Arab American University
Faculty of Graduate Studies

Explainable Deep Learning Methods for Neuroscience Data to Analyze the Extracted Features in The Hidden Layers

By

Ahmad Khatib

Supervisor

Dr. Ahmad Hasasneh

Co- Supervisor

Dr. Jürgen Dammers

This thesis was submitted in partial fulfillment of the requirements for the Master`s degree in Data Science and Business Analytics

February /2024

Thesis Approval

Explainable Deep Learning Methods for Neuroscience Data to Analyze the Extracted Features in The Hidden Layers

By

Ahmad Hasan Khatib

This thesis was defended successfully on 20/02/2024 and approved by:

Committee members — Signature

1. Dr. Ahmad Hasasneh: Supervisor
2. Dr. Jürgen Dammers: Co-Supervisor
3. Dr. Mahmoud Obaid: Internal Examiner
4. Dr. Anas Samara: External Examiner

Declaration

I, Ahmad Khatib, hereby declare that this thesis, titled "Examining the Features in Hidden Layers of Deep Learning Models Applied to Neuroscience Data," is my own original work and has not been submitted previously for any degree or qualification at any other institution. All sources of information and data used in the research have been duly acknowledged. I affirm that the work presented herein is the result of my independent effort and has been completed in accordance with academic integrity standards. I understand that any breach of these standards may result in disciplinary action. I sign this declaration in full acknowledgment of its contents and implications.

The Name of The Student: Ahmad Hasan Ibrahim Khatib

ID: 202012390

Signature: Ahmad Khatib

Date: 16.7.2025

Dedication

To my wife, kids, parents, and friends—your unwavering support has fueled my journey through this master's program. Thank you for being my anchor, my motivation, and my source of strength. This achievement is as much yours as it is mine.

Acknowledgments

The success of this research is intricately linked to the combined efforts of the Human Connectome Project and the Institute of Neuroscience and Medicine at Forschungszentrum Jülich. The total number of 372 MEG recordings from 126 different subjects provided a robust foundation for our investigations.

This work would not have been possible without the commitment and collaboration of this institution, and we express our sincere thanks for their contributions to the advancement of our understanding of neural processes.

Abstract

In recent years, deep learning models have revolutionized neuroscience by uncovering the complex layers of hidden information. However, unraveling the meaning of features within these hidden layers remains a challenge. This study, titled "Examining the Features in Hidden Layers of Deep Learning Models Applied to Neuroscience Data" employs explicable deep learning techniques to elucidate the decision-making processes of convolutional neural networks (CNNs) when applied to neuroscience data.

Understanding hidden layers in deep learning models, especially in applied applications such as connectivity analysis in neuroimaging, is challenging due to their opaque representations, non-linear transformations, and increased dimensionality. In such cases, this opacity hinders the interpretation of neural representations learned by the model, limiting insights into brain function. Balancing model complexity is crucial to accurately capturing meaningful patterns in neuroscientific data. Overcoming these challenges requires specialized interpretability tools and techniques to unravel hidden layer representations and gain deeper insights into neural systems and their computational mechanisms.

When analyzing Magnetoencephalography (MEG) data with six frequency bands as important features, the complexity of MEG connectome images hinders the visual inspection of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) results. Undeterred, the study examines high-accuracy CNN models for dogs, cats, and the MNIST dataset, showcasing the efficacy of LIME and SHAP. This is

achieved through visual inspection and mathematical equations for SHAP values. Crucially, the research aims to demystify the traditionally elusive hidden layers in CNN models, offering insight into their decision-making processes and enhancing overall model transparency.

Across the MEG dataset, the study identifies Gamma1 as having the highest SHAP values, indicating a significant influence on CNN predictions. This nuanced understanding contributes to interpreting the decision-making process of the model and provides insight into the hierarchical influences between frequency bands within the MEG dataset. In conclusion, this study addresses the challenge of interpretability of neuroscience data by unlocking the black box of hidden layers, thereby promoting informed decision making in neuroscience applications.

Keywords: XAI, CNN, Neuroscience, LIME, SHAP, Deep learning Interpretation, MNIST, MEG, Black box.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Term	Definition
AI	Artificial Intelligence
CNN	Convolutional Neural Networks
DCNN	Deep Convolutional Neural Networks
DL	Deep Learning
ECG	Electrocardiogram
EEG	Electroencephalogram
EOG	Electrooculogram
FC	Functional Connectivity
FGT	Fibroglandular Tissue
FMRI	Functional Magnetic Resonance Imaging
GCN	Graph Convolutional Network
GPU	Graphics Processing Unit
GRAD-CAM	Gradient-Weighted Class Activation Mapping
LIME	Local Interpretable Model-Agnostic Explanations
MEG	Magnetoencephalography
ML	Machine Learning
MNIST	Modified National Institute Of Standards And Technology
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography

Term	Definition
RNN	Recurrent Neural Networks
SHAP	Shapley Additive Explanations
TPU	Tensor Processing Unit
VGG16	Visual Geometry Group from Oxford
XAI	Explainable Artificial Intelligence

Chapter One: Thesis Overview

1.1 Introduction

We live in a world where data is everywhere, influencing how we see and interact with our reality. Three key elements make this possible: Data Science, Big Data, and Artificial Intelligence (AI). They work together to discover hidden patterns, reveal important insights, and drive innovation. One of the most amazing fields where this happens is healthcare, especially in neuroscience.

Consider the sheer volume of data we generate daily – from electronic health records to genetic information, it's an overwhelming amount that traditional tools struggle to handle (Manyika et al., 2011). That's where Data Science steps in, armed with mathematics, computing power, and ingenious problem-solving to extract meaning from this data deluge (Russell, 2021).

Artificial Intelligence, often referred to as AI. Imagine it as the brain behind the magic – using specialized computer programs, it learns from data, identifies intricate patterns, and makes decisions much like humans (Russell, 2021). When Data Science and AI combine forces, they unlock the potential to reveal remarkable insights from data, such as predicting the onset of diseases.

Recent years have seen remarkable advancements in healthcare, particularly within the realm of neuroscience, courtesy of AI.

For instance, AI can analyze brain images with incredible precision and detect early signs of neurological disorders like Alzheimer's or Parkinson's disease (Manyika et al., 2011). This breakthrough empowers healthcare professionals to initiate treatment at an earlier stage, significantly improving patient outcomes.

However, AI, especially within a framework called Convolutional Neural Networks (CNNs), can sometimes appear as a "Black Box." In other words, it performs admirably, but understanding its inner workings can be elusive. In neuroscience, where precision and interpretability are paramount, unraveling this mystery is crucial (Tirumala, 2021).

In conclusion, the collaboration of Data Science, Big Data, and AI is revolutionizing healthcare, particularly within neuroscience. This introduction has laid the groundwork for delving deeper into these transformative technologies. Additionally, it underscores the significance of comprehending AI, particularly in the intricate world of CNNs, as we aspire to gain a better understanding of the brain and enhance patient well-being.

1.2 Problem Statement

In recent years, deep learning models have made significant advancements in the field of neuroscience, offering unprecedented insights by leveraging the complexity of their hidden layers. However, extracting, interpreting, and discerning the importance of features within these layers remains a challenge. Furthermore, understanding how these features influence the decision-making processes of CNN is crucial. This study delves into **'Examining the**

Features in Hidden Layers of Deep Learning Models Applied to Neuroscience Data.' By employing Explainable Deep Learning Techniques, our aim is to bridge the interpretability gap, offering a clearer understanding of how these models process, analyze, and are influenced by neurological data features. Such insights are critical not only for the transparency of the models but also for validating and potentially refining the findings they produce.

However, our proposed methodology in explaining the model for Neuroscience Data will make use of a novel approach.

1.3 Research Objectives

In the pursuit of deciphering the intricacies of CNNs when applied to neuroscience data, this research aims to dive deep into the hidden layers to extract and understand their feature roles and representations (Tirumala, 2021). A pivotal aspect of our inquiry is evaluating the significance and contribution of each unearthed feature to CNN's overarching performance and decision-making processes (Zhang & Gao, 2020). By harnessing Explainable Deep Learning Techniques, we aspire to amplify the transparency and understanding of CNNs, ensuring that their analysis of neuroscience data stands up to analysis by experts in the domain (Salahuddin et al., 2021). This deep dive will also enable an assessment of the direct impact of these extracted features on CNN decisions, potentially illuminating biases or areas ripe for model improvement (Alelyani, 2021). Furthermore, armed with insights about feature importance and their decisional effects, there's an opportunity to refine these models, ensuring their efficacy while preserving interpretability (Daniele et al., 2023). A benchmark comparison, contrasting the performance and explainability of CNNs before and after the

application of our techniques, will conclude our research trajectory, establishing the robustness of our chosen methods (Brandt et al., 2023; Daniele et al., 2023).

1.4 Significance of the study

1.4.1 Introduction to Deep Learning

Deep Learning, a subset of machine learning, has taken the world of artificial intelligence by storm in recent years. Utilizing multiple layers of neural networks, it excels in modeling and processing large datasets, especially those with complex hierarchies or that are unstructured in nature. The architecture of deep learning is inspired by the structure of the human brain, specifically the neural networks within it, allowing computers to recognize patterns and make decisions in a manner analogous to human cognition (Tirumala, 2021).

The fundamental building block of deep learning is the artificial neuron or node, inspired by biological neurons. These artificial neurons are organized in layers: an input layer, multiple hidden layers, and an output layer. The depth, which refers to the number of hidden layers, is where deep learning gets its name (Salahuddin et al., 2021). The intricate web of connections and the depth of these networks enable the processing of data with high levels of abstraction, allowing for nuanced feature detection and data representation (Salahuddin et al., 2021).

CNNs and Recurrent Neural Networks (RNNs) are among the prominent architectures within deep learning, each tailored for specific types of data and tasks. For instance, CNNs are especially powerful for image processing due to their unique architecture that identifies

spatial hierarchies in images (Alelyani, 2021). On the other hand, RNNs excel in tasks requiring the analysis of sequences, such as time series prediction or natural language processing (Salahuddin et al., 2021).

As promising as deep learning is, it is not without its challenges. One of the main criticisms is the "Black Box" nature of these models. While they can achieve remarkable accuracy, understanding the internal workings and decision-making processes can be daunting, leading to a call for more transparent and explainable models (Brandt et al., 2023).

1.4.2 Explaining deep learning in neuroscience research.

Deep learning, while a powerful tool, is often likened to a "Black Box" due to the uncleaer nature of its internal workings. The algorithms can process vast amounts of data and produce results with incredible accuracy, yet the exact manner in which they derive these conclusions remains obscured even to experts (Tirumala, 2021). This lack of transparency poses challenges in trust, reliability, and applicability.

In many fields, especially those with high stakes like medicine or finance, just getting the right answer isn't enough. Stakeholders need to understand how a model arrived at its conclusion to ensure its decisions are reliable and not based on unintended biases or anomalies (Salahuddin et al., 2021). Explaining deep learning, therefore, isn't a luxury but a necessity, ensuring these powerful models are used responsibly, ethically, and effectively.

