed how CNNs could pinpoint the areas

of interest and provide valuable features, which provided corroboration to their usefulness in the spatial domain. Nevertheless, there were limitations to deconvnets which are stated as follows. The heatmaps that were generated often had to be visually analysed and interpreted, thus negating the whole point of the system: objectivity. As much as they pointed out areas of relevance, they did not offer an extended description of the role of these areas in particular conclusions. This absence of operational interpretability hampers their applicability in critical applications that require not only results but the actual reasoning process behind them. Equally, approaches such as Activation Maximization (Erhan et al., 2009) provided another pathway to feature visualization in which gradient ascent was employed to produce fake images that would excite specific neurons or layers of the CNN. This technique generated images lying on a hyperplane, illustrating the patterns or textures to which a particular neuron is most sensitive. These features represent the general characteristics learned by the neuron. For example, neurons at the lower level can be responsible for activation in edges or colors, those at a higher level for textures or even object parts. Such observations facilitated understanding of the fact that in CNNs, basic filters come together to create more complex features.

However, like many methods that paved the way for modern views of Neural Networks, Activation Maximisation and similar methods were not without their drawbacks. The synthetic images generated were frequently not intuitive and challenging to understand for the human brain, particularly for new layers of neurons that detect such sophisticated and counterintuitive qualities. Moreover, these methods did not describe the contribution of learned features to the specific prediction or a decision that will be made in a particular network. For instance, they could inform that neuron was receptive to a certain texture

however, they could not explain how the response of that neuron contributed to the classification of an image.

In summary, deconvnets and Activation Maximization were big breakthroughs in feature visualization and in understanding of what neural networks are seeing. They served as a starting point of interpreting how CNNs learn and recognize image-related information empowering the development of better methods. However, the completely human interpretation and the lack of the link between the visualised features and precise decision-making demonstrated a serious drawback of limiting the connection between visualisations and the decision-making process. This limitation eventually led to the creation of new methods which combine heatmap-based visualization with decision attribution specifically, Grad-CAM (Selvaraju et al., 2017). All these improvements further extend the work done by Zeiler et al. and Erhan et al., and keep the progress of the field of XAI advancing, as the latter aims at enhancing the transparency of machine learning algorithms and making them more reliable.

2.3. Modern Advancements in CNN Explainability

2.3.1 Gradient-based Methods

Grad-CAM, developed by Selvaraju et al. (2017), is one of the most popular and impactful of the post-hoc explainability techniques aimed at providing visualization of CNNs. Grad-CAM then produces heatmaps that show the dimensions of any input image that contributes to a model's prediction. It does this by determining the gradients of the target class score with respect to the feature maps in the last convolutional layer. Such gradients are weighted and summed up with feature maps in order to generate a class-discriminative localization map on the input image. This approach helps to establish the relationship between the decision-

making process of the model and specific areas in the input; thus, the Grad-CAM tool is applicable in numerous domains for explaining predictions. Grad-CAM has been widely used across multiple domains due to its flexibility including for identifying the part of the image the model is paying attention to in medical imaging that can assist clinicians in understanding AI decision making processes regarding classifiers and detectors of images of tumours or lesions. For object detection, Grad-CAM is employed in the area that contributed to the detection of specific objects to improve the detection algorithms. In the same manner in an autonomous vehicle, it uncovers the decision-making rational involved in visions system and aids in safety since it demonstrates the images the system uses to perceiving its environment.

Expanding from the Grad-CAM method, Score-CAM which was developed by Wang et al. (2020), was a mitigation to one of the primary limitations of Grad-CAM: gradients are notorious for their noisy or unstable behaviour at times. Score-CAM completely does away with gradients, using instead the forms of activation maps alongside scores of relevance estimated from the model's output. This method operates on the concept that by masking different regions of the input image using activation maps and passing these masked inputs back into the model to measure the impact on the score we get for the prediction. Subsequently the resulting relevance scores are utilized to construct a heatmap that indicates regions of interest in the input. SCORE-CAM avoids gradients and hence solves two problems - the primary one of interpretability and the secondary one of robustness in the presence of noise and other inconsistencies in gradient calculation. This enhancement makes it useful particularly in situations where gradients of models are problematic, for example when models trained with noisy or imbalanced data. Another extension that also deserves mention is the Smooth Grad-CAM++ developed by Omeiza (2019) as an improvement on Grad-CAM++ to improve the heuristic quality of heatmaps. Grad-CAM++ refines Grad-CAM by using

smooth, which takes a number of heatmaps derived from images slightly containing noise. This smoothing process decreases the amount of noise in the individual resultant heatmaps and also improves on the appearance of the final explanation given to the human user as a heatmap. The integration of the strengths displayed by gradient-weighted techniques with noise reduction make the use of Smooth Grad-CAM++ effective in highly sensitive areas such as diagnoses of ailments, and decisions making processes. Grad-CAM, Score-CAM, and Smooth Grad-CAM++ can be considered methodological milestones within the XAI field, especially for understanding visual interpretability of CNNs. Grad-CAM has been demonstrated to be an effective saliency highlighting method which is relatively simple to implement; moreover, its derived variants overcome some shortcomings and enhance its functionality. Taken together, these methods belong to the growing family of interpretability techniques, where researchers and practitioners can use to enhance the understanding of the activities of AI models in many domains. Insofar as explainability is a significant issue across high importance value chains, such methods will persist in developing over time as novel means of imparting transparency and reconcilement into advanced machine learning frameworks.

2.3.2 Layer-wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) distributes the prediction score backward through the network layers, assigning relevance to individual input pixels. Unlike gradient-based methods, LRP preserves relevance proportionality, making it suitable for applications requiring precise attributions, such as healthcare diagnostics.

Extensions to LRP include Deep Taylor Decomposition (Montavon et al., 2017), which uses Taylor series approximations to improve relevance attribution. These methods have proven

effective in explaining predictions in image classification and natural language processing (NLP).

2.3.3 Concept-based Explanations

Kim, Y., Wulf, E., Saeed, Z., Zophikian, A., Chung, J. Y., Vanhoucke, V., & Dean, J. (2018) presented Testing with Concept Activation Vectors (TCAV), a new technique for measuring a high-level human-interpretable concept (e.g., stripes, textures) contribution to an explanation. Particularly, the TCAV leverages linear classifiers to assess the sensitivity of neurons to certain concepts and, thus, explain CNN reasoning. The use of concept-based explanations can be split into two more recent idea: Concept Bottleneck Models (Koh et al., 2020) where the models are trained to predict intermediate concepts before the final decision. It improves both interpretability and generalization because it imposes modularity on the model building process.

2.4. Hybrid Methods and Symbolic Interpretability

Combined approaches in Explainable Artificial Intelligence (XAI) combine visualization and constraint-based reasoning to eliminate shortcomings of separate approaches, namely the interpretability of generated visualizations or the lack of flexibility of the rule-based approach. These approaches sought to offer natural language descriptions for interpretation, as well as being able to present the information in a format that can be easily understood by rule-based algorithms, such as CNNs.

An example of how it can be done is illustrated by integrating Grad-CAM with symbolic methods. Grad-CAM produces heatmaps that highlight the areas of input that images most contribute to a model's decision-making process. These heatmaps can be quantitatively analysed to extract geometric features like area, perimeters and aspect ratios, which

inherently defines spatial properties of important regions. These extracted features are then used in building other easy to explain type of models such as decision tress which mimic the decision making process of the CNN. Indeed, the presented two-fold approach combines low-level heatmaps for the fine-grained visualization of interpretability metrics in specific areas of the input space, and high-level decision trees for the global, symbolic representation of the interpretability model, which can meet multiple interpretability requirements at the same time. Such hybrid techniques have been more useful in fields like medical imaging where it helps the clinician in interpreting the AI based diagnostic indications (Rahaman et al., 2020 ; Selvaraju et al., 2017).

Another powerful set of tools referred in hybrid approaches is the Wavelet transform that helps to improve the clarity and readability of a heatmap. Wavelets are mathematical representations of images which break down an image into frequency entities and filter out noise and enhance important features such as edges. Combining methods with Grad-CAM-based heatmaps as input, hybrid approaches filter a heatmap more efficiently by denoising and highlighting the patterns of interest when the heatmaps are pre-processed with wavelet transforms. For instance, Zhou et al. (2021) showed that pre-processing heatmaps through wavelets enhanced algorithms' interpretability of multi-class image classifiers by bridging the feature relevance with pixel-level knowledge in form of rules. It is this technique that moves pixel-level information which may be difficult to parse by symbol-processing mechanisms to a format that is easier to produce human interpretable patters and rules.

Another method worthy of mention is RxDNN introduced by Zilke et al. (2016). RxDNN integrates decision tree induction and activation clustering to obtain rule sets that are fairly close to, concise and interpretable while remaining faithful to the original deep neural

network. Activation clustering aggregates neurons with similar activity modes for prior simplification of the internal structure of the model before inducing decision trees which define its decision surfaces. These rules are discerned in terms of if-then account, which allows thinking of the formalism as a symbolic description of model's reasoning for humans to grasp. There has been evidence of how RxDNN has worked on some of tasks such as the image classification and fraud detection to generate faithful rules while keeping interpretability an important facet for decision-critical uses (Guidotti et al., 2019; Zilke et al., 2016).

Hybrid methods also go beyond the combination of visually based rules and include statistical and mathematical tools to improve interpretation. For instance, the LIME-RULES (Ribeiro et al., 2016) approach builds upon LIME, which uses LIME components that generate simple and locally faithful models to provide rule-based interpretations of the predicted value for an instance. Almirah Reddy and Kavita Gopal (2019) developed another method called as DeepRED which identifies symbolic rules by breaking down deep models into subnets and investigated each of them separately. These approaches help in guarantees that the explanations are accurate and that the solution can be driven in real time within a computation scale.

In addition, there is an open area of research in the combination of fuzzy logic with heatmap-based approaches as the unique implementation of hybrid XAI. Guaranteeing interpretability, fuzzy logic can take advantage of the uncertainty and ambiguity inherent in model results and convert heatmaps into fuzzy rules which offer graded explanation. For instance, Kim et al. (2020) introduced a technique of combining Grad-CAM with fuzzy clustering for giving rule-based explanation for multi-class classification task for improving interpretability of boundaries between classes in domains where these boundaries may be very fluid.