Having established the importance of transparency in deep learning, its intersection with neuroscience becomes particularly intriguing. Neuroscience, dedicated to unraveling the mysteries of the nervous system, is a field that greatly benefits from computational tools capable of handling the complexity of neural data. With the surge in available neural data, notably from advancements like brain imaging, the application of deep learning has grown exponentially (Zhang & Gao, 2020).

Models like CNNs are instrumental in processing imaging data such as Functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET) scans, identifying brain activity patterns or mapping neurological conditions. On the other hand, RNNs shine in analyzing time-series neural data, predicting neural responses to stimuli or mapping disease progression (Alelyani, 2021).

One of the most promising areas is neural decoding, where deep learning attempts to predict or even reconstruct stimuli from neural activity. By training on vast datasets of paired neural activity and stimuli, there's potential to 'reverse engineer' the brain's information processing (Brandt et al., 2023).

Yet, in such sensitive research, the need for model transparency is heightened. Unraveling the brain's secrets with a "Black Box" can be counterproductive. Thus, the push for explainable deep learning models in neuroscience ensures that insights are not just accurate but also understandable and reliable.

Chapter Two: Background and Related Work

2.1 Introduction:

This literature review aims to provide a comprehensive understanding of the landscape of interpretability of deep learning models, focusing on their application to neuroscience data, specifically magnetoencephalography (MEG) data. The review is structured around the core research objectives, which include clarifying the use of Convolutional Neural Networks (CNNs) in image classification, improving the interpretability of CNNs on the pet's dataset (dogs and cats) and MNIST dataset, and exploring the interpretability of CNNs when analyzing MEG data (Alzubaidi et al., 2021a; Sarker, 2021a). The primary focus is on the application of interpretability techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) in these contexts.

2.2 Theoretical Background

2.2.1 Introduction to Deep Learning

Deep Learning (DL) (LeCun et al., 2015) is a type of machine learning (ML) (Murdoch et al., 2019) that uses neural networks with multiple layers to automatically extract hierarchical representations from raw data without manual feature engineering as shown in Figure 1. Unlike classical ML techniques, DL models can discover complex patterns and features by abstracting relevant features through successive layers of computation. Deep learning models can model complex relationships within data, provided there is sufficient labeled data and significant computational resources for training (Sarker, 2021b).

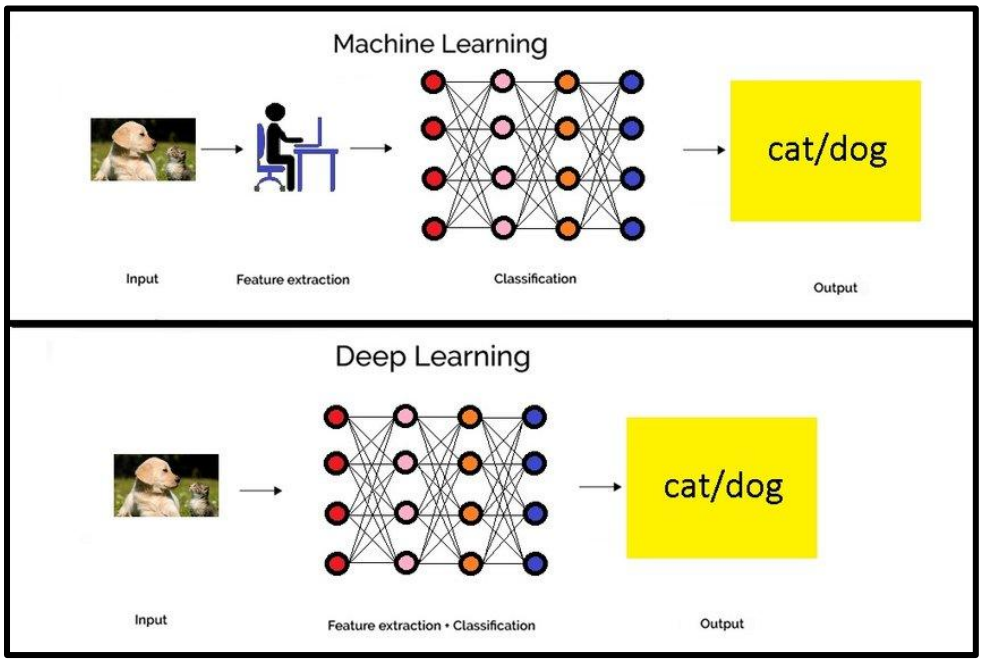


Figure 1 The difference between ML and DL for image classification.

The training process of deep learning involves optimization algorithms such as gradient descent. Optimization algorithms are mathematical techniques used to adjust the parameters of a deep learning model in order to minimize the difference between the predicted output and the actual output, which often requires specialized hardware such as graphics processing unit (GPU)s or Tensor Processing Unit (TPU)s, these hardware accelerators are designed to perform parallel computations and speed up the training process. (Sarker, 2021b). Despite their powerful capabilities, deep learning models are often considered opaque, making their decision-making processes challenging to interpret compared to the more transparent nature of classical machine learning models. Transfer learning, a strength of deep learning, enables the fine-tuning of pre-trained models on related tasks, leveraging learned representations (Nishani & Çiço, 2017).

Applications of deep learning span diverse domains. In computer vision, deep learning excels in tasks such as image recognition, object detection, and segmentation. DL is used in various applications such as sentiment analysis, machine translation, chatbots, disease diagnosis, drug discovery, fraud detection, algorithmic trading, robotics, gaming, recommendation systems, and generative models for content creation. It also plays a crucial role in anomaly detection across various domains, including cybersecurity. The success of DL models has revolutionized industries by pushing the boundaries of what is achievable in machine learning. However, it also presents challenges such as interpretability and increased computational requirements (Goodfellow et al., 2016; Sarker, 2021b; Tirumala, 2021).

The fundamental building block of deep learning is the artificial neuron or node, inspired by biological neurons. These artificial neurons are organized into layers: an input layer, several hidden layers, and an output layer. The depth, which refers to the number of hidden layers, is where deep learning gets its name (Salahuddin et al., 2021). The intricate web of connections and the depth of these networks allow data to be processed at a high level of abstraction, enabling nuanced feature extraction and data representation (Salahuddin et al., 2021).

CNNs and RNNs are among the prominent architectures within deep learning, each tailored for specific types of data and tasks. For instance, CNNs are particularly powerful for image processing due to their unique architecture that identifies spatial hierarchies in images (Alelyani, 2021). On the other hand, RNNs excel in tasks requiring the analysis of sequences, such as time series prediction or natural language processing (Salahuddin et al., 2021).

2.2.2 Convolutional neural networks

CNNs is a specialized type of deep learning neural network designed for tasks such as image and video analysis. Key features of CNNs include hierarchical feature learning, convolutional layers, pooling layers, and a fully connected layer at the end of the network as shown in figure 2. CNNs have been widely used for image recognition and classification tasks, such as distinguishing between dogs and cats, due to their ability to learn and recognize complex patterns and shapes in images. CNNs consist of several layers, such as convolutional layers, pooling layers, and fully connected layers, which perform different operations on the input images. Convolutional layers apply filters to the images and produce feature maps that capture the local information in the images. Pooling layers reduce the size and complexity of the feature maps by applying a function, such as max or average, to a region of the feature map. Fully connected layers connect all the neurons from the previous layer to the output layer, where the final classification is performed (Dhillon & Verma, 2020).

In Figure 2, the input layer, convolutional layer, and pooling layer are employed for feature extraction, whereas the fully connected layer and output layer at the end of the CNN network are dedicated to classification.

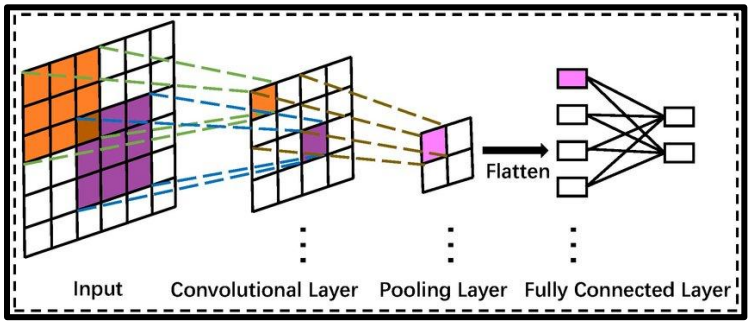


Figure 2: Convolutional Neural Network (CNN) layers (Lu et al., 2018).

CNNs are designed to learn spatial hierarchies of features automatically and adaptively from input data. This ability makes them particularly effective in image analysis tasks (Singh & Kumar, 2023), where local patterns and relationships are crucial. The convolutional layers, a fundamental component of CNNs, allow the network to detect different features in the input, such as edges, textures, and more complex structures (Yamashita et al., 2018).

One key advantage of CNNs lies in their parameter sharing mechanism. In convolutional layers, filters (also known as kernels) are applied across the entire input, sharing weights to detect similar patterns across the image. This parameter sharing not only significantly reduces the number of parameters, making the network more efficient, but also increases the network's ability to generalize and recognize patterns in different spatial locations (Szegedy et al., 2014).

Furthermore, the use of pooling layers adds another layer of abstraction to the learned features. Pooling helps to down sample the spatial dimensions of the input, reducing computational complexity while retaining essential information. Common pooling operations include max pooling, which selects the maximum value from a region, and average pooling, which calculates the average. This downscaling helps to make the learned features more robust to variations in scale and orientation (Simonyan & Zisserman, 2014a).

CNNs have shown remarkable performance in several computer vision tasks beyond image classification. They excel in tasks such as object detection, where the network not only classifies the objects in an image but also localizes and delineates their boundaries. Another application is semantic segmentation, where CNNs assign a class label to each pixel in an image, providing a detailed understanding of the scene's content (L.-C. Chen et al., 2014; Long et al., 2015).

Another notable aspect of CNNs is transfer learning. Pre-trained models on large datasets, such as ImageNet (Mishkin et al., 2016), can be fine-tuned for specific tasks using smaller datasets. This leverages the knowledge gained from the broader dataset, leading to improved performance, particularly when labelled data for a specific task is limited.

In recent years, CNN architectures have evolved with advances such as residual connections, attention mechanisms, and capsule networks (Sabour et al., 2017), addressing challenges like vanishing gradients and allow for better capture of long-range dependencies in images. These developments contribute to the continuous refinement and improvement of CNNs, making

them even more powerful tools in the field of computer vision (Simonyan & Zisserman, 2014b).

Mathematical modelling plays a crucial role in understanding the inner workings of CNNs. Let's look at the mathematical aspects of key components such as convolutional layers and pooling layers.