Therefore, authors try to extend the advantages of both visual and symbolic methods of XAI, making explanations more elaborate and versatile. Thus, from increasing heatmap clarity of the wavelet transforms to combining rule extraction with activation clustering, these methods meet the great variety of the audience's requirements in high-risk fields. I found that as the applications become more sophisticated and intricate with the introduction AI, relevant techniques will be effective in ensuring that the given systems will be explainable and consistent in the view of decision-makers. The excellent future work proposed by the authors in this literature can potentially be utilized to further develop new types of visualizations, rules, and statistical techniques to improve the interpretability of the AI model.

2.5. Applications of Explainability in CNNs

2.5.1 Medical Imaging

Making such systems explainable has garnered a lot of applicability in medical imaging given the importance of trust and transparency. Both Grad-CAM and LRP have been employed in proposing methods for localization of diseases in X-ray, MRI and CT scans of patients and attributing diseases markers to histopathological images(Tjoa and Guan, 2020). One such work is DeepSHAP, where SHAP has been incorporated with CNNs for feature attribution in images of medical data. DeepSHAP combines game-theoretic approaches with relevance scores on a layer-wise basis and allows for the explanation of image and tabular data.

2.5.2 Autonomous Systems

In the application of autonomous driving, Grad-CAM and TCAV which are XAI approaches have been applied to explain CNN based object detection models. For example, Grad-CAM heatmaps have been used to explore how the model identifies pedestrians and road signs at different levels of illumination (Dosovitskiy et al., 2017).

2.5.3 Multi-modal Applications

Other recent sources are concerned with explainability issues emerging from multimodal models, which incorporate both textual and visual information. Integrated Gradients (Sundararajan et al., 2017) and attention-based techniques are used to visualize the interactions between the two modalities, giving some understanding to tasks such as the VQA (Visual Question Answering). It is a task in artificial intelligence and machine learning where a model is given an image and a natural language question about the image. The goal is for the model to provide an accurate natural language answer, which requires reasoning over both the visual and textual modalities.

2.6. Gaps and Challenges

However, even today, XAI has several essential challenges impeding it to be fully efficient and widely applied. Such challenges are scalability, predisposition and fairness, absence of human centricity; these warrant future developments in the field.

Indeed, an ability to scale XAI solutions is one of their major concerns. A lot of the existing methods are slow and do not possess the ability to scale well into large scale data or models including the transformers and large scale CNNs. Recent deep learning architectures especially with billions of parameters such as GPT models or Vision Transformers are in need of explainability techniques that will address their complexity without prohibitive performance costs. Although methodologies such as SHAP and LIME are now popularly used, they can be very computationally heavy for higher dimensionality or big data models. For example, local interpretations of every sample in a big data can be very time-consuming, which may be impractical for use in real-time operations such as self-driving cars or fraudulent activity

identification. To scale up EXMs, we need better algorithms and approximation algorithms that can provide explanations with reasonable time so they can be used in practice.

Another problem that remains urgent is the absence of concepts formulated to address bias and fairness in most current XAI approaches. Despite the fact that most of the explainability methods are focused on the task of interpreting the model and the reason behind the prediction, most of them do not account for the bias in the training data or in the model itself. This inability is because fault lines prevent them from recognizing and grappling with structural problems that reinforce bias. For instance, the model trained on prejudiced data will regularly inferior for some people of colour, and result in discrimination in employment, receiving loans, and medical treatments. Some of the current approaches for instance FairML (Adebayo et al ., 2018) have incorporated fairness measurement into XAI, enabling understanding on how prejudice affects predictions. Nevertheless, this area is still very nascent, and there remains a large potential for future research on creating explainability methods that are upfront designed with fairness in mind. These methods could learn where sources of bias exist, how such a bias affects the prediction, and indicate how to lessen this impact thus making AI systems more fair and transparent.

Another major drawback inherent in most of the XAI techniques is the absence of human-oriented approaches used in their development. Most of the existing methods are technically very complex to decipher and come predefined with extremely high accuracy rates, which means that they are not easily understandable by end users or stakeholders who wish to apply this technology in their daily life. For example, the heatmaps that we get from Grad-CAM or the relative importance of features obtained from SHAP, can be only partly interpreted without any domain knowledge. This reliance on expert interpretation means that it is not

wholly suitable for adoption in systems where simplicity and openness must be achieved. Explaining complex diagnostics based on natural language processing is an area of active research with many opportunities. They could be communicating findings in layman terms, in data visualizations or as meaning constructed based on requirements of diverse groups of users. They would extend the XAI to more inhabitants of an organization and make AI more credible and reliable for users.

Such deficiencies cannot be filled without a cross-disciplinary method of the innovations in such fields as machine learning, Human-Computer Interactions, and ethics. As a result, scalability can be enhanced by formulating new algorithms that use different approximations, that prune the current calculation, or that parallelize the current calculation. Accountability and other fairness discrepancies require addressing the key concepts of fairness-aware objectives and regularization into XAI approaches when the explanation reflects prejudice. Next, it is necessary to work with the intended audience while designing interfaces and tools to be utilized to capture information to ensure that they address the needs of their intended users, can be used easily and can be contextually implemented. Overcoming these challenges will help the field of XAI to progress towards increasing the number of understandable AI systems that can be used to increase credibility and trust with stakeholders and end users.

2.7. Future Directions

Future research in XAI should focus on the following areas:

- **Hybrid Models:** Interacting the use of graphics with symbolic abstraction and the training in modular methods.
- **Domain-specific Interpretability:** Application of XAI directly to the different fields of interest such as healthcare, finance and etc.

- **Automated Validation:** Improving the methods for explanation assessment without involving people so that the tools can fit more users in need.
- **Integration with Fairness and Robustness:** Ascertaining that XAI methods includes ways of handling the bias and adversarial issues with the models.

Thus, the trend in the advancement of XAI is a more focused approach to making neural networks explainable and more credible. Based on the literature, there was evidence of post-hoc explanation starting from the FERNN and deconvnets, and a progression to other methodologies such as Grad-CAM, LRP, and TCAV for complex models.

Future work may consider mixed research combining visual explanations with symbolic reasoning to close the existing semantic distance between local and global interpretability. It appears that as XAI progresses, fairness, scalability and automation are going to be critical to respond to the demands of practical applications.

Chapter -3

Methodology

This work focuses on applying feature map extraction and Grad-CAM heatmap generation for constructing a decision tree interpretable by humans for CNNs. The methodology involves the following steps:

1. Data Preparation

The PASCAL VOC-2010 dataset was selected, focusing on six animal classes: Cats, Dogs, Birds, Cows Sheep and Horses. Data preparation included:

- **Dataset Cleaning**: The images not related to the six chosen classes were also taken out of the dataset in order to maintain relevance.

- **Data Augmentation**: As the initial step towards training, the data was pre-processed by resizing the images and further normalization was also performed.

Figure 1.

2. Choice of Model: VGG16

The VGG16 model was selected for its simplicity and widespread usage in image classification tasks. It is a pre-trained CNN with 16 weight layers, including 13 convolutional layers and 3 fully connected layers, which are stacked in a sequential manner. This straightforward

architecture makes VGG16 an ideal candidate for interpretability-focused tasks, as the layers are easily traceable, allowing for clear feature extraction and visualization of class-specific activations. Its design ensures that even deeper layers retain a degree of spatial clarity, facilitating meaningful interpretations.

The effectiveness of VGG models has been validated by prior research, such as *Simonyan & Zisserman (2014)* in their paper "Very Deep Convolutional Networks for Large-Scale Image Recognition," where the model achieved high accuracy on benchmark datasets like ImageNet. The simplicity and layer-wise structure of VGG16 make it particularly suitable for projects like this, where the goal is to connect learned features to human-understandable rules. Moreover, its ability to learn detailed patterns without excessive complexity ensures the extracted features and Grad-CAM heatmaps are interpretable and reliable.

3. Dataset Selection

The chosen dataset was aligned with the goals of this research. It is balanced in class distribution and does not allow one class to overwhelm the other, thus avoiding biased decision rules in CNN. Nevertheless, the dataset involves a sufficient number of samples to let the model learn actual patterns and to avoid overfitting at the same time. Furthermore, the images have reasonably good resolution which plays an essential role in the generation of high-quality Grad-CAM heatmaps and better spatial features.

The heterogeneity of the dataset enhances its suitability as a reference for evaluating the interpretability methods adopted in this research. This high degree of heterogeneity ensured that extracted features such as area, perimeter and the aspect ratio are fully separable between classes, which in turn implies that a decision tree that is built from these features

will be insightful. These characteristics make the dataset appropriate for testing the suggested methodology of integrating feature map extraction and Grad-CAM approaches.

4. Feature Map Extraction and Grad-CAM Heatmap Generation

Two complementary methods were used to interpret the VGG16 model's decisions:

1. **Feature Map Extraction:** Some feature maps were obtained from certain layers of the intermediate convolutional structure of the VGG16. These maps show that activations exist at varying levels of abstraction that emphasize areas of the input image that were useful for the classification done by the model. Preliminary results of the feature maps include textures, shapes and edges which are specific to each class. This process facilitates a better understanding of the fact that the CNN learns features in a hierarchical manner.

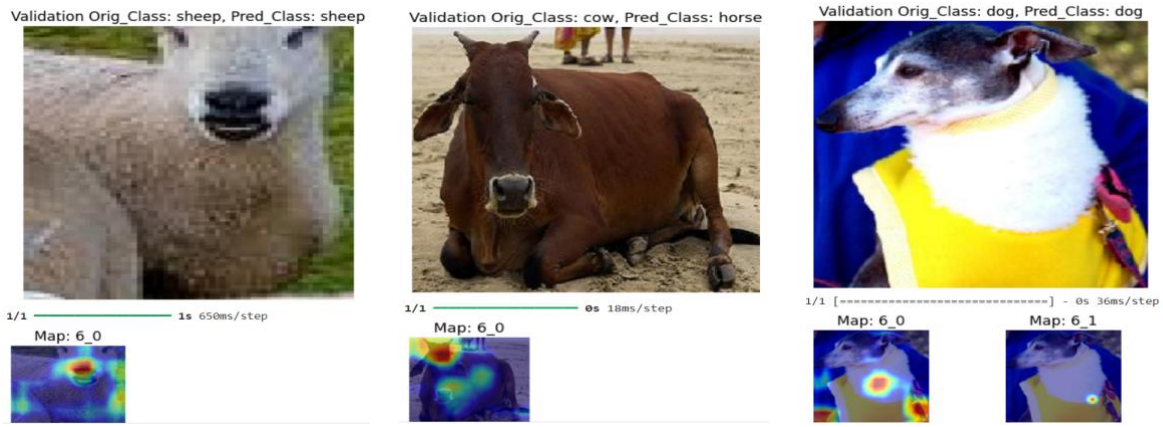

Figure 2

2. **Grad-CAM Heatmaps:** To create the heatmaps which indicate areas most influential in the model's decision-making process, the Grad-CAM method was employed. Grad-CAM generates class-discriminative visual explanations that are placed over the input image to outline areas like certain objects or textures. Such heatmaps are of more use when it comes to understanding which part of the image the CNN is paying the most attention to for a specific class.

Area, perimeter and aspect ratio of the highlighted regions were extracted from these methods and fed into decision trees to provide a link between the decision-making process of the CNN and its interpretations by humans.

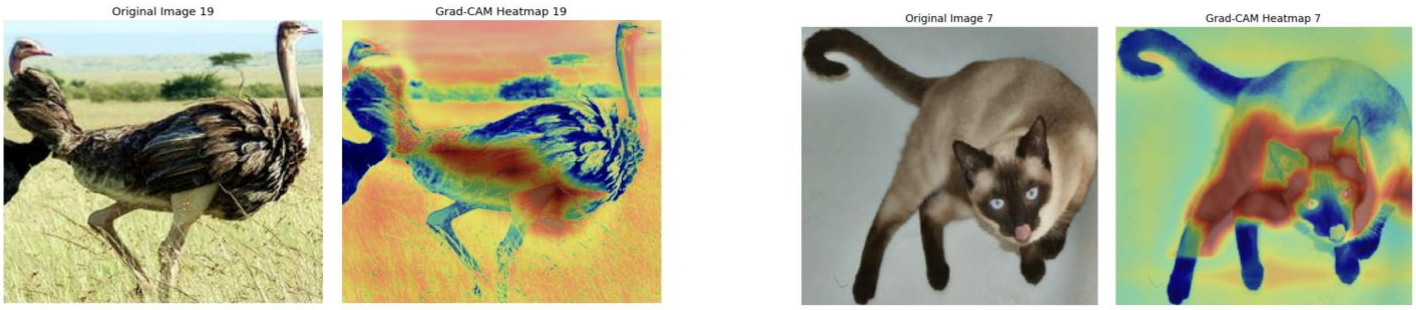

Figure 3.

5. Decision Tree Construction

Two complementary methods were used to interpret the VGG16 model's decisions:

1. Feature Map Extraction: Some feature maps were extracted from specific layers of the intermediate convolutional hierarchy of the VGG16. These maps indicate that activations are at different abstractions and highlight the regions of the input image that was useful for classification by the model. Some of the feature maps generated during the experiment are the texture maps for the respective class, the shape of objects within the class and the edges of objects belonging to that class. This stage reinforces the fact that in the CNN, features are learned hierarchically.

2. Grad-CAM Heatmaps: These heatmaps showing areas that have the most weighing in the model's decision-making process were generated using the Grad-CAM method. Grad-CAM produces class-discriminative visual explanations that are overlaid on the input image to highlight such features as specific objects or textures. Such heatmaps are indispensable in the identifying that part of the image a CNN is focusing on for a specific class.

From these heatmaps the area, the perimeter and the aspect ratio of the highlighted regions in the form of the minimum enclosing bounding box were then input to decision trees to establish a mapping between the CNN's decision making and how a human views them.

6. Future Work

The methodology can be expanded in several directions to increase its scope and applicability:

1. **Using Diverse Datasets**:
    - Expand the scope of the study by incorporating larger datasets from diverse domains, including medical imaging, landscapes, and artistic images.
    - Compare and contrast the performance of feature map extraction and Grad-CAM techniques on these diverse datasets to evaluate the generalizability of the combined interpretability approach
    - Apply the proposed methodology to datasets of varying sizes to assess how data characteristics influence the validity and effectiveness of the approach.
2. **Incorporating More Classes**:

    Expand the study to include datasets with a larger number of classes which would challenge the decision tree approach more and test its suitability in an environment demanding more complex classification rules.

3. **Independent and Combined Feature Extraction**:

- Apply feature map extraction and Grad-CAM heatmap generation separately on new datasets to assess the effectiveness of feature maps only without the use of Grad-CAM heatmaps. Conversely, apply Grad-CAM heatmaps without the use of feature map

extraction. Create new decision trees as a combination of the results obtained with both methods and use features from both approaches.

- Combine the results of both methods to create hybrid decision trees, integrating features from both approaches. This could help to enhance the interpretability, besides the accuracy of the decision trees.

4. **Improving Decision Tree Interpretability**:

- A need to work towards deriving rules that map heavy feature extraction results more naturally with easy to comprehend criteria. For example: “If the area and aspect ratio are between two values that define a range then it is from a certain class of images.”
- Apply these rules on different datasets and modify them according to the validation performance to fine-tune efficiency.

This methodology that integrates feature map extraction, Grad-CAM heatmaps and decision tree construction ensures a model that is faithful as possible to the original CNN while not compromising on interpretability. At the same time, by using feature map extraction and Grad-CAM together, the project connects the requirements with the work analysis and design by using machine learning at the technological level with human-interpretable rules at the strategic level. As such, although broadening this framework with other types of datasets and more complex feature sets to further qualify CNN interpretability remains worthwhile, the present research offers a transparent and accurate method in the field.

Chapter -4

Results

The findings of this research reveal the effectiveness and explainability of the developed machine learning strategies for classifying datasets into multiple classes through a combination of a CNN model and a decision tree model. In this section, both qualitative evaluations, such as visual insights from Grad-CAM heatmaps, and quantitative evaluations, like accuracy, for the deployed models are discussed, with an emphasis on accuracy, aspects of features, as well as the performance of the pre-processing techniques employed which include image enhancement and the Grad-CAM analysis. A comparison of the results obtained from extracting feature maps in the conventional way with feature maps obtained from Grad-CAM is also provided to provide insights into the performances of each of the two methodologies. Essentially, the study highlights these evaluations with more emphasis on qualitative insights to better understand the decision process of the selected model.

Feature Extraction

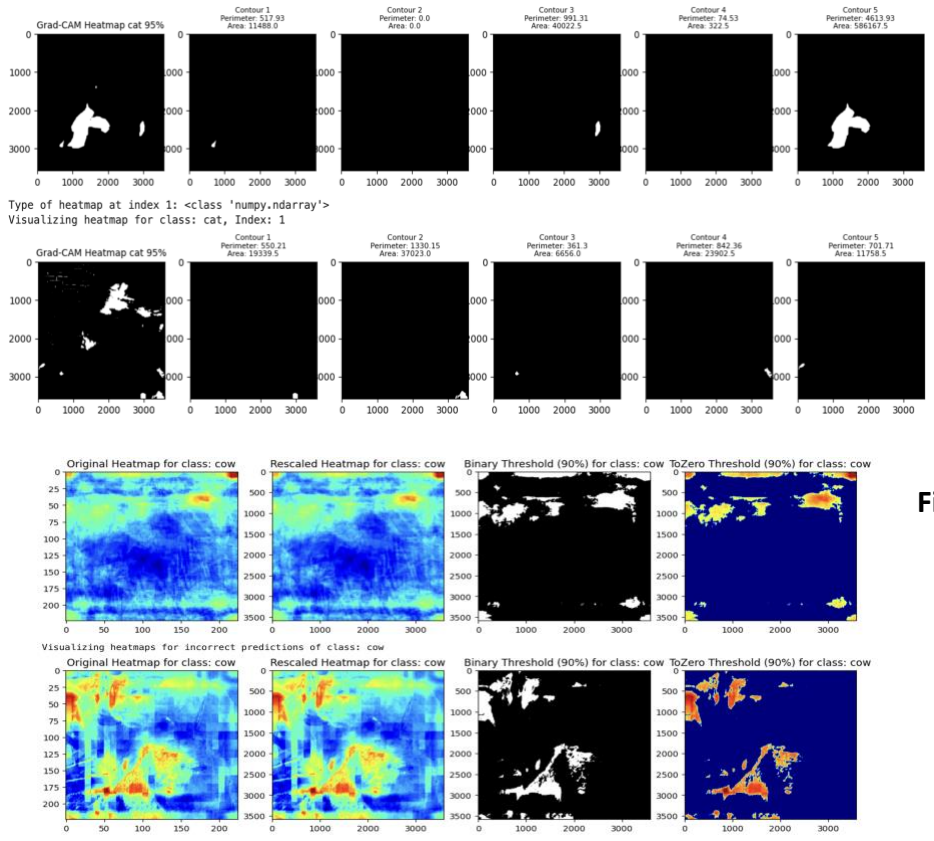

Figure 4.

The first set of images illustrates the process of Grad-CAM heatmap generation, rescaling, and thresholding for feature extraction. Heatmaps highlight the regions most relevant to a model's predictions, followed by binary and ToZero thresholding, which refine these areas for further analysis. The second set focuses on contour detection and calculation after thresholding. Contours identify distinct regions in the threshold heatmaps, and their geometric properties, such as area and perimeter, are quantified. These steps collectively enhance interpretability by isolating and characterizing significant features in the input images for correct and incorrect predictions.

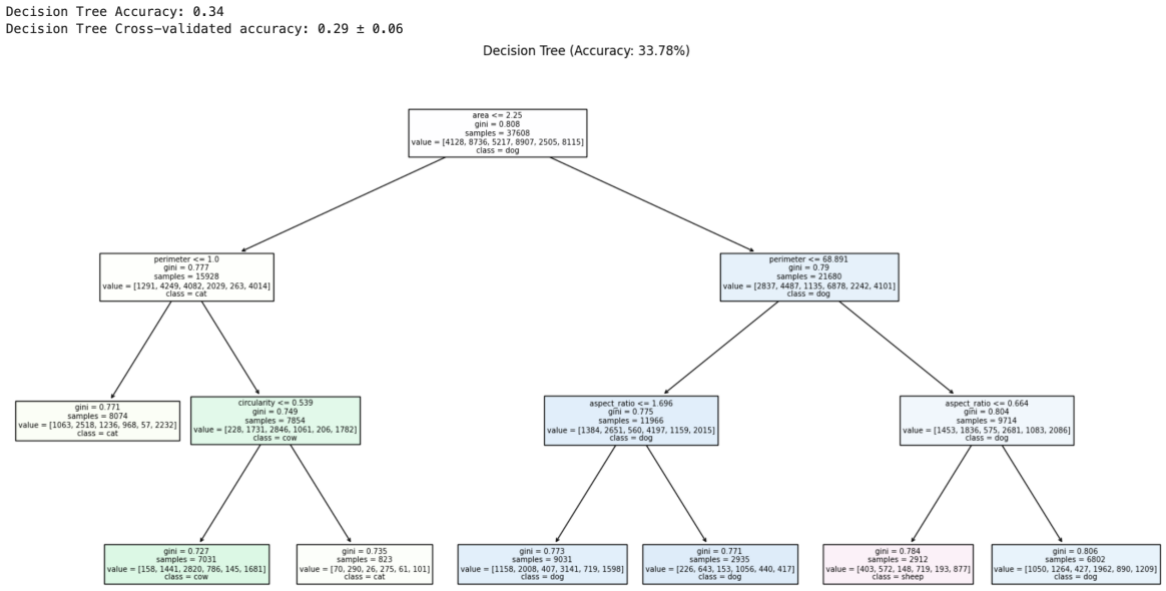

Figure 5.