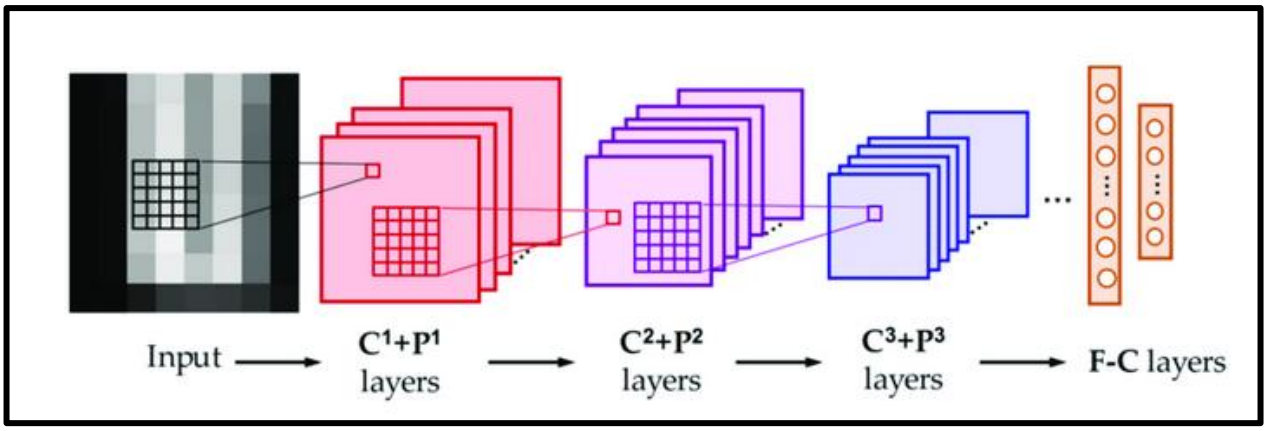


Figure 3 The structure of a traditional convolutional neural network (CNN) (Chu et al., 2020). C and P denote a convolutional and pooling layer block, respectively and included the layer number. FC denotes a fully-connected layer.

CNNs are comprised of convolutional layers and pooling layers arranged in an alternating configuration to extract features, along with fully connected layers dedicated to constructing semantic features. Figure 3 illustrates a conventional two-dimensional CNN and its feature extraction process within three different layers (Chu et al., 2020).

CNNs leverage the power of convolutional processes through the application of specialized filters on input data, thereby unlocking a myriad of advantages (Jang, 2023; Zeiler & Fergus, 2013). One notable benefit lies in their ability to capture spatial hierarchies and patterns

within the data. By employing convolutional layers, CNNs systematically scan input information with filters that detect specific features, enabling the network to discern complex spatial relationships. This capability proves particularly valuable in image recognition tasks, where CNNs excel at recognizing patterns like edges, textures, and shapes (Guo et al., 2016; G. Huang et al., 2016). The hierarchical nature of the convolutional process further enables the network to learn increasingly abstract representations as it progresses through deeper layers, enhancing its ability to understand and classify intricate structures within the input data. This hierarchical feature extraction not only contributes to improved accuracy but also enhances the network's efficiency in handling large-scale datasets, making CNNs indispensable in various fields, including computer vision, image processing, and pattern recognition (Guo et al., 2016; G. Huang et al., 2016).

2.2.2.1 Convolutional Layers

In a convolutional layer, the input image (or feature map from the previous layer) is convolved with a set of filters or kernels. The convolution operation can be represented mathematically as follows:

$$H_{i,j}^{(l)} = f\left(\sum_{m=0}^{k-1}\sum_{n=0}^{k-1} I_{i+m,j+n}^{(l-1)} \cdot W_{m,n}^{(l)} + b^{(l)}\right) \quad (1)$$

Here,

- $H_{i,j}^{(l)}$ is the value at position (i,j) in the feature map of the $l-th$ layer.

- $I_{i+m,j+n}^{(l-1)}$ is the value at position $(i+m,\ j+n)$ in the feature map of the $(l-1)-th$ layer.

- k is the size of the convolutional filter used in the convolution operation.

- $W_{m,n}^{(l)}$ is the weight at position (m, n) in the $l\ -th$ layer's filter.

- $b^{(l)}$ is the bias term for the $l\ -th$ layer.

- f refers to an activation function.

- i and j refer to positions in the output feature map.

- m and n are indices used for the convolutional kernel, representing positions within the kernel.

The convolution operation involves sliding the filter over the input and computing the dot product at each position, followed by applying the activation function (Alzubaidi et al., 2021b; Goodfellow et al., 2016).

The size of the convolution filter plays a critical role in determining the receptive field of the convolutional operation within convolutional neural networks (CNNs). It directly influences the spatial information captured from the input data. Larger filter sizes encompass more extensive areas of the input, enabling the network to extract complex features spanning multiple pixels. However, this comes at the cost of increased computational complexity and the potential for overfitting if not appropriately managed. Hence, selecting an optimal filter size is crucial for balancing the trade-off between model complexity and feature extraction capability (LeCun et al., 2015).

In CNNs, the weights and biases associated with convolutional filters are fundamental parameters that dictate the network's ability to learn and extract meaningful features from the input data. The weights define the learned features, while the biases provide flexibility by

allowing the network to shift activation functions' outputs. Through iterative training processes such as backpropagation, these parameters are adjusted to minimize the error between predicted and actual outputs. However, improper initialization or learning of weights and biases can lead to suboptimal performance and hinder the network's generalization ability (Springenberg et al., 2014).

The influence of weights and biases on CNN performance and interpretability is profound. Properly trained parameters enable CNNs to achieve high accuracy in various tasks such as image classification, object detection, and segmentation. However, their interpretation is equally crucial for understanding how the network makes decisions, particularly in applications where model transparency and interpretability are essential, such as medical diagnosis or autonomous systems. Careful consideration and management of weights and biases are therefore indispensable for ensuring the efficacy and interpretability of CNNs in real-world applications (Krizhevsky et al., 2017).

2.2.2.2 Pooling Layers

Pooling layers are used to downsample the spatial dimension of the feature maps. Max pooling is a common pooling operation, and can be expressed mathematically as:

$$P_{i,j}^{(l)} = max_{m,n}\,(H_{i.s+m,j.s+n}^{(l-1)}) \quad (2)$$

Here,

- $P_{i,j}^{(l)}$ is the value at position (i,j) in the pooled feature map of the $l\ -th$ layer.

- s is the stride, specifying the step size of the pooling operation (Goodfellow et al., 2016).

- (i, j) represent positions in the output feature map after max-pooling.

- (m, n) are indices used for iterating over the pooling window in the input feature map.

- s is the stride, defining the step size for moving the pooling window during the operation.

2.2.2.3 Fully Connected Layers

Fully connected layers connect neurons from the previous layer to the output layer. The output of a neuron in a fully connected layer can be expressed as:

$$O_i^{(L)} = f(\sum_{j=1}^{N^{(L-1)}} I_j^{(L-1)}.W_{ij}^{(L)} + b_i^{(L)}) \quad (3)$$

Here,

- L: Represents the index of the current layer where the computation is taking place. In the context of neural networks, the layers are typically numbered starting from 1 for the input layer, increasing consecutively through the hidden layers, and reaching the output layer.

- $O_i^{(L)}$ is the output of the $i-th$ neuron in the output layer.

- $I_j^{(L-1)}$ is the output of the $j\ -th$ neuron in the $(L-1)-th$ layer.

- $W_{ij}^{(L)}$ is the weight connecting the $j\ -th$ neuron to the $i-th$ neuron in the output layer.

- $b_i^{(L)}$ is the bias term for the i-th neuron in the output layer (Goodfellow et al., 2016).

These equations provide insight into the mathematical foundations of CNNs, where weights and biases are learned through backpropagation during the training process. The activation function f introduces non-linearity, allowing the network to capture complex relationships

within the data (Alzubaidi et al., 2021b; Goodfellow et al., 2016; Long et al., 2015). The overall architecture and success of CNNs in image-related tasks are deeply rooted in these mathematical formulations.

2.2.3 Visual Geometry Group from Oxford (VGG16) Model

The VGG16 (Visual Geometry Group 16) (Mahardi et al., 2020) CNN model is a prominent architecture in the field of computer vision, specifically designed for image classification tasks. It was introduced by the Visual Geometry Group at the University of Oxford and has gained widespread popularity due to its simplicity and effectiveness (Simonyan & Zisserman, 2014c).

VGG16 is characterized by its deep architecture, consisting of 16 weight layers, hence the name. The core building blocks of the model are 3x3 convolutional layers, with max-pooling layers interspersed to downsample the spatial dimensions of the input. The repeated stacking of small-sized filters allows the network to capture complex hierarchical features in a more efficient manner compared to larger filters. The model's architecture follows a straightforward and uniform pattern, making it easy to understand and implement (Olisah et al., 2024).

One distinctive feature of VGG16 is its use of a simple and uniform structure throughout the network. The convolutional layers are configured with small receptive fields and a fixed-size 3x3 filter, maintaining a consistent stride and padding. This regularity contributes to the ease of training and enables the model to learn hierarchical features effectively. The deep

architecture of VGG16 facilitates the extraction of intricate patterns and representations, making it well-suited for image recognition tasks (Mahardi et al., 2020; Simonyan & Zisserman, 2014c).

The VGG16 model achieved state-of-the-art performance on various benchmark datasets at the time of its introduction, showcasing the effectiveness of deep convolutional neural networks in image classification. Its impact on the field is evident in subsequent research and model architectures. Researchers and practitioners continue to draw inspiration from VGG16, adapting and extending its principles for a wide range of computer vision applications (Olisah et al., 2024).

2.2.4 Convolutional Neural Network Interpretability

Convolutional Neural Networks (CNNs) have emerged as powerful tools for solving complex computer vision tasks, with remarkable success in image classification, object detection, and segmentation (Tirumala, 2021). However, the intricate and opaque nature of these deep neural networks poses challenges in terms of interpretability, making their adoption in critical applications such as healthcare, finance, and autonomous systems very difficult. This discourse explores the limitations of CNN models in terms of interpretability and highlights the indispensability of interpretability techniques such as LIME and SHAP (Chu et al., 2020; Ribeiro et al., 2016).

Deep Learning models such as CNNs are often referred to as "black box" models due to their inherent complexity, which renders their internal workings inscrutable. Hierarchical learning

of features within the network is automatic, making it difficult for users to understand the rationale behind certain decisions. As the depth and complexity of the model increases, the interpretability of CNNs decreases, leading to a lack of transparency in decision-making processes (Buhrmester et al., 2021).

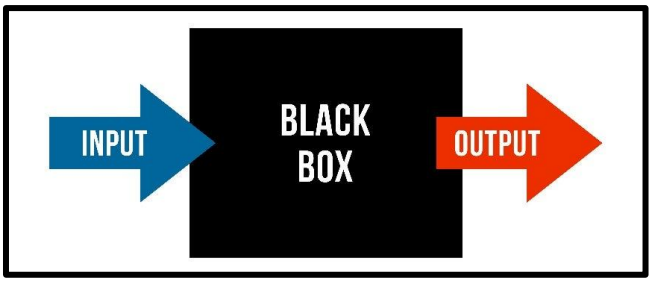


Figure 4 CNN Model as Black Box with input and output (A. Huang et al., n.d.)