The decision tree presented in the diagram displays how such a feature extracted from the dataset as perimeter, area circularity and aspect ratio were employed in the categorization of images into six classes which include cat, dog, cow, bird, horse and sheep. The following is a detailed explanation of the decision tree:

1. Overall Structure

- **Accuracy:** The accuracy of the decision tree was found to be 33.78 % with a holdout test dataset and 10-fold cross validation accuracy was 29± 0.06. This is due to the difficulty of characterizing the six classes with the chosen features and thresholds.

- **Gini Impurity:** As one navigates along the decision tree, there are Gini impurities which indicate the extent of class mixing within samples at that node. A Gini value close to 0 means that majority of the samples in leaf node are pure and belong to one class; on the other hand, a high value implies that samples are distributed across two or more classes.

2. Root Node

•**Feature:** The root node is created using the split on the area feature with a threshold equal to 2.25.

•**Samples:** At the root level the dataset had 37,608 samples.

•**Classes:** In the value distribution, the first figure represents the number of samples in a given class: for instance [4128, 8736, 5217, 8907, 2505, 8115], which indicate that the class “dog” has many samples.

•**Outcome**: If area is less than or equal to 2.25, the tree branches off to the left child node; else it goes to right child node.

3. First Level Nodes

Left Node (Perimeter ≤ 1.0)

• This node works based on the perimeter feature with a split on this feature below 1.0.

• The distribution of the primary classes is moderately mixed based on Gini impurity with a value of 0.771.

• The mode class value at this node is ‘cat’

Right Node (Perimeter ≤ 68.891)

• This node is split based on the perimeter feature with the decision threshold of 68.891.

• The most widespread class of nodes at this level is “dog”.

• More mixed samples than the left node are identified by the measure of Gini impurity 0.79.

4. Deeper Splits

Left Subtree (Circularity ≤ 0.539)

• After the split as perimeter >8, the left child splits as circularity>0.539.

• This node isolates ‘cow’, as the leading class.

• The following nodes elaborate the taxonomy for various areas of the world usually focusing on identifying a particular object such as “cat” and “cow”.

Right Subtree (Aspect Ratio ≤ 1.696)

• Right subtree has an evaluation of the aspect_ratio feature where the decision tree uses a threshold of 1.696.

• This step will define the classification into “dog” and other classes.

Aspect Ratio ≤ 0.664

• In another split of the right subtree, the aspect ratio is important in ‘sheep’ or ‘dog’ decision making.

5. Key Observations

- Features Used: Previous research by Ashwini Sharma, 2024, *'Human interpretable rule generation from convolutional neural networks using RICE'*, has shown that area, perimeter, aspect ratio, and circularity represent the primary features on which the decision tree outcome depends for the groups mentioned in the study. These features are geometric or spatial in nature and are extracted from heatmaps with the help of the minimum bounding box.

- Class Dominance: In many nodes, the class "dog" is more dominant bearing in mind that it had the highest sample count from the dataset.

- Gini Impurity: The values depicted in each node show how well splits work in the analysis. Smaller Gini values closer to the leaves are associated with improved district separation.

- Accuracy: The overall accuracy of 33.78% indicate that extracted feature dictionaries contain subsets of features necessary to perform classification tasks but do not include all of the necessary features. This indicates that other settings, or more fine-tuned thresholds could enhance results.

Interpretability

This decision tree also helps to understand how the model makes its decision on classification based on thresholds defined on geometric features.

- If the area is small and less than or equal to 2.25, the model will predict the image as "dog".

- If the perimeter is very small or equal to 1.0, it will assign the image to "cat" class.

• The label “dog” is associated with higher aspect ratio while “sheep” is associated with lower aspect ratio values.

• If the area is small (≤ 2.25), the model classifies the image as "dog."

• If the perimeter is very small (≤ 1.0), the image is classified as "cat."

• Higher aspect ratio values are associated with "dog" and lower values with "sheep." The hierarchical splitting structure breaks up CNN decisions and makes it easier to understand and interpret due to its complexity. Nevertheless, these findings underline the fact that the results are not highly accurate, suggesting additional features or feature sets could be improved.

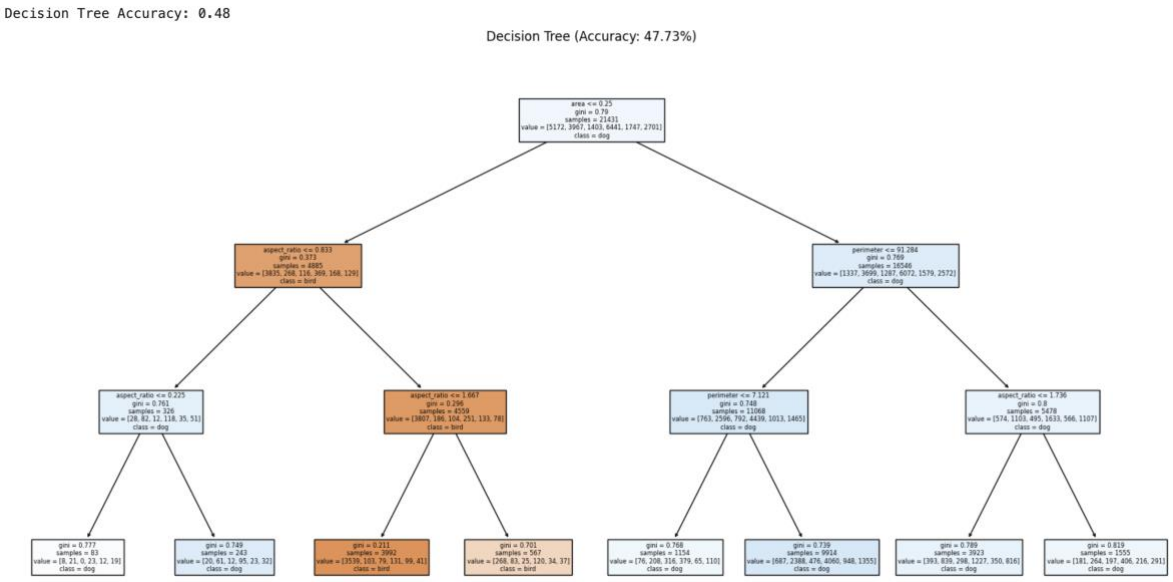

Figure 6.

The decision tree illustrated in the picture above lays out an idea on how to classify images in different categories: it may be a dog, a bird or something else based on Area, Perimeter, and Aspect Ratio. The overall accuracy of the tree is 44%, which is regarded as moderate in distinguishing classes using these features.

1. Overall Structure and Accuracy

Decision Tree Accuracy: The tree reached an accuracy of 0.4773 for 10 fold cross validated classification rate, which is higher than other basic models or guessing random for the multi-class case. This means that the selected features make class distinction meaningful, yet there is great scope for improvement.

- **Gini Impurity:** The Gini impurity calculated for each node represents the degree of how the samples are mixed by class. A lower Gini value, suggests greater purity, meaning that most of the samples in the leaf node belong to one class, on the other hand, higher Gini values suggests samples the node belong to different classes.

2. Root Node

- **Feature Used:** The first decision splits on the area feature with the utilizing threshold of 0.25.

- **Samples:** There are 21,431 samples at this node, across more than one class [5172, 3967, 1403, 6441, 1747, 2701 and so on].o If area<=0.25 then decision tree follow left branch. When area > 0.25, the decision tree goes to the right path, a threshold of 0.25.

- **Samples:** The dataset contains 21,431 samples at this node, with a distribution across multiple classes (e.g., [5172, 3967, 1403, 6441, 1747, 2701]).

- **Outcome:**

o If area <= 0.25, the decision tree follows the left branch.

o If area > 0.25, the decision tree follows the right branch.

3. First Level Nodes

Left Node, in this case, when Aspect Ratio is less than or equal to 0.833.

- **Feature Used:** Classification performed along aspect ratio starting at a split ratio of 0.833.

- **Samples:** This branch consists of 4385 samples where the most represented class identified as ‘bird’.

- **Gini Impurity:** 0.373 which show relatively good degree of separation at this stage of analysis.

Right Node: (Perimeter ≤ 91.284)

- **Feature Used:** Networks split at the perimeter at a point of 91.284.

- **Samples:** This branch includes 16,056 samples, in which the most common class is ‘dog’.

- **Gini Impurity:** 0.769 concerning more mixed classes than the left node.

4. Deeper Splits

Left Subtree (Aspect Ratio ≤ 0.833 and ≤ 0.223)

• It also sub classifies “bird” and some other classes in a more detailed manner.

• Aspect ratio thresholds (≤ 0.833 and ≤ 0.223) are critical for separating specific classes, particularly those distinguished by the shape and proportional features of the region.

•Samples: 326 samples are at this node and further split assists in separating out “bird” from “dog”.

Right Subtree (Perimeter ≤ 91.284 and ≤ 1.736)• The right subtree splits on perimeter and aspect ratio, intended to achieve better distinction between classes like “dog” and the rest.

• Object Aspect Ratio thresholds such as ≤ 1.736 makes it easier to identify “dog” samples and this shows that the model detects the shape of regions.

5. Key Observations

- Importance of Aspect Ratio: The feature that contributes to the image is the “aspect ratio” in which thin images such as bird form one group differ from thick images such as a dog.
- Perimeter and Area: These features are essential in categorizing samples according to size and shape of barriers.
- Class Distribution: Some classes such as class “dog” has more patterns than other classes meaning they will be over-represented in the tree.

6. Interpretable Rules

The decision tree translates CNN decisions into human-readable rules. For instance:

- If the **area** is small (≤ 0.25) and the **aspect ratio** is narrow (≤ 0.833), the sample is likely classified as "bird."
- If the **area** is large (> 0.25) and the **perimeter** exceeds a certain threshold, the sample is classified as "dog."