In addition, the processing of high-dimensional input data, such as images, is a fundamental aspect of CNNs. However, this very characteristic poses a significant challenge in tracking how specific features or pixels contribute to the final decision. The high dimensionality of the input data exacerbates the difficulty of understanding the decision boundaries of the model, impeding efforts to gain insights into its functioning (Ventura et al., 2018).

2.2.5 Local and Global Explanation of CNN Models

Local explanation of CNN models is a critical aspect to improve the interpretability and transparency of these complex deep learning architectures. One prominent technique for local explanation is LIME (Ribeiro et al., 2016). LIME works by troubling input data instances and observing the corresponding model predictions, and then fitting a locally interpretable

model to approximate the behavior of the underlying CNN in the vicinity of the instance of interest. This method allows for a more explicit understanding of the CNN's decision-making process on a case-by-case basis, helping users to understand why a specific prediction was made (Shah & Sheppard, 2020a).

LIME is particularly effective in the context of image-based CNNs, where understanding the features that contribute to a prediction is critical. By generating perturbed samples and observing the resulting predictions, LIME constructs a simpler, more interpretable model that provides insight into the CNN's decision boundary for a given input. This local interpretability is valuable in scenarios such as medical imaging, where understanding the factors influencing a diagnosis can be critical (Kun et al., 2022).

In addition to LIME, there are several other eXplainable AI (XAI) techniques that contribute to the local explanation for CNN models. SHAP is one such technique, which assigns a value to the contribution of each feature to a particular prediction. This enables a finer understanding of the impact of individual features on the model's output. Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2016) focuses on visualizing the importance of different regions in an input image by exploiting the gradients of the target class with respect to the model's convolutional feature maps (Kun et al., 2022; Shah & Sheppard, 2020a).

Interpretable models, such as decision trees and rule-based systems, also play a role in local explanations for CNNs. These surrogate models serve as simplified approximations of the CNN's decision boundaries and help interpret specific predictions (Ribeiro et al., 2016).

Global explanation of CNN models is essential to gain a holistic understanding of the model's behavior across the entire dataset. SHAP is a powerful technique that provides insight into the overall feature importance and the contribution of each input feature to the model's predictions. SHAP values are rooted in cooperative game theory, specifically Shapley values, and provide a principled approach to assigning a value to each feature based on its impact on the prediction across all possible combinations of features (Murdoch et al., 2019).

The application of SHAP in the context of CNNs allows the quantification of feature importance on a global scale. By aggregating SHAP values across the entire dataset, it is possible to identify which features consistently influence model predictions. This global perspective is particularly valuable for understanding the general behavior of CNN and for identifying patterns and trends in feature importance that may not be apparent in a local context.

In addition to SHAP, several other explainable AI (XAI) techniques contribute to the global explanation of CNN models. Layer-wise relevance propagation (LRP) (Miller, 2017) is another method that assigns relevance scores to different layers of the neural network, shedding light on the contribution of each layer to the final prediction. Sensitivity analysis,

in which input features are systematically varied and the effect on model predictions is observed, provides a global perspective on feature sensitivity (Miller, 2017).

Ensemble methods, such as model stacking and bagging, also contribute to global explanations by combining multiple models to improve overall predictive performance. The interpretation of an ensemble model can provide insights into the diversity of individual models and their collective decision-making.

2.2.6 Local Interpretable Model-agnostic Explanations (LIME)

LIME, or Local Interpretable Model-agnostic Explanations, is a method designed to provide faithful explanations for the predictions made by any classifier or regressor. It achieves this by locally approximating the black-box model with an interpretable model. The algorithm alters individual data samples by adjusting their feature values and assesses the ensuing changes in output. Acting as an "explainer," LIME explains predictions for each data sample. The result is a collection of explanations that delineate the influence of each feature on the prediction for a specific sample, embodying a form of local interpretability. (Ribeiro et al., 2016).

LIME can be used to better explain the predictions of a Convolutional Neural Network (CNN) model on images such as discriminating between images of dogs and cats. The CNN model is trained on a dataset of images of dogs and cats, and it learns to distinguish between the two classes. LIME can be used to explain the predictions of this model by locally approximating it with an interpretable model, such as a linear regression or decision tree model (Ribeiro et al., 2016).

The mathematical equation for LIME is as follows:

$$explanation(x) = \ argmin_{g \in G} \ L(f, g, \pi_x) + \Omega \ (g) \tag{4}$$

where f is the model to be explained, g is the interpretable model, π_x is the distribution over the training data, L is the loss function, and Ω is the complexity penalty. The goal of LIME is to find an interpretable model g that approximates the behavior of the original model f in the vicinity of the data point of interest (Ribeiro et al., 2016).

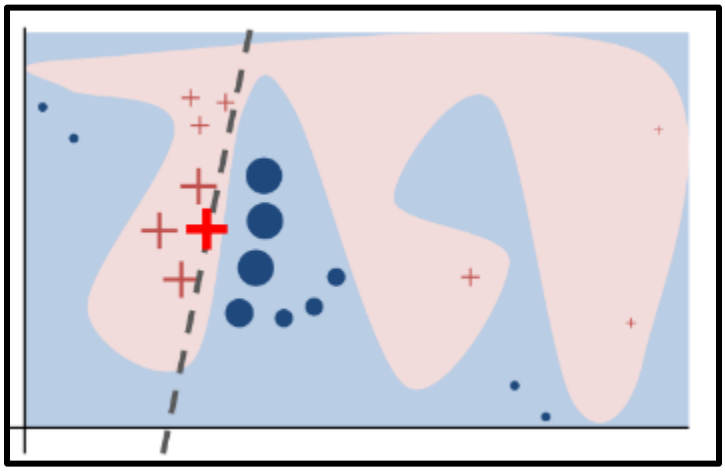


Figure 5 Present intuition for LIME (Ribeiro et al., 2016)

In Figure 5, the background represents the complex decision function f of the black-box model in blue/pink. Challenged by its non-linearity, LIME aims to approximate it. The bold red cross indicates the instance to be explained. LIME samples instance predicts with f, and weights them by proximity (indicated by size). The resulting dashed line signifies the locally faithful explanation generated by LIME, which effectively captures the complex decision boundary around a particular instance (Ribeiro et al., 2016).

The workflow of LIME involves selecting an instance from the dataset, perturbing it to generate a set of similar but slightly modified instances, and then obtaining predictions from the black-box model for these perturbed instances (Ribeiro et al., 2016). These predictions, along with the corresponding perturbed instances, are used to train a simple, interpretable model (e.g., linear regression) that locally approximates the behavior of the black-box model. The coefficients of this interpretable model serve as explanations for the black-box model's prediction at the selected instance (Datta et al., 2016).

Mathematically, let $f: X \rightarrow Y$ be the black-box model, where X is the input space and Y is the output space. Given an instance $x^0 \in X$ and a perturbed dataset $\{(x_i, f(x_i))\}_{i=1}^{n}$ LIME aims to find an interpretable model $Z \rightarrow Y$ that approximates f locally around x_0 where Z is the space of interpretable features. The interpretable model g is often a simple linear model, and its coefficients provide insights into the contribution of different features in the prediction and n is the number of perturbed instances in the dataset. (Ribeiro et al., 2016).

Advantages of LIME include its model-agnostic nature, which allows it to be applied to different machine learning models without requiring knowledge of their internal architectures. It provides human-understandable explanations for individual predictions, which is crucial in applications where interpretability is essential, such as healthcare or finance (Zafar & Khan, 2021).

However, LIME has certain limitations. The choice of perturbation and sampling strategies can affect the explanations, and the local nature of the interpretation might not accurately capture the global model behavior. In addition, LIME may struggle with high-dimensional data, and the quality of the explanations may be sensitive to the choice of hyperparameters (Zhao et al., 2020).

When applied to CNN for image classification, visual inspection of LIME results is crucial. Explanations provided by LIME can be visually inspected to ensure that they are consistent with human intuition. For instance, in the context of distinguishing between images of dogs and cats, a valid explanation would highlight the presence of distinct features such as ears, tails, or fur patterns that are indicative of the predicted class. Visual inspection helps to validate the fidelity and relevance of LIME explanations, making them more trustworthy in real-world applications (Gashi et al., 2022).

2.2.7 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) is a powerful technique for explaining the output of machine learning models, providing insights into feature contributions on both local and global levels. SHAP values are based on the Shapley values of cooperative game theory, which aim to distribute a value equitably across a group of contributors (S. Lundberg & Lee, 2017a).

SHAP techniques offer several advantages over traditional methods in feature extraction and feature importance analysis. Firstly, their interpretability and transparency stand out. Unlike

conventional approaches that often produce opaque models, SHAP values provide clear insights into how each feature influences a machine learning model's prediction. This transparency is crucial, especially in critical decision-making scenarios such as loan approvals, where stakeholders require explanations for model outcomes. Secondly, SHAP employs a model-agnostic approach, making it compatible with various machine learning algorithms ranging from linear regression to neural networks. This flexibility is absent in many traditional feature selection techniques, which are often tailored to specific models. Thirdly, SHAP techniques excel in handling high-dimensional datasets efficiently without sacrificing interpretability. Unlike some traditional methods that may struggle with large feature sets, SHAP can effectively manage complex data structures. Finally, SHAP values offer objective importance scores based on game theory principles, ensuring consistent evaluation of feature importance across different models and datasets. In conclusion, SHAP techniques provide a robust and versatile framework for enhancing model interpretability, presenting a compelling alternative to conventional feature selection and importance methods (Marcilio & Eler, 2020).

In the context of a CNN discriminating between images of dogs and cats, and also applied to the MNIST dataset for handwritten numbers, SHAP operates by evaluating the impact of each feature (e.g., pixels in an image) on the model's prediction. To obtain a measure of global importance, we compute the average absolute Shapley values for each feature over the entire dataset:

$$I_j = \frac{1}{n}\sum_{i=1}^{n}\left|\emptyset(i)_j\right| \quad (5)$$

Here, I_j represents the global importance of feature J, and $\emptyset(i)_j$ denotes the Shapley value for feature j in the $i-th$ instance and n is the number of instances in the dataset over which the average absolute Shapley values are computed. It helps in obtaining a measure of the global importance of each feature by considering the Shapley values across all instances in the dataset.(S. Chen, 2021).

One of the advantages of SHAP is its ability to provide both local and global interpretability. Locally, it can explain the prediction for a specific instance, while globally, it provides insight into overall behavior of the model across the entire dataset. This makes it a versatile tool for understanding model decisions (Nohara et al., 2021).

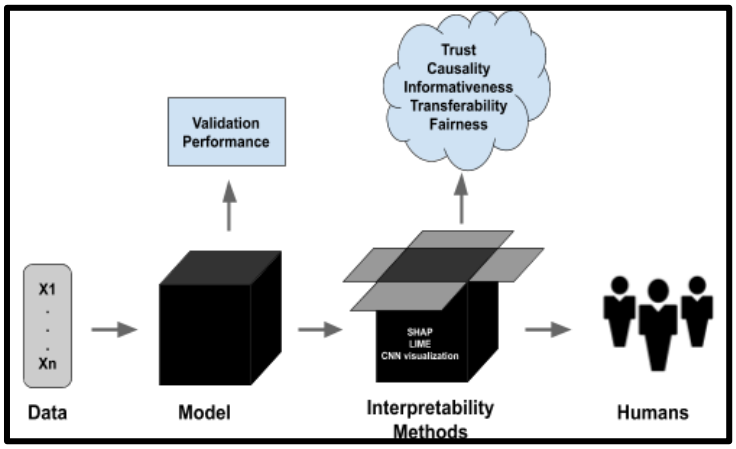


Figure 6 The interpretation of CNN using LIME and SHAP enhances the trustworthiness of the model's predictions for humans. (A. Huang et al., n.d.)

Visual inspection plays a crucial role in validating SHAP results. For images of dogs and cats, SHAP values can be visualized as heatmaps overlaid on the original image, highlighting the pixels that contribute the most to the model's decision. In the case of the MNIST dataset,

visualization can show the importance of each pixel in determining the predicted digit. These visualizations provide an intuitive understanding of the model's reasoning (Krause et al., 2016).

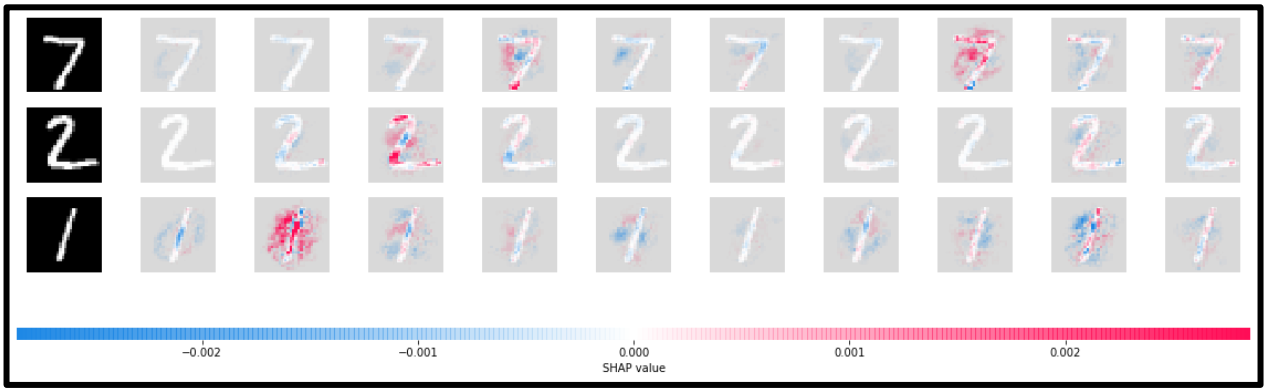


Figure 7An illustration of elucidating Convolutional Neural Networks (CNNs) through the application of SHAP (Shapley Additive exPlanations) on the MNIST dataset. (Scott Lundberg, 2018)

Figure 7 illustrates the SHAP values for the MNIST dataset, showcasing a sample of three digits. The first column displays the original images, while the subsequent 10 columns provide SHAP values for each class ranging from 0 to 9. In this visual representation, elevated SHAP values are represented by an abundance of red pixels, while negative SHAP values are depicted in blue. The blue and red regions in the visualization explain the model's predictions, offering insights into both correct and incorrect predictions and elucidating the underlying reasons for these outcomes.

However, there are some drawbacks to consider. Computationally, the calculation SHAP values can be resource-intensive, especially for complex models or large datasets (Molnar et al., 2020). Furthermore, the interpretation of SHAP values relies on the assumption that

model behavior can be accurately represented as an additive combination of feature contributions, which may not always be true (S. Lundberg & Lee, 2017b).

In terms of global interpretation, SHAP values can be aggregated across all instances in the dataset to reveal the overall influence of each feature. This aggregated view helps to identify features that consistently influence the model across different data points, contributing to a more comprehensive understanding of the model's decision-making process on a global scale (C. Huang et al., 2017; Molnar et al., 2020).

2.2.8 SHAP Values vs Prediction Accuracy vs Features Importance

Shapley values, prediction accuracy, and feature importance are key concepts in machine learning for understanding and evaluating models.

Shapley values, originating from cooperative game theory, are employed in machine learning to equitably assign the impact of individual features on predicting a particular instance. They consider all possible feature combinations and calculate the average marginal contribution of each feature, aiding in understanding individual feature importance, particularly in cases where features work together in a cooperative manner (Lee et al., 2023; Saarela & Jauhiainen, 2021).

Prediction accuracy, a crucial metric, gauges the overall effectiveness of a machine learning model by evaluating the percentage of accurately predicted instances among the total

instances. While it provides a global view of model performance on a dataset, it does not reveal the contribution of each feature to individual predictions (Lee et al., 2023).

Feature importance, another crucial metric, gauges the contribution of each input feature to a model's predictions. Various methods, such as tree-based feature importance, permutation importance, or coefficients in linear models, can be employed to assess feature importance. Unlike Shapley values, these methods may not consider interactions between features and typically assess the influence of each feature independently (S. M. Lundberg & Lee, 2017; Mokhtari et al., 2019).

In other words, Shapley values distribute credit for a prediction among features, prediction accuracy measures overall model performance, and feature importance identifies the influence of individual features. These concepts play distinct roles in comprehending and evaluating machine learning models (S. M. Lundberg & Lee, 2017; Mokhtari et al., 2019; Saarela & Jauhiainen, 2021).

Additionally, (Fryer et al., 2021) emphasizes the significance of feature importance estimation in tree-based ensemble methods. The authors discuss how SHAP values serve as an effective tool for interpreting models and provide insights into feature contributions. The paper compares SHAP values with other methods and highlights their accuracy and consistency.

Similarly, (Marcilio & Eler, 2020) explores SHAP values as a feature selection mechanism, by highlighting the importance of feature selection in machine learning preprocessing, the authors demonstrate that SHAP values outperform other feature selection algorithms in terms of accuracy and consistency. The paper addresses the limitations of traditional feature selection algorithms and emphasizes the utility of SHAP values in overcoming these limitations.

2.2.9 Explaining Deep learning Results in neuroscience research

As promising as deep learning is, it is not without its challenges. One of the main criticisms is the "black box" nature of these models. While they can achieve remarkable accuracy, understanding the inner workings and decision-making processes can be daunting, leading to a call for more transparent and explainable models (Brandt et al., 2023).

The algorithms can process vast amounts of data and produce results with incredible accuracy, but the exact manner in which they derive these conclusions remains opaque even to experts (Tirumala, 2021). This lack of transparency poses challenges in terms of trust, reliability, and applicability.

In many areas, particularly high-risk areas such as medicine or finance, it's not enough to get the right answer. Stakeholders need to understand how a model arrived at its conclusion to ensure its decisions are reliable and not based on unintended biases or anomalies (Salahuddin et al., 2021). Explaining deep learning is therefore not a luxury, it's a necessity to ensure these powerful models are used responsibly, ethically, and effectively.

Having established the importance of transparency in deep learning, its intersection with neuroscience becomes particularly intriguing. Dedicated to unravelling the mysteries of the nervous system, neuroscience is a field that benefits greatly from computational tools that can handle the complexity of neural data. With the increase in available neural data, particularly from advances such as brain imaging, the application of deep learning has grown exponentially (Zhang & Gao, 2020).

Models such as CNNs are essential for processing imaging data such as MEG scans, identifying patterns of brain activity or mapping neurological conditions, predicting neural responses to stimuli or mapping disease progression (Alelyani, 2021).

One of the most promising areas is neural decoding, where deep learning attempts to predict or even reconstruct stimuli from neural activity. By training on large datasets of paired neural activity and stimuli, there's potential to 'reverse engineer' the brain's information processing (Brandt et al., 2023).

However, in such sensitive research, the need for model transparency is heightened. Unraveling the brain's secrets with a "black box" can be counterproductive. Thus, the push for explainable deep learning models in neuroscience ensures that insights are not just accurate but also understandable and reliable (Ventura et al., 2018).

2.3 Literature review: CNN Interpretations

Understanding the performance of CNNs in this section sheds light on the evolving version of CNN. This literature review examines LIME and SHAP, the integrated efficiency strategies. The focus extends beyond the traditional boundaries of accuracy measurement, emphasizing the importance of understanding and interpreting model predictions.

LIME and SHAP emerged as key frameworks that explored semantics, enabling practitioners to bridge the gap between complex, black-box CNNs and human understanding LIME, with local alternative models, though simplified crystals through which specific predictions for SHAP are understood, by assigning values to each component, provides a holistic approach, thereby quantifying their contribution to the model results LIME and integrating SHAP into the training of CNNs not only enhances our understanding of model decisions but also opens up ways to validate the model and refine architectural designs.

This literature review examines studies that use visual inspection as a means of understanding CNN definitions. By explaining how LIME and SHAP generate striking images or reveal key areas in early images, researchers attempt to explain the cognitive processes of CNNs. Visual inspection bridges abstract mathematical explanations and human intuition and facilitates a better understanding of how CNNs discriminate between shapes and features in images.

Alongside visual analysis, mathematical simulation plays an important role in revealing the mechanisms embedded in CNN interpretation. This review examines the statistical foundations of LIME and SHAP and provides insight into how these methods quantify the

contribution of individual components to the overall prediction. By linking visual and mathematical algorithms together, the researchers aim to go beyond mere accuracy analysis to achieve a comprehensive and accurate understanding of CNN versions.

As this research progresses through this literature review, a combination of LIME, SHAP, visual modeling and computational algorithms will become apparent, providing a comprehensive view of the semantic capabilities of CNNs in image classification. By examining both the dimensional and quantitative aspects of model interpretation, this section contributes to the growing discourse on how LIME and SHAP give us the ability to describe and rely on decisions made by CNNs in complex image classification tasks.

Starting the review of LIME. this paper (Garreau & Mardaoui, 2021), studied LIME, a popular interpretability method used to understand how certain predictions are made in computer vision tasks such as object recognition. They show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which they give an explicit expression. They extend this study to elementary shape detectors and linear models. As a consequence of this analysis, they uncover a connection between LIME and integrated gradients, another explanation method. More specifically, the LIME explanations are similar to the sum of integrated gradients over the superpixels used in the preprocessing step of LIME (Garreau & Mardaoui, 2021).

Superpixels are commonly used in computer vision tasks such as image segmentation, object recognition, and image analysis. By grouping pixels into superpixels, the computational load

is reduced, and the resulting segments often align more closely with the boundaries of objects or regions in an image. This makes them particularly useful in tasks where understanding the structure and content of an image is crucial, such as in the preprocessing step of interpretability methods like LIME (Schallner et al., 2020; Wang et al., 2017; Yang et al., 2020).