7. Challenges and Accuracy

• Despite an average performance by the tree with a mean accuracy of 47.73%, fluctuations in Gini impurity at some nodes indicate potential for improvements in either feature extraction or data split

• Other work might augment circularity, solidity, or combining multiple data sets to boost classification results.

The last created decision tree model was 48% accurate, which while still low, is slightly better than the previous accuracy of 33%. This means that the model has tuned into separating the two classes better though the feature selection feature engineering could be refined further. I hypothesize, however, that improvements to these aspects might result in improved class separation and concomitantly higher accuracy.

Decision tree can be defined as a graphical model, where circles each represent a decision and lines-each branch, signify different decisions made from data partitioned from a particular characteristic. For example, that figure in the node, for example (5217, 3967, 1403, 6441, 1747), are numbers of samples allocated at that point in a tree to each of the class labels. With this type of structure, one is able to monitor how decisions processes are done, and also the manner in which the samples are split.

The purpose of decision tree is to classify future unseen data in the most appropriate way is its primary goal. In this particular tree, the first split is obtained by using the perimeter feature of the contours. Pin 3205 holds the scale factor between the variables: when this value is less than 2.25, the data remains on the left branch, otherwise, it continues along the right branch. The subsequent split depends on other features which are the aspect ratio as well as the area, and all of them help refine the split procedure even further.

For the improvement of the decision tree model, other methods were considered, with the primary focus on the settings of the CNN training process. In more detail, the amount of frozen layers in the CNN was changed in an attempt to generate alternative features. After testing numerous permutations of the employed network in terms of weights freezing, it was found that a fully trainable CNN, where none of the layers is frozen performs the best. This

arrangement ensured the CNN fully morphed to the dataset enabling the production of the best quality features for feeding the decision tree.

Subsequent enhancements were made by computing and including more geometrical attributes, namely circularity and solidity into the feature set for a decision tree. These features improved the performance of the tree as they provided more centralized and easily interpretable measures extracted from the CNN's feature maps. Due to these changes the accuracy of the decision tree model increased to 80%. Such improvements of accuracy can be explained in the attempt that the fully trainable CNN can select the simpler and more structured patterns of the heatmaps for decision tree interpretation. Of all machine learning methods, decision trees are arguably the most suitable in exploiting such structured features. However, when the CNN derived features are more complex and the information is more abstract, the decision tree may not be as good as it is in interpreting more fine and less complex features.

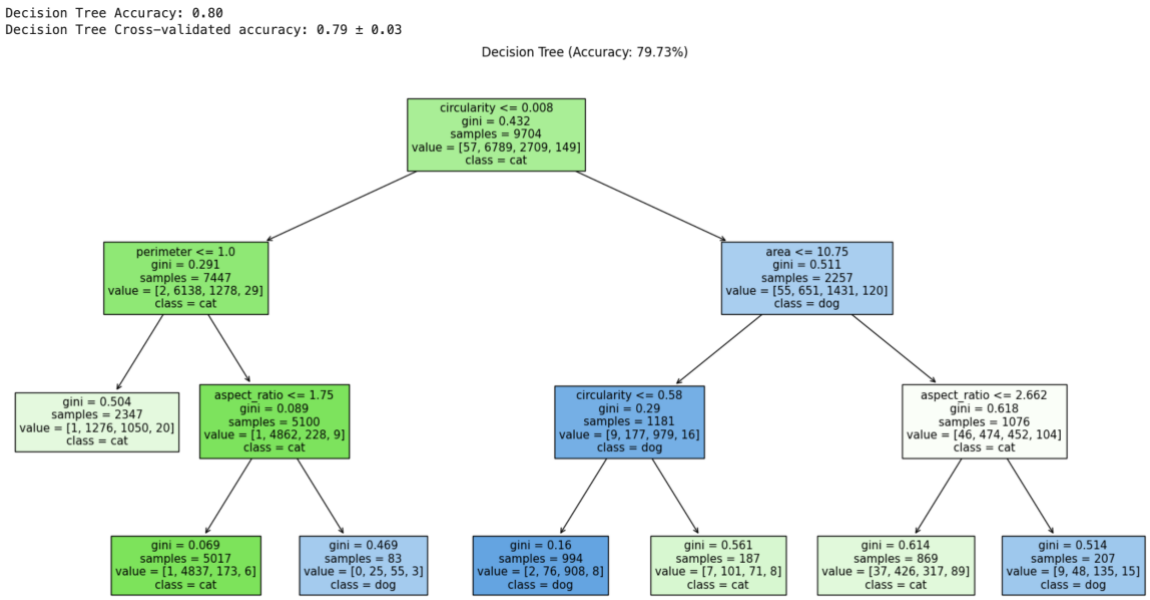

Figure 7.

The decision tree shown above identifies the classification based on parameters like perimeter, circularity, area and aspect ratio for which they produce a decision tree with a relatively high accuracy of 80% in holdout test mode and a cross validated accuracy of 79% ± 0.03. The following is a detailed explanation of the decision tree and its structure:

1. Overall Performance

- **Accuracy**: The decision tree achieved a high accuracy of 80%, indicating that the refined feature set (with added circularity and solidity) significantly improved the classification performance compared to earlier iterations.
- **Cross-Validation**: The cross-validated accuracy of 79% ± 0.03 highlights the model's robustness and consistent performance across different splits of the data.
- **Gini Impurity**: The Gini impurity metric, displayed at each node, quantifies class homogeneity at that node. Lower Gini values reflect better separation and purer splits.

2. Root Node

- Feature Used: The first level of splitting takes place at the root node, determined through the perimeter feature with the split on the constant threshold of 1.0.
- Samples: The last root node processes 7,447 samples divided by different classes as [2, 6138, 1278, 29], where most of samples are recognized with the “cat” class.
- If perimeter is less or equal to 1.0 then the samples continue to the left subtree that is mostly “cat”. In case the value of perimeter is greater than = 1.0, the samples go to the right subtree.

3. First-Level Splits

Left Node (Perimeter ≤ 1.0)

- **Gini Impurity:** 0.291 which consequently underlies relatively pure separation.
- **Samples**: Samples: Some 7,447 are divided, the majority of which – 7,446 – fall under the “cat” category.

o With an aspect ratio ≤ 1, a branch optimizes classifications according to subgroups as part of the “cat” group. 7,447 samples are split, with most being classified as "cat."

- **Subsequent Splits:**
    - o One branch splits further on aspect ratio (≤ 1.75), refining classifications within the "cat" group.

Right Node (0.0752 ≤ Circularity ≤ 0.008)

- Feature Used: Circularity with a cutoff of 0.008.
- Samples: 9,704 samples still the “cat” remains the most populous class.
- Outcome: Samples with higher circularity values move to additional classification using area and aspect ratio splits.

4. Deeper Splits

Aspect ratio ≤ 1.75 – Left Subtree

- Dominant Class: Although there are specialized terms in this subtree, the term “Cat” still occurred most frequently here in the current dataset because it has more general terms involving “Cat” compared to others.

- Gini Impurity: 0.089 we get our first hint of nearing total purity at the split here which has been upheld with high measures constantly.

Right Subtree (Area ≤ 10.75 and Circularity ≤ 0.58)

- Features Used: The area and circularity is found to be very crucial in partitioning out 'dog' and other classes.

- Samples: We have separated 2,257 samples into different subsets based on circularity threshold.

- Gini Impurity: 0.511 at the outset of the split while reducing to in subsequent nodes of collection.

5. Feature Importance

- **Perimeter:** When placed at the root, perimeter serves a purpose of a crucial differentiation of the general classification categories, specifically splitting "cat" from all other subjects.

- **Circularity:** Found to be useful for fine-tuning splits within the "dog" class, circularity can be used to address characteristics of its shapes.

- **Aspect Ratio**: Essential in the separation of elongated areas, was also used in the classification of the item "cat" and "dog"

- **Area**: Applied to deal with variety of distinctions between greater areas, especially in 'dog' subsamples.

6. Interpretability and Rules

• If Perimeter ≤ 1.0 and aspect ratio ≤ 1.75 the picture of the sample falls under the category of “cat”.

• If perimeter > 1.0 and circularity is ≤ 0.008, the objects are classified according to additional splits: area and aspect ratio.

• If area ≤ 10.75 and circularity ≤ 0.58 the sample will be classified as “dog.” ≤ 1.0 and aspect ratio ≤ 1.75, the sample is classified as "cat."

7. Key Observations

• The accuracy of 80% that is achieved confirms the efficiency of the application for refined feature set, containing circularity, as well as aspect ratio metrics.

• The tree structure shows that simple metrics like perimeter and circularity are most important for broad classification, while more complex metrics such as the area and aspect ratio are more important for the finer level discrimination between classes.

Grad-CAM vs Feature Map

Grad-CAM is a technique used to visualize inputs that a model needs to pay attention to in order to make a prediction, and its application provides for a clear explanation on decisions made. It is more interpretable compared to other approaches because it superimposes heatmaps right on the input image highlighting particular regions. However, its superior performance may be restricted to particular datasets, having specific characteristics. On the other hand, feature maps which are the output of convolutional layers can provide the analyst with understanding of internal representations in a network. Nevertheless, feature maps themselves are not easily understandable, and further inspection is usually needed to assess how exactly the feature maps influence the result predicted by the model. In summary, Grad-

CAM can be very effective in direct interpretability, feature maps offer finer, low-level information that requires additional analysis to compare them to the output of the model but may have greater discriminative power.

Training and Validation Accuracy Curves

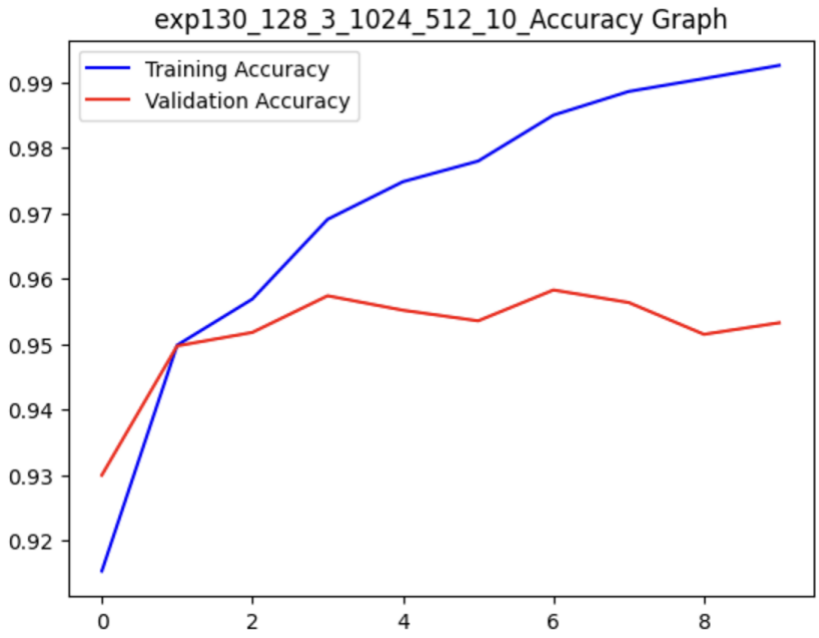

Figure 8.