Furthermore, (Garreau & Mardaoui, 2021), the authors study the image version of LIME and explain how it works. They decompose an example into superpixels and generate a dataset of perturbed examples around the original example. They then train a linear model on the perturbed dataset to approximate the behavior of the underlying model. Finally, they use the linear model to explain the predictions of the underlying model. The authors also show that when the number of generated examples is large, LIME explanations are concentrated around a limit explanation for which they give an explicit expression. They extend this study for elementary shape detectors and linear models. As a consequence of this analysis, they uncover a link between LIME and integrated gradients, another explanation method (Garreau & Mardaoui, 2021).

In addition to the LIME method, this paper (Schallner et al., 2020) explores the effect of different superpixel methods such as Felzenszwalb, SLIC, and Compact-Watershed (these commonly employed image segmentation algorithms are used to divide an image into significant components.) (Schallner et al., 2020) on the generated visual explanations. The authors compare the resulting relevance areas with the image parts marked by a human reference and show that the image parts selected as relevant vary strongly depending on the

method applied. Quick-Shift produced in the lowest and Compact-Watershed in the highest correspondence with the reference relevance areas (Garreau & Mardaoui, 2021; Schallner et al., 2020).

In addition, LIME creates a new dataset containing perturbed samples and the corresponding predictions made by the black box model. Subsequently, LIME then trains an interpretable model on this new dataset, assigning weights based on the proximity of the sampled instances to the target instance. In addition, LIME uses a surrogate model to approximate the behavior of the underlying model. The authors of (Garreau & Mardaoui, 2021) demonstrate that LIME explanations closely resemble the sum of integrated gradients over the superpixels employed in LIME's preprocessing step. They also show a correlation between LIME and integrated gradients, another method of explanation (Molnar, 2022).

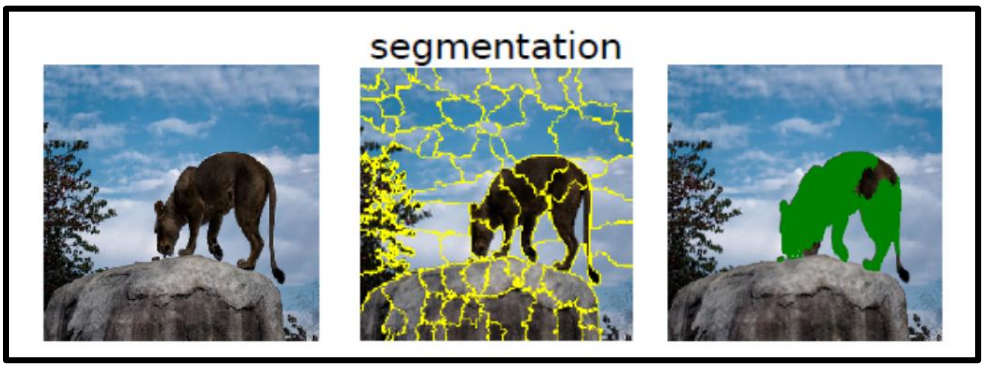


Figure 8 Describing a prediction using LIME, using Segmentation and finding the superpixels (Garreau & Mardaoui, 2021)

Another paper (Shah & Sheppard, 2020b) proposes a framework for evaluating the quality of generated explanations by measuring their sufficiency and salience. The authors propose two experiments to test the quality of generated explanations. The first experiment tests the ability of a trained CNN to correctly classify the generated explanation to test whether it

contains sufficient information for classification. The second experiment tests the salience of the generated explanation by comparing the classification precision and recall of two new CNNs, one using raw image data and the other using explanations as training data, on a common set of test data. The authors use their new evaluation framework to test their hypothesis that LIME and SHAP are able to generate explanations that are both sufficient and salient (Shah & Sheppard, 2020b).

LIME works by fitting a local model using sample data points that are similar to the observation being explained. LIME obtains parameters for feature contribution at the observation level, which means that it provides explanations for the prediction relative to each observation (Shah & Sheppard, 2020b).

Moving on to SHAP, the second technique to be utilized in this research. SHAP is a model-specific method that provides global explanations using a game-theoretic approach. SHAP values are based on Shapley values, which are a way of fairly distributing the total contribution of each feature to the prediction across all possible orders of the features. SHAP values can be used to explain the output of any machine learning model (Shah & Sheppard, 2020b).

Here in this paper (Bhandari et al., 2023) presents a deep learning model that can accurately identify renal abnormalities such as cysts, stones, and tumors. The proposed model is based on a convolutional neural network (CNN) and uses Explainable Artificial Intelligence (XAI) methods such as LIME and SHAP to provide clinicians with conclusive and understandable

results (Bhandari et al., 2023). The main purpose of this paper is to explore the capabilities of a lightweight CNN model in accurately identifying renal abnormalities using LIME and SHAP methods. The authors aim to provide clinicians with conclusive and understandable results (Bhandari et al., 2023).

The authors used a dataset of 1,000 CT images of the kidney, which were collected from different sources. The dataset was divided into training and test sets using a 10-fold stratified sampling method. The proposed model was trained on the training set and evaluated on the test set. The results showed that the proposed model achieved an accuracy of 98.5% in identifying renal abnormalities (Bhandari et al., 2023).

The authors also used LIME and SHAP methods to provide interpretable explanations for the predictions of the proposed model. The results showed that the proposed model could accurately identify the regions of interest in the CT images that contributed to the predictions which can be recognized in Figures 9 and 10, In SHAP analysis, the red regions explain the model results, while in LIME, it is the white region that serves the same purpose (Bhandari et al., 2023), Figure 9 the presence of a tumor was deduced based on the high concentration of red pixels in the fourth explanation image (located in the fifth column) of the CT scan, emphasizing the result derived from the model prediction.

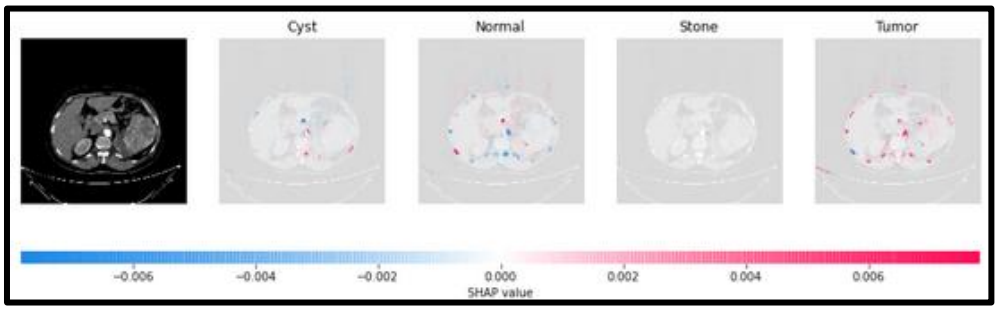


Figure 9 SHAP images illustrating the CNN model's explanation for classifying kidney conditions, including Cysts, Stones, or Tumors, and based on the density of red pixels in the image provided in the fifth column the CT scan depicted the existence of a tumor.

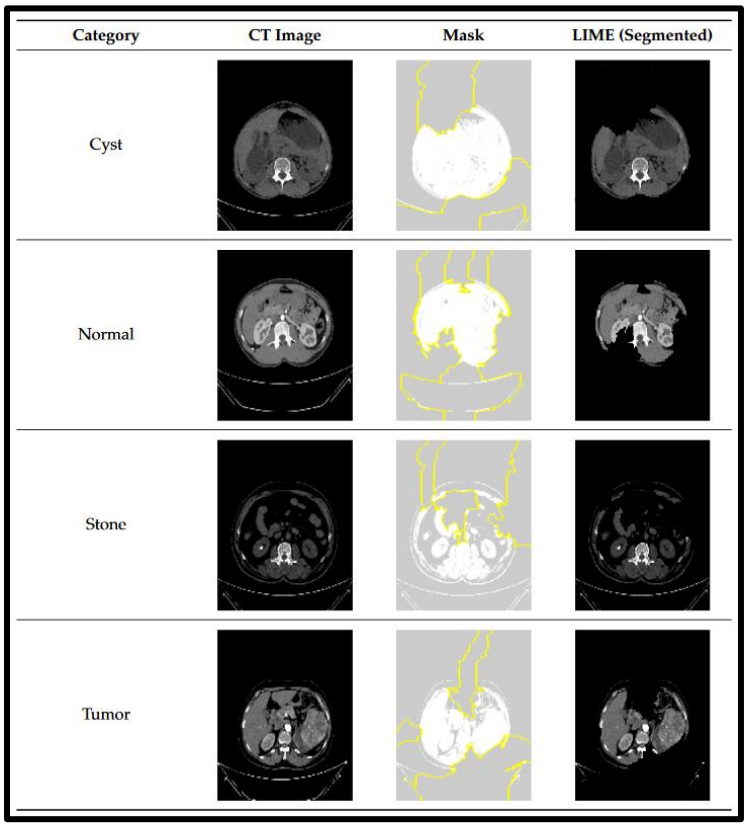


Figure 10 LIME images illustrating the CNN model's explanation for classifying kidney conditions, including Cysts, Stones, or Tumors (Bhandari et al., 2023)

As illustrated in Figure 10, LIME provided a visual representation of the model's decision-making process, highlighting the regions in the image that significantly contributed to a specific class prediction. For example, in Figure 10, the segmented LIME results displayed

only those portions of the image that played a substantial role in classifying the image as a cyst.

Finally, the proposed lightweight CNN model can accurately identify renal abnormalities such as cysts, stones, and tumors using LIME and SHAP methods by visual inspection. The proposed model provides interpretable explanations for its predictions, which can help clinicians make informed decisions. The proposed model can be used as an alternative approach to achieve higher accuracy in the diagnosis of renal abnormalities (Bhandari et al., 2023).

Interpretation techniques are crucial, particularly in the medical field, where the accurate diagnosis of diseases is of paramount importance. Medical applications should prioritize the incorporation of interpretation techniques into their models to enhance the reliability and transparency of disease diagnosis. (van der Velden et al., 2020) proposes a method to directly assess breast density on MRI using a regression CNN. The CNN model was trained on 350 patients, validated on 75 patients, and tested on 81 patients with breast cancer. The CNN received input in the form of 128 x 128 voxel slices of the breast MRI, and it produced an output as a continuous density value ranging from 0 (indicating a fatty breast) to 1 (indicating a dense breast). The study found that the density predicted by the CNN on the test set was significantly correlated with the ground truth densities ($N = 81$ patients, Spearman's rho = 0.86, $P < 0.001$) (van der Velden et al., 2020).

The paper also discusses the use of Deep SHAP to investigate why the CNN came to its predicted density. The authors found that voxels in fibroglandular tissue (FGT) often had positive SHAP-values, voxels in fatty tissue often had negative SHAP-values, and voxels in non-breast tissue often had SHAP-values close to zero. This means that the prediction of density is based on the structures we expect it to be based on, namely FGT and fatty tissue (van der Velden et al., 2020).

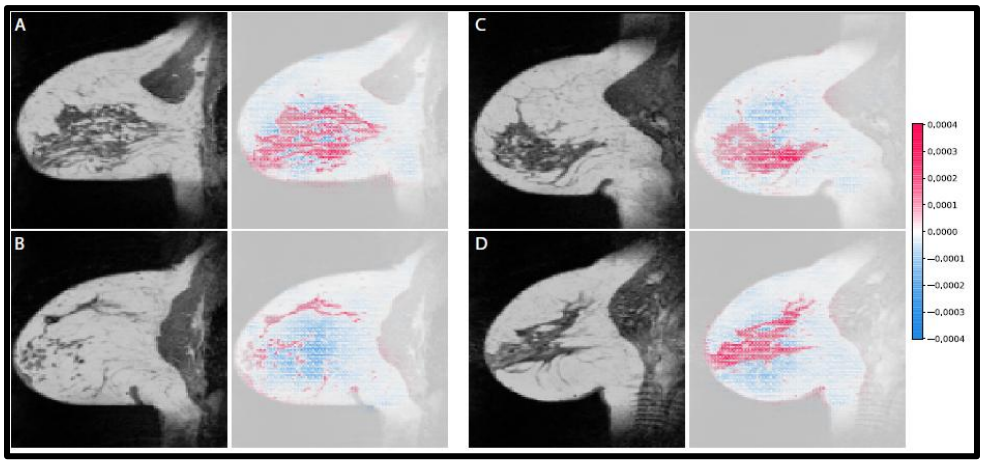


Figure 11 SHAP Analysis, Authors are using the visual inspection by human to confirm the results of the CNN (van der Velden et al., 2020)

Figure 11 presents four instances where the CNN accurately predicted the slice density. Each pair of images consists of the original slice used for density prediction on the left and a SHAP-map on the right. The underlying slice, shown with 50% opacity, reveals anatomical details. Positive SHAP-values (in red) appear in fibroglandular tissue, while negative SHAP-values occur in fatty tissue. Notably, areas like air and pectoral muscle that do not influence density are not indicative of CNN's explanation.

Our focus shifts towards the interpretation of CNN models in the context of neuroscience data. Understanding the application of CNNs to analyze and interpret data in the field of neuroscience is crucial for unraveling complex patterns and gaining insights into the intricate structures inherent in neurological information. This section aims to explore existing research, methodologies, and advancements in utilizing CNNs for the interpretation of neuroscience data, shedding light on the contributions and challenges in this interdisciplinary domain.

This literature review aims to provide an overview of the current state of research on the use of CNN models with MEG data and the interpretation of their predictions using LIME and SHAP.

A novel approach by (Giovannetti et al., 2021) that presents an innovative method for predicting early signs of Alzheimer's disease through the analysis of magnetoencephalography (MEG) data. The approach involves representing MEG data by measuring functional connectivity (FC) between bio-magnetic signals originating from distinct brain regions. This representation is then integrated with image-based MEG data representations using ensemble classifiers based on deep convolutional neural networks (CNNs) in the proposed Deep-MEG architectures (Giovannetti et al., 2021). To validate the effectiveness of the proposed method, Deep-MEG was tested on a dataset comprising resting-state MEG recordings and corresponding magnetic resonance imaging scans from an 87-subject. The findings demonstrate that Deep-MEG serves as a robust tool for identifying early changes in spectral-temporal connectivity profiles and their spatial relationships. The paper

provides a comprehensive insight into the spatio-temporal CNN features and multi-band ensemble classification employed in the approach. It also details the spatial and temporal data preprocessing steps, including band-filtering operations, various FC indicators, their transformation into FC images, and the extraction of deep spatio-temporal features based on AlexNet (Giovannetti et al., 2021).

Another paper that used an MEG Data, (Mouches et al., 2023) presents an innovative method for identifying epileptic spikes in magnetoencephalography (MEG) data. The proposed approach involves the use of a 1D temporal convolutional neural network (Time CNN) in conjunction with a graph convolutional network (GCN) to categorize short time intervals of MEG recordings as either containing or not containing an epileptic spike. In addition, a GCN is used to consider the spatial relationships between MEG sensors. The proposed method demonstrates clinically relevant results and surpasses the performance of existing state-of-the-art deep learning methods, achieving a classification f1-score of 76.7% on a balanced dataset and 25.5% on a realistically unbalanced dataset specifically for the spike class (Mouches et al., 2023).

Moreover, (Hasasneh et al., 2018) presents a deep learning method for categorizing and eliminating ocular and cardiac artifacts in neuromagnetic recordings. The authors propose an artifact classification system based on a combined deep and convolutional neural network (DCNN) model. This model automatically identifies cardiac and ocular artifacts in neuromagnetic data without the need for additional electrocardiogram (ECG) and electrooculogram (EOG) recordings. The model utilizes both spatial and temporal

information from decomposed magnetoencephalography (MEG) data to extract features. The method is fully automated, independent of user input, and once trained, it doesn't rely on auxiliary signal recordings. The flexibility of the model's design and training accommodates different modalities (MEG/EEG) and sensor types (Hasasneh et al., 2018). In general, Convolutional Neural Networks (CNNs) are neural networks particularly effective for image recognition tasks. They work by applying filters to input images, producing a feature map that is used for image classification or further processing. CNNs also excel at feature extraction, identifying relevant features from input images. Transfer learning, a technique used in CNNs, involves using a pre-trained CNN to extract features from new images, leveraging knowledge from a larger dataset, which is beneficial when training data is limited (Hasasneh et al., 2018). The paper provides a comprehensive overview of CNNs for feature extraction, highlighting key techniques and approaches in this research area (Hasasneh et al., 2018).

This literature review will now present an expanded collection of papers using MEG data, emphasizing the importance of **frequency bands** as key features, and highlighting the importance of these specific bands. The focus remains on exploring the central role of these frequency bands in the context under consideration.

(da Silva Castanheira et al., 2021) highlights the importance of frequency bands in neurophysiological activity, suggesting that the preservation of brain oscillations and cross-frequency coupling effects across species indicates a fundamental role in the temporal coordination of neuronal activity. Emphasizing evidence from neurophysiological

recordings, the paper highlights that neuronal γ-band (30–70 Hz) rhythms play a crucial role in attention and perception processes, involving binding by synchronization or communication by coherence. The paper concludes by stating that the presented technique is promising for identification of neurophysiological biomarkers for psychiatric disorders (da Silva Castanheira et al., 2021). Using magnetoencephalography (MEG) to record brain activity in 158 participants, the study shows that neurophysiological functional connectomes, similar to fMRI approaches, allow individual differentiation at rates comparable to fMRI results. The paper shows that individual differentiation is equally successful using simpler measures, such as the spatial distribution of neurophysiological spectral signal power. Furthermore, the study shows that differentiation can be achieved with short 30-second brain recordings, and that this distinct neural fingerprint remains robust over time, even in recordings conducted weeks after the baseline data collection (da Silva Castanheira et al., 2021). While the paper lacks detailed information on the modeling and feature extraction techniques used, it suggests that the presented approach could be valuable in identifying neurophysiological biomarkers for psychiatric disorders (da Silva Castanheira et al., 2021).

However, (Demuru & Fraschini, 2020) presents an innovative method for biometric identification using EEG signals. This approach involves the extraction of phase synchronization (PS) features for subject identification in various tasks, quantified by the phase locking value (PLV) in different frequency bands. Experimental results, based on the application of the proposed method to three datasets with different cognitive tasks, show consistently high classification accuracies and commendable stability (Demuru & Fraschini, 2020). Specifically, in the beta and gamma bands, the average accuracies exceed 97%, with

standard deviations equal to or less than 10^{-2} for both Dataset 1 and Dataset 2. In the case of Dataset 3, the PS features vectors across all bands that achieve high classification accuracies, exceeding 97%, and exhibit standard deviations of the same magnitude. It is noteworthy, however, that the paper does not emphasize the importance of frequency bands in the results; but instead draws attention to the importance of PS features for subject identification (Demuru & Fraschini, 2020).

Moreover, (Demuru et al., 2017) investigates whether the functional (FC) or effective (EC) connectivity patterns of one monozygotic twin can be utilized to identify the co-twin within a larger sample. The research assesses the overlap in functional fingerprints within monozygotic (MZ) twin pairs using resting-state magnetoencephalography (MEG). The study involved 32 cognitively normal MZ twin pairs from the Netherlands Twin Register participating in the EMIF-AD preclinAD study, with an average age of 68 years. By combining EC information across multiple frequency bands, the researchers achieved an identification rate exceeding 75%. These findings, particularly in genetically identical MZ twin pairs, suggest a substantial genetic contribution to MEG-based EC patterns, leading to significant similarities in brain connectivity patterns between individuals even after 60 years of life or more (Demuru et al., 2017). The paper emphasizes the importance of frequency bands in the obtained results. By consolidating EC information across various frequency bands, the researchers achieved an identification rate exceeding 75% (Demuru et al., 2017).

Finally, (Sareen et al., 2021) investigates the functional connectome (FC) of resting-state magnetoencephalography (MEG) data obtained from the Human Connectome Project. The

authors examine MEG FC fingerprinting and its correlation with various factors, including amplitude- and phase-coupling functional connectivity measures, spatial leakage correction, frequency bands, and behavioral significance. Two identification scoring methods, namely differential identifiability, and success rate, are employed to assign quantitative fingerprint scores to each FC measurement (Sareen et al., 2021). The study examines the patterns of edgewise and nodal MEG fingerprinting across different frequency bands (delta, theta, alpha, beta, and gamma). The results indicate that the performance of fingerprinting is significantly influenced by the functional connectivity measure, frequency band, identification scoring method, and spatial leakage correction. The authors observe higher MEG fingerprints in phase-coupling methods, central frequency bands (alpha and beta), and in specific networks such as visual, frontoparietal, dorsal-attention, and default-mode networks (Sareen et al., 2021).

2.4 Conclusion

The literature reviewed demonstrates the growing importance of CNNs in the analysis of MEG data, providing innovative approaches for the early prediction of neurological disorders, identification of epileptic spikes, and artifact classification. The integration of CNNs with MEG data provides new insights into the spatio-temporal features and connectivity patterns of brain activity, offering promising tools for both clinical and research applications.

The deep MEG model proposed by (Giovannetti et al., 2021) illustrates the potential of CNNs to capture spectral-temporal connectivity profiles for early detection of Alzheimer's disease.

The use of ensemble classifiers and deep spatio-temporal features highlights the robustness of the model in dealing with complex MEG data. Furthermore, the study emphasizes the importance of preprocessing steps, including band-pass filtering operations, in improving the interpretability of the model.

(Mouches et al., 2023) (Mouches et al., 2023) contribute to the literature by introducing a time CNN and graph convolution network to detect epileptic spikes in MEG data. The inclusion of spatial relationships between MEG sensors through a Graph Convolution Network demonstrates the effectiveness of a multimodal approach in improving classification results. The emphasis on achieving clinically relevant results further highlights the potential of CNNs to address real-world challenges in neurological diagnosis.