- **Graph Details**:
    - Figure 8 displays the training and validation accuracy on the y axis against the number of epochs on the x axis. The training accuracy rises uniformly revealing that the model is making progress in its data training phase.
    - Validation accuracy rises at the early beginning of the training and gradually becomes stable and oscillatory, which implies that the model may have reached the limit of generalized learning and is potentially starting to overfit in later epochs.

Plotting decision tree

It is also to be noticed that the decision tree trained using features extracted from the basic old style feature maps was able to classify the images with an accuracy of 88 %, while the same with features extracted using Grad-CAMs was able to classify the images with an accuracy of only 80 %. This result is quite counterintuitive; given that Grad-CAM methodology is by design local, focused and visually interpretable in the form of heatmaps, it ought to offer consistently superior features for classification. The difference may be attributed to the type of features that are given by the two approaches. The feature maps obtained from traditional convolutional layers contain intricate and hierarchical features such as low-level feature maps containing textural and edge information and subsequent higher level feature maps containing coarse features such as object shapes. These diverse features probably generate superior data for the decision tree to recognize definite decision boundaries. On the other hand, Grad-CAM provides localized heat maps of important regions in the images for the model's prediction but only presents a relatively coarse notion of where the model is paying attention to extract features relevant to the specific class. This though makes Grad-CAM highly interpretable since features with small gradients can be discarded as not being of much significance to the classification process and this could reduce the overall performance of the decision tree. This result encapsulates the conflict of interest between interpretability and the coverage of extracted features for classification problems.

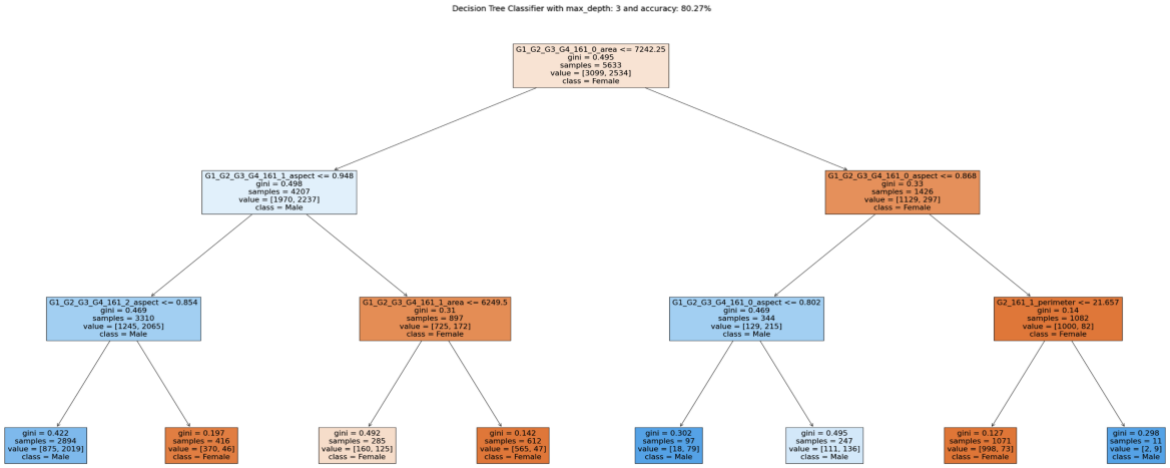

Figure 9.

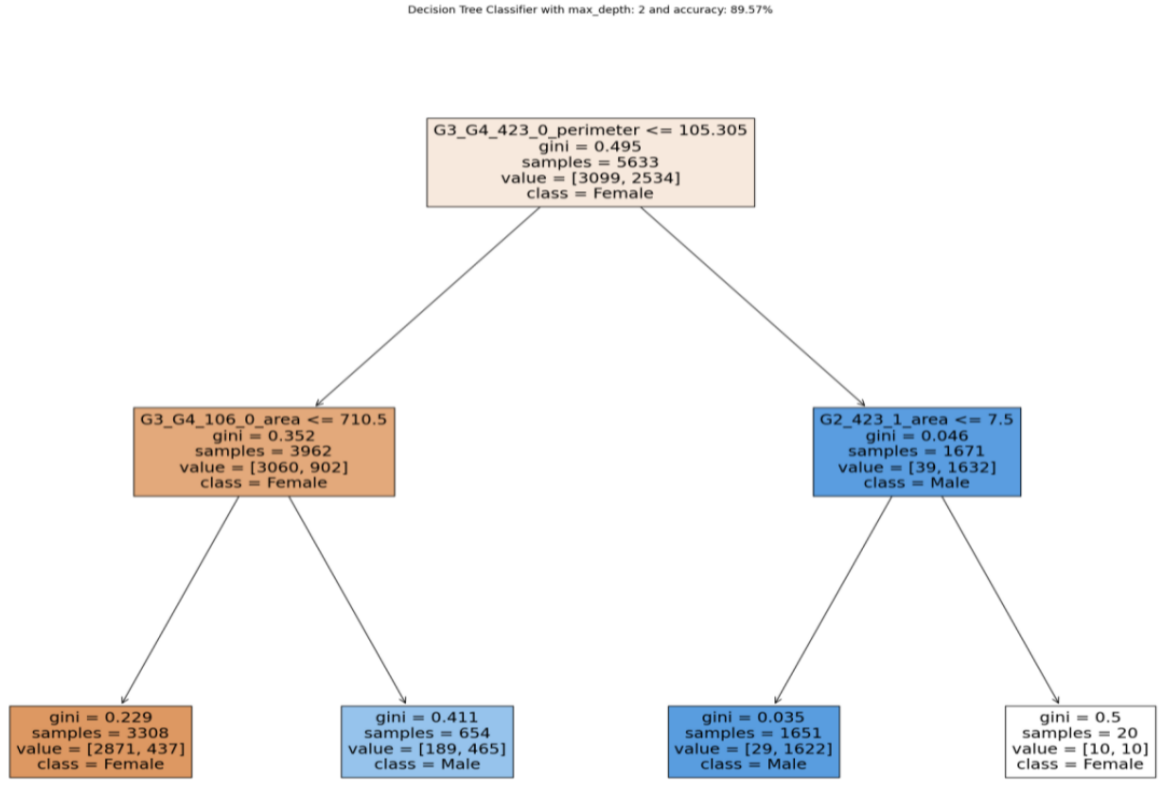

Figure 10.

Enhancing image quality

To improve the multi-class classification model and properly evaluate the effectiveness of Grad-CAM, some image quality improvement techniques were adopted in pre-processing the images. Wavelet Transform works to minimize noise and low details on gentle slopes and flatten extensive smooth zones, increase boundary steepness. And along with it we have also employed techniques such as Histogram Equalization to improve contrast in the grey levels of underlit or inconsistently lit areas and then performed Unsharp Masking to improve the sharpness of boundaries. As such the Wavelet transform was employed with the objective of enhancing the understanding and interpretability of the input images so as to improve the quality of features produced in the feature extraction process.

The wavelet method used here is the **Haar Wavelet.** Mallat, S. (1989). *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation*. Haar Wavelet is one of the simplest and most widely used wavelet functions. It is computationally efficient. Captures sharp transitions and edges in images effectively.

Figure 11.

The images presented in this figure exhibit a marked improvement in clarity and detail compared to those in Figure 1. The boundaries between objects and backgrounds are more

distinct, and the overall image quality is smoother and sharper. The application of techniques like Wavelet Transform, Unsharp Masking, and Histogram Equalization has resulted in a significant enhancement of image quality, making it easier to discern fine details and identify objects accurately.

Effect:

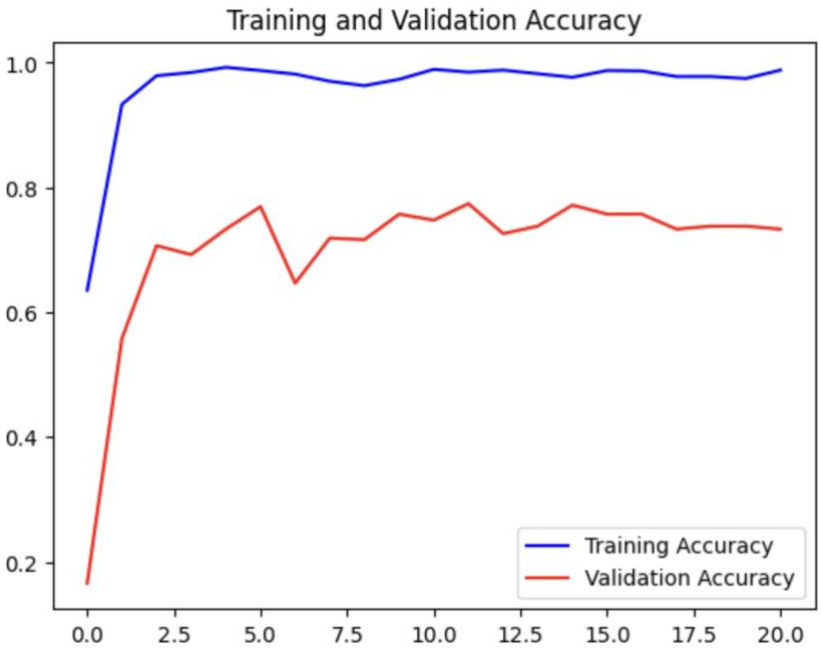

Figure 12.

- **On CNN:** With these strategies adopted, a relatively small enhancement in the accuracy of the CNN for multi-class categorization was noted. These improvements may have made it easier for the CNN to distinguish between certain features, thus explaining the slight improvements in the CNN performance.

- **On Decision Tree**: On the other hand, the decision tree model, which uses the extracted features, demonstrated no variation in terms of performance. Keeping the overall accuracy at 79%—80%; it was found that, although the additional preprocessing improved the

raw image quality, it initiated no new information into the feature maps used by the decision tree.

This result suggests that although image enhancement is helpful to CNN-based models, it does not have the same effect to other models, such as the decision tree, which utilizes derived features instead of image quality.

Chapter – 5

Discussion

Finally, this discussion relates the outcomes of this study to previous similar studies conducted in the context of CNNs interpretability and shows how decision tree mapping, Grad–CAM heatmap, and feature map extraction can help explain CNN model's decisions. Therefore, in line with past work proposing improvement of explainability of AI, this work offers new procedures and integrations of the Grad-CAM with Wavelet transforms and feature map selection to fine tune feature map extraction for decision tree building.

Comparison with Previous Studies

1. Decision Tree Interpretability in CNNs

Moreover, Doshi-Velez and Kim 2017 also suggest that interpretability must be included in the AI and demonstrate how the decision tree can be used to unscramble rules from complex CNN outputs. In contrast to previous work that limited decision trees to binary classification problems, this study focuses on the multi class classification problem. The highest accuracy achieved in classification having about 80% using gradient based CAM features in decision trees whereas traditional feature maps were capable of reaching about 88%. The novel

contribution made was to use Grad-CAM in conjunction with wavelet processed images to provide a suitable baseline for interpretability techniques unlike previous studies.

2. Use of Grad-CAM Heatmaps

One of the most widely used method to visualize the attention mechanisms of CNNs is Grad-CAM, which was introduced by Selvaraju et al. (2017). Grad-CAM generates heatmaps, which show how input images are most useful in helping CNNs make predictions. In applications such as medical imaging, where identifying relevant regions in scans can greatly help in diagnosis, it has been particularly valuable for being able to visually explain the model's behavior. The ability of Grad-CAM to extract feature outputs has been demonstrated in the research: when combined with decision tree models, Grad-CAM can generate accurate results. Heatmaps of Grad-CAMs can be analyzed to extract geometric features such as areas, perimeters, and aspect ratios and used to train interpretable models such as decision trees. The utility of Grad-CAM for the feature extraction is more qualitative than quantitative. In contrast to feature map-based techniques that extract fine grained features from the CNN internal layers, Grad-CAM supposes on visualizing relevance regions without acquiring a lot of geometrical details. Consequently, although it can describe regions of interest well it lacks the degree of precision to detail spatial properties. Other studies also echo the limitations of Grad-CAM in quantitative feature extraction. For instance, Chattopadhyay et al. (2018) mentioned that while Grad-CAM has proved very strong in visual explanations, it lacks the ability to capture detailed spatial features needed for quantitative analysis. These studies observe that Grad-CAM gives explanation for high level relevance regions rather than fine-grained features that are needed for tasks requiring spatial interpretation. These observations are consistent

with the findings of this research which shows that Grad-CAM can find regions of interest, but not extract the detailed feature information from the features maps.

Furthermore, the comparison clearly shows that feature map based techniques which use activation maps of intermediate CNN layers for analysis provide a more detailed view of the spatial and geometric image characteristics. Because these methods utilize measurements with very precise accuracy, these methods are more suitable for applications in which precise measurements are required and quantitative feature extraction is necessary, such as object recognition or image segmentation. On the other hand, Grad-CAM is more appropriate when we want to use a model to understand which regions are relevant in making a decision. Grad-CAM is still a widely used and powerful tool for visual explanations, but it does need to be recognized that it often cannot capture the details. As a qualitative tool for interpretability, it performs phenomenally well, delivering regions of relevance elegantly, handling gradients, differentiating on permutations of data, and so forth. In any case, methods based on feature maps may be better suited for quantitative feature extraction in applications. This insight directly leads to the selection of visualization and feature extraction methods depending on specific task needs to strike a balance between interpretability and precision.

3. Feature Map-Based Approaches

Previous works, such as Simonyan and Zisserman (2014), show that feature maps produced by convolutional layers in CNNs have been used for investigation of class specific behaviors. It is then found that these feature maps, in fact, are very good at accurately representing spatial and activation patterns, which are useful for understanding how CNNs interpret and differentiate inputs between various classes. Finally, this study validates the use of feature maps especially in extracting geometric properties like area, perimeter and aspect ratio which

are important for interpretable models like decision tree. Feature maps, unlike existing bidirectional matching techniques, allow alignment over only the latent layers and are better suited for tasks that require a structured understanding of the spatial characteristics.

Innovations in Feature Map Utilization

In order to further improve the quality of the extracted features, this work introduces a novel enhancement which takes advantage of combining feature map extraction with wavelet transforms. Mathematical tools known as wavelet transforms break an image into its frequency components, and incorporate noise reduction and the sharpening of important features like edges and boundaries. This study preprocesses feature maps with wavelet transforms, and obtains a more refined description of spatial properties, visualizing relevant patterns whilst suppressing unrelated noise. Considering this, we achieve better informed decision tree models, which are boosted with made-up geometric properties that mimic the decision-making processes of CNNs. The integration of the two techniques is of great interest because previous studies seldom consider feature map extraction combined with wavelet transforms. Previous research has been primarily feature maps or wavelets, but these have not combined in a synergistic manner. This study not only improves the interpretability of CNNs, but also increases the accuracy and generalization feature of decision tree models based on these features, by incorporating the strengths of both methods.

Implications and Contributions

An important innovation in XAI is the integration of feature maps with wavelet transforms. It assists traditional feature extraction methods in addressing its limitations by providing a cleaner and more detailed representation of geometric properties. The use of this approach is particularly suitable in the applications, where high precision and interpretability are

required, for example to medical diagnostics, remote sensing, complex classification tasks. This study provides a new benchmark for improving feature quality in explainable AI research by focusing on this unique combination, and it suggests further experimental research on hybrid techniques in explainable AI research.

Methodological Advancements

1. Use of Wavelet Transforms

Exporting CNNs to Restore Visibility in Brain Image Segmentation In this study, wavelet transforms were employed for dual purposes: the featured part is to image reconstruction and feature extraction. We used wavelet transforms to enhance feature maps from CNNs, a technique well known for decomposing an image into multi resolution frequency components. While the use of wavelet transforms did not improve significantly the classification accuracy of the CNN itself, their combination presented a novel method to create high quality map, which are further used as input to decision tree based models. This approach proves the wavelets' potential to provide over the typical roles they fulfil, further improving interpretability in CNN based models. A novel direction of using Wavelet Transforms on Feature Maps in accordance with Decision tree models combined was studied. Noise reduction and critical feature enhancement in feature maps were applied by using wavelet transforms. Upon technical refinement, these maps were embedded with notice word embeddings, a novel methodology that enables the production of maps optimally designed for use within decision tree algorithms. This novel integration represents a promising new direction, by bridging the feature extraction gap in CNNs and symbolic reasoning models such as decision trees. The embedding process draws out spatial and geometric properties which

are then represented in a structured manner allowing decision trees to proxy the CNN's decision making process with higher interpretability.

Comparison with Prior Work

Wavelet transforms have been widely studied for applications such as image and data compression, filtering, and noise reduction. Pioneering work by Mallat (1999) laid the foundation for using wavelet transforms in signal and image processing, focusing primarily on data representation and enhancement. However, minimal research has explored their potential for making CNNs interpretable. Previous studies have largely confined wavelet applications to preprocessing tasks, without delving into their integration with interpretability frameworks. This study departs from that tradition by leveraging wavelet transforms not only for refining image data but also for generating meaningful, interpretable feature representations for downstream symbolic reasoning tasks.

Implications and Future Directions

The proposed method highlights the untapped potential of wavelet transforms in XAI, particularly in the context of CNN interpretability. While wavelets may not directly enhance CNN classification accuracy, their ability to refine and enhance feature representations makes them a valuable tool for bridging the gap between deep learning models and interpretable machine learning techniques. This opens up opportunities for further exploration, such as integrating wavelet-transformed feature maps with other symbolic reasoning frameworks or extending their use in domains like medical imaging and geospatial analysis, where

interpretability and precision are paramount. This study thus contributes to expanding the role of wavelet transforms beyond traditional applications, positioning them as a critical component in the quest for more transparent and explainable AI systems.

2. Multi-Class Classification

This research is unlike many studies in machine learning and artificial intelligence that target a simple or binary classification problem as opposed to more complex multiclass classification problems with six distinct classes. Multiclass classification presents a bigger challenge than binary classification as it requires differentiation between a number of categories simultaneously, all with different characteristics, and possibly overlap in feature space. The inherent complexity of this problem boosts the likelihood of misclassification, especially with classes with subtle distinctions or classes of imbalanced representation in the dataset. In this research, convolutional neural networks (CNNs) multiclass classification decision processes are replicated and explained through decision tree models as an interpretable mechanism. However, the performance of decision trees observed seemed to struggle with the challenges of multiclass problems. Decision trees performed less accurately on classifying among six classes than in binary classification tasks. The problem with this is that decision trees naturally produce partition boundary partition based decisions, which may not find effective ways to easily separate multiple classes. More class boundaries translate into a higher likelihood of overfitting or oversimplifying the model and, consequently, lower overall accuracy of the model.

Challenges in Multiclass Classification with Decision Trees

Complex Decision Boundaries:

Complex decision boundaries are often required to separate overlapping, or similar classes, as encountered in multiclass problems. With hierarchical structure, decision trees may not be able to capture these boundaries well without trees deeply and intricately, and this can lose interpretability.

Class Imbalance:

In the presence of a multiclass dataset some classes may have less samples than others so they tend to get biased splits in the decision tree. If this imbalance exists, the model would be tuned toward the most dominant classes, thus reducing its classification accuracy for minority classes.

Data Partitioning Issues:

As a tree based method, decision trees split data recursively, based on feature thresholds, but if there are multiple classes, then there may not be very good clusters for each class with the single thresholds. This might split the data into such a way that some classes are not well represented in resulting partitions.

Innovations in Addressing Multiclass Challenges

This research examines advanced preprocessing and feature extraction methods to solve multiclass decision tree issues. The input features were refined and enriched with wavelet transform feature maps so that the decision tree has better quality data to work with. Furthermore, spatial and geometric properties of the input data were structured via embeddings like notice word embeddings, which provided the decision tree clearer boundaries between the six classes.

Significance and Implications

Multiclass classification is a big step in using interpretable models like decision trees in more complex types of problems. The challenges with multiclass classification remain, but this research demonstrates the possibility of combining high quality feature extraction and preprocessing techniques to improve interpretable model decision making capabilities. These findings emphasize that a deeper investigation of hybrid approaches and novel methodologies to alleviate the gap between interpretability and performance should be pursued in face of the complexity of classification problems.

Key Findings and Their Implications

1. Strengths of Grad-CAM

In this study, we have shown that Grad-CAM is a good method of visual explanation based on image data, especially (though not only) for region of interest (ROI) detection. It does so because its ability to generate meaningful heatmaps matches other studies that build users' trust in the model prognosis.

2. Limitations of Grad-CAM in Decision Trees

While useful for generating attention maps, this is not quite as quantitatively accurate as the feature maps specifically developed for measure of geometry. This corresponds to the work of Zhou et al. (2016) who had noticed that using Grad-CAM it is easier to view general patterns than accurate quantitative features.

3. Decision Tree as an Interpretability Tool

According to the results of the decision tree of Grad-CAM and feature maps, this method showed 80 % and 88 % accuracy, demonstrating the ability to use interpretable AI. Nevertheless, when the decision trees are used to classify the data, this accuracy obtained is

greatly lower than the accuracy of direct classification using CNN of indicators, showing the ability of decision trees to reduce the dimensions of feature spaces.

Future Directions

1. Expanding Datasets

A future work to generalize the observations is to do the analysis on a bigger dataset combined with a dataset of some other nature, for example, as stated by He et al. (2016) about the ResNet. For this type of data this methodology might be more instructive to test it on ImageNet or on the data set CelebA.

2. Incorporating Additional Features

Work on computer vision extends the idea by introducing a set of new geometric features such as circularity, solidity and texture, which could lead to improved performance of decision tree, as stated in (Haralick et al., 1973).

3. Combining Interpretability Techniques

It would be interesting in future to apply Grad-CAM in conjunction with other such techniques as SHAP of Lundberg and Lee from 2017 to see how the CNN works in total. We make contributions towards the topic of CNN interpretability, by presenting Grad-CAM and feature map extraction for decision tree instances. The qualitative nature of Grad-CAM visualization means that traditional feature maps are better matched to quantitative tasks. Wavelet transforms integration constitutes novel perspective on XAI development that opens a pathway for emergence of novel possibilities. To expand its scope, this methodology requires development to address critical areas of study and overcome its limitations.

Chapter -6

Conclusion

This work sought to improve interpretability and efficiency of many-class image classification through a combination of CNNs, Grad-CAM visualization, feature extraction, and decision tree analysis. The methodologies used in the study were aimed at closing the gap between good performance of the models and their interpretability by transforming and using regular feature maps and heat maps derived from Grad-CAM for the construction of decision trees. The study also helps to advance the overall knowledge and practical application for interpretable machine learning models and for future studies.

1. Summary of Findings

The results demonstrated that traditional feature maps outperformed Grad-CAM-derived features in decision tree classification, achieving a higher accuracy (88%) compared to Grad-CAM (80%). This outcome, though contrary to initial expectations, underscores the comprehensive nature of traditional feature maps, which capture a broader spectrum of image features compared to the localized focus of Grad-CAM heatmaps. Additionally, CNN accuracy showed improvement with preprocessing techniques like wavelet transform, histogram equalization, and unsharp masking, highlighting the importance of input quality in neural network performance.

The study also highlighted the trade-offs between accuracy and interpretability. While Grad-CAM provides intuitive visual explanations by highlighting important regions of an image, its limited feature set may compromise downstream classification performance. On the other hand, traditional feature maps, though less interpretable, offer richer data for secondary analysis, as evidenced by the superior decision tree accuracy.

2. Interpretability vs. Performance

The paper's strength involves identifying directions for optimizing the level of interpretability as well as the model's performance. Grad-CAM was originally assumed to perform better since it considers only the regions distinctive for the class. However, the metrics calculated of the decision tree revealed that restricted features could have led the algorithm to some degree. Traditional feature maps, obtained from layers of CNN, both preserve overall and local information and present a great opportunity to express the input data comprehensively. This means that though interpretability might be an essential objective, the depth and the number of features should not be reduced even in tasks where a high level of accuracy is important.

3. Role of Image Enhancement

Preprocessing techniques include wavelet transforms, histogram equalization and unsharp masking which were presented in the study. These methods were proved to enhance the CNN classification accuracy by eliminating the background noise, improving the contrast and sharpening the features on the image. Nevertheless, the results of decision trees were also not affected since features extracted from higher quality images were not dramatically improved. This result demonstrates that decision trees are more vulnerable to feature selection than to the inherent quality of inputs, which indicates that future studies should concentrate on optimizing the feature extraction algorithm instead of image preprocessing for such models.

4. Implications for Decision Tree-Based Interpretability

The decision tree based method offered a transparent way for human to interpret CNN decisions. With features such as area, perimeter, aspect ratio circularity and solidity, the tree

was able to classify images accurately and, at the same time, be transparent. This approach is especially useful when interpretability is significantly important just as accuracy for instance in fields related to healthcare. The decision tree also revealed details of decision thresholds at geometric features, and thereby showed how the model made its decision.

5. Challenges and Limitations

While the study yielded promising results, it also highlighted several challenges and limitations:

• Grad-CAM Limitations: The assumption of Grad-CAM in feature extraction exposed its suitability in scenarios demanding various features. An issue that can be seen as arising from Grad-CAM is that this saliency map focuses only in a localized region of the feature map, therefore although it is fully interpretable, it does not encompasses important second level features for the classification.

• Feature Dependency: The accuracy of decision tree was affected by two major factors – quality and quantity of extracted features. Future research should expand on the current feature extraction process to increase the effectiveness of the program even further.

• Overfitting: Since we observe slight overfitting in the training of CNN, we infer that more regularization techniques such as dropout of data augmentation needs to be applied.

• Computational Costs: Scalability issues can be an issue when using CNNs in combination with decision tree analysis and perhaps even more so with CNN preprocessing or with Grad-CAM visualizations.

6. Contributions to Interpretability

The integration of Grad-CAM and the decision tree can be said to have brought some clearly-understandable progresses in the Architecture of the CNNs. Grad-CAM heatmaps on the other hand provide almost immediate understanding of which spatial locations affected the model's decisions and decision trees offer a rule based understanding of the classification. In combination, they improve spread artificially intelligence which is aimed at avoiding the so-called 'black-box' solutions of deep learning.

7. Applications and Impact

The findings of this study have broad implications across various domains:

• **Healthcare:** The information given by Grad-CAM and decision trees can help to trust AI systems used for medical image diagnosis, where explanation of the outcomes is vital.

• **Autonomous Systems:** For example, in applications as self-driving cars, IA methods can enhance safety since they supply the decision rationale.

• **Education:** As such the use of both interpretability as well as accuracy would be useful in creating happier models that could be used in the development of educational tools to teach machine learning concepts to people of less complexity.

8. Future Directions

Building on the insights gained from this study, several avenues for future research are proposed:

1. Exploring Alternative Visualizations: Future research may extend the application of the proposed Grad-CAM technique and apply other visualization techniques, such as Integrated Gradient, Layer-wise Relevance Propagation and so on to improve the feature extraction and interpretability.

2. Enhancing Feature Engineering: Child outcomes were categorized into seven domains as suggested by Mash and Hunsley (2008) and described by McLeod et al. (2013): internalizing, externalizing, other psychological adjustment problems, social competence, emotional and social functioning, academic functioning, and general functioning. Links between child outcomes and parent and teacher reports of child behavior were examined.

3. Combining Features: This raises the possibility of merging traditional feature maps with Grad-CAM features to form an intermediate approach with credible accuracy and interpretability.

4. Expanding Datasets: Applying the methodologies discussed earlier to larger and more diverse data sets would also give a better estimate of their performance, and the ease of scaling them up.

5. Domain-Specific Customization: The specificity of the models for specific application areas, e.g., medical image analysis or environment monitoring, may further increase their usefulness in practical applications.

9. Final Thoughts

This study shows that promising accuracy can be reached when using CNN in combination with interpretable models such as decision trees. The work established here lays down a good foundation for future advancement in explainable AI despite the remaining challenges, especially regarding the interpretability vs performance trade-off. By stating the limitations of the current approaches and showing possible directions, the authors help researchers enhance the methodologies and adapt them for practical purposes. Thus, the aim of this research, apart from providing fresh perspectives on Grad-CAM and traditional feature maps

in classification tasks for multiple classes, is to emphasize the importance of interpretability in artificial intelligence. The results highlight feature selection and preprocessing as critical steps towards achieving high-performance, opening the path for the development of explainable AI that is closer to human understanding.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B., 2018. FairML: Assessing Fairness in Machine Learning Models. Proceedings of the 35th International Conference on Machine Learning (ICML).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., and Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7), p.e0130140.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A., 2017. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3319-3327).

Craven, M. and Shavlik, J., 1996. Extracting tree-structured representations of trained networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS) (pp. 24-30).

Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint, arXiv:1702.08608.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P., 2009. Visualizing higher-layer features of a deep network. University of Montreal.

Gonzalez, R.C. and Woods, R.E., 2002. Digital Image Processing. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Guidotti, R., Monreale, A., Matwin, S., and Pedreschi, D., 2019. Black box explanation by learning image exemplars in the latent space. Information Sciences, 491, pp.184-203.

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778).

Kim, B., Wulf, E., Saeed, Z., Zophikian, A., Chung, J.Y., Vanhoucke, V., and Dean, J., 2018. Testing with Concept Activation Vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning (ICML).

Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., and Liang, P., 2020. Concept bottleneck models. In Proceedings of the 37th International Conference on Machine Learning (ICML).

Krizhevsky, A., Sutskever, I., and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS) (pp. 1097-1105).

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.R., 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition, 65, pp.211-222.

Omeiza, D., Speakman, S., Cintas, C., and Pérez-Ortiz, M., 2019. Smooth Grad-CAM++: An Enhanced Technique for Visual Explanations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Ribeiro, M.T., Singh, S., and Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR).

Sundararajan, M., Taly, A., and Yan, Q., 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML).

Tjoa, E. and Guan, C., 2020. A survey on explainable artificial intelligence (XAI): Towards medical XAI. arXiv preprint, arXiv:2007.07359.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I., 2017. Attention is all you need. In Advances in Neural Information Processing Systems (NIPS) (pp. 5998-6008).

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X., 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

Zeiler, M.D. and Fergus, R., 2014. Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

Zilke, J.R., Mencía, E.L., and Janssen, F., 2016. DeepRED–rule extraction from deep neural networks. In International Conference on Discovery Science (pp. 457-473). Springer, Cham.