(Hasasneh et al., 2018) highlight the versatility of CNNs in artifact classification, with a particular focus on ocular and cardiac artifacts. The proposed DCNN model demonstrates the ability to automatically identify and eliminate artifacts in MEG data, thereby increasing the reliability of neuromagnetic recordings. The adaptability of the model to various modalities and sensor types adds to its practical utility.

The literature also highlights the importance of frequency bands in the interpretation of CNN models applied to MEG data. Studies by (da Silva Castanheira et al., 2021; Demuru et al., 2017; Demuru & Fraschini, 2020; Sareen et al., 2021) collectively highlight the pivotal role of frequency bands as key features in differentiating individuals, identifying neurophysiological biomarkers, and characterizing functional connectivity patterns. The

findings underscore the need for a nuanced understanding of frequency-specific information for accurate interpretation and diagnosis.

To enhance the interpretability of CNN models on MEG data, researchers have explored the use of SHAP and LIME. While specific details of these techniques are not provided in the literature reviewed, the inclusion of interpretability methods suggests a growing awareness of the importance of understanding model decisions. Visual inspection and mathematical approaches using SHAP and LIME can provide valuable insights into the features that contribute to CNN predictions, aiding clinicians and researchers to build confidence in these models.

In light of the existing literature, it is advisable for future investigations to further and improve the interpretability of CNN models applied to MEG data. Furthermore, given the crucial role of frequency bands highlighted in the literature, there is a need for further exploring the physiological significance of specific frequency bands, with the potential to uncover novel biomarkers for neurological disorders. My master thesis aims to contribute to this objective by using an Explainable Deep Learning approach tailored for neuroscience data, focusing on the analysis of features extracted from the hidden layers of the model. By elucidating the interpretability of deep learning models applied to MEG datasets, the research aims to provide valuable insights for neuroscientists, clinicians, and machine learning experts. This collaborative effort is expected to drive advances in the field and facilitate the translation of innovative approaches into practical clinical applications.

Chapter Three: Methodology of Design Deep Learning Models and their Interpretation

3.1 Introduction

The methodology chapter serves as a guiding framework for the comprehensive exploration and interpretation of the interpretability of deep learning models in the context of various datasets, with a particular focus on magnetoencephalography (MEG) data. This chapter outlines the systematic approach used to achieve the overall objectives of the thesis, namely, to clarify the use of CNNs in image classification, to verify the interpretability of CNNs on different datasets, and explore the interpretability of CNNs when applied to MEG data.

3.2 Dataset Selection and Description

The first phase of the methodology involves the careful selection and description of the datasets to be studied. This includes elucidating the characteristics of the Dogs and Cats dataset (Sachin, 2019) or image classification, the Modified National Institute of Standards and Technology (MNIST) dataset (sherena johnson, 2010) for digit recognition, and the MEG dataset (Kampel et al., 2023) for neuroscience applications. The unique features and challenges of each dataset define the stage for the subsequent steps in the research process.

The choice of the Dogs and Cats dataset is motivated by the inherent human ability to visually interpret and differentiate structural details between images of dogs and cats. This dataset's simplicity allows for straightforward visual inspection by humans, enabling researchers to easily understand and validate the interpretation results generated by the CNN model. The

transparency in interpreting the model's predictions enhances its explainability, making it more accessible to non-experts and facilitating collaboration across interdisciplinary teams.

Moreover, the use of the Dogs and Cats dataset contributes to the seamless development and improvement of CNN models. The straightforward nature of the dataset allows for quick iterations in model development, making it easier to identify and address challenges. Researchers can efficiently fine-tune the model by changing the model's parameters, until it reaches its maximum accuracy potential. This iterative process is crucial for achieving optimal performance and ensuring that the CNN model is well-adapted to the specific characteristics of the dataset.

In addition to the Dogs and Cats dataset, the inclusion of the MNIST dataset for digit recognition serves a specific purpose in our research methodology. MNIST, consisting of handwritten digits, is chosen for its inherent advantages in developing CNN models with high accuracy. The simplicity and clarity of the dataset make it an excellent choice for training models that excel in recognizing and classifying handwritten digits.

The MNIST dataset, consisting of handwritten digits, stands as a cornerstone in the realm of CNN research and development. This dataset, with its collection of 28x28 grayscale images representing digits from 0 to 9, has become the ideal benchmark for evaluating the performance of machine learning models, particularly CNNs. The simplicity and clarity of MNIST make it an ideal playground for exploring and understanding the complex workings of neural networks, as well as for advancing interpretability techniques.

In this context, the application of SHAP to interpret CNN models trained on the MNIST dataset adds a layer of transparency and comprehensibility to the neural network's decision-making processes. SHAP excels in unraveling the black box nature of deep learning models, providing insights into feature importance, and contributing factors for specific predictions (Beohar & Rasool, 2021).

What makes MNIST particularly conducive to SHAP-based interpretation is the ease with which features, in this case, pixel values, can be extracted and analyzed. Each pixel in the handwritten digits carries valuable information contributing to the overall digit classification which can easily be presented as visual images and mathematically using SHAP. SHAP enables a straightforward extraction of the importance of each pixel, shedding light on the neural network's focus areas and aiding in the development of a more nuanced understanding of its decision criteria.

In this exploration, we delve into the application of SHAP in interpreting CNNs trained on the MNIST dataset. By uncovering the significance of individual pixels and visualizing the decision-making process, we aim to demystify the inner workings of the CNN, fostering a clearer comprehension of how these models discern and classify handwritten digits.

Furthermore, the MNIST dataset's suitability for determining the features' importance is heightened through its compatibility with SHAP analysis, a tool employed to interpret the features importance in machine learning models. The straightforward nature of digit images

in MNIST simplifies the process of attributing importance to features, making it easier to interpret the results using SHAP. This advantage in determining feature importance is particularly notable when compared to the Dogs and Cats dataset.

To further enhance the interpretability of results, within this research an augmented MNIST dataset has been done by introducing additional digits with specific font types as features within each class (digit). This deliberate addition facilitates the interpretability of results not only through visual inspection by humans but also through the calculation of SHAP values. The integration of computer-generated digital fonts creates a structured and controlled environment for examining the impact of different features on model predictions.

The use of SHAP allows for a dual approach to interpretation—visual inspection by human observers and a mathematical analysis of SHAP values. This interpretation strategy aims to establish a strong correlation between visual and mathematical results. Demonstrating the consistency between visual inspection and SHAP values is a critical step in building confidence in the interpretability of the model.

Importantly, this groundwork in interpretation methodology sets the stage for subsequent applications on more complex datasets, such as the MEG dataset for neuroscience applications. By establishing the reliability and alignment of visual and mathematical interpretation results, we pave the way for extending this approach to more intricate datasets, fostering a robust and comprehensive methodology for MEG dataset interpretation in the later stages of this research.

Analyzing neural data, particularly in the context of MEG, presents challenges, particularly when extending the analysis across various frequency bands. In these cases, determining the most crucial frequency band (feature) remains uncertain. The intricacies of MEG data, especially coherence (COH) images, add complexity to this task, making it difficult for human observers to identify the key frequency bands through visual interpretation alone. To tackle this issue, our research methodology strategically initiates with the interpretation of more accessible datasets, such as the Dogs and Cats dataset for image classification and the MNIST dataset for digit recognition.

The rationale behind commencing the interpretation procedure with these simpler datasets lies in the necessity to establish a foundation for the application of interpretability tools such as SHAP and LIME. These tools prove invaluable in elucidating the decision-making processes of complex machine learning models. While the Dogs and Cats dataset facilitates visual interpretation due to its inherent simplicity, the MNIST dataset offers a platform to understand and refine the application of SHAP and LIME, given its suitability for feature extraction and incorporation of digital fonts.

By unraveling the intricacies of SHAP and LIME through these preliminary datasets, this research aims to lay the groundwork for overcoming the challenges associated with interpreting MEG COH images. The utilization of SHAP allows for a quantitative and mathematical analysis of feature importance, bridging the gap between visual inspection and algorithmic understanding. This approach is crucial for establishing a robust methodology

that ensures the reliability and consistency of interpretability tools across varying complexities of datasets.

In essence, the journey from Dogs and Cats to MNIST datasets serves as a steppingstone towards unraveling the complexities inherent in interpreting MEG data. Through this systematic approach, we seek to build a comprehensive framework that empowers researchers to make sense of the intricate neural connectivity patterns captured in MEG COH images. This iterative and structured methodology enables a seamless transition from more straightforward datasets to the nuanced challenges presented by neuroscientific data, fostering a deeper understanding of the neural processes under investigation.

3.3 CNN Model Development Workflow

The development of CNN models for the Dogs and Cats, modified MNIST, and MEG datasets involves a comprehensive and tailored approach for each dataset. Detailed dataset information, including characteristics and challenges, is carefully considered. This is followed by meticulous dataset preparation procedures to expose the model to a representative set of images. The architecture of the CNN model is then crafted, considering the specific requirements of image classification and digit recognition tasks for Dogs and Cats and modified MNIST datasets, respectively.

Similarly, for the MEG dataset, a specialized CNN model is meticulously crafted, considering the unique features (frequency bands) and challenges inherent in neuroscience datasets. The dataset preparation steps are tailored to ensure the model captures the intricate

neural connectivity patterns present in MEG data. The architecture of the CNN model is designed with a specific focus on accommodating the complexities of MEG data analysis.

Once the models are developed, a thorough evaluation is conducted for each dataset. This involves employing relevant metrics and validation techniques to assess the performance of the CNN models. The structured approach ensures effective development and assessment, providing insights into how well the models perform in their respective tasks.

This meticulous development process, tailored to the specific characteristics of each dataset, lays the groundwork for subsequent interpretability analyses. The CNN models, which are well-suited to the intricacies of Dogs and Cats, MNIST, and MEG datasets, serve as the basis for gaining insights into their decision-making processes using interpretability techniques.

3.4 Interpretability Techniques

Interpretability techniques play a crucial role in unraveling the decision processes of developed CNN models. In our study, two prominent eXplainable Artificial Intelligence (XAI) methods, LIME and SHAP, have been selected for their effectiveness in shedding light on the intricacies of these models.

LIME is a model-agnostic method chosen for its versatility across different datasets and models. It operates by generating perturbed instances of input data, obtaining predictions from the black-box model for these instances, and fitting a local interpretable model, such as a linear model, to approximate the complex model's behavior in the vicinity of a specific data

point. In our case, LIME was adapted for application to both the Dogs and Cats dataset and the MNIST dataset, providing locally faithful explanations for the predictions made by our CNN models (Ribeiro et al., 2016).

On the other hand, SHAP values, rooted in cooperative game theory, offer a principled approach to attributing the output of any machine learning model to its individual features. SHAP is also model-agnostic and ensures a fair distribution of contributions among features. SHAP values were applied to both the Dogs and Cats dataset and the MNIST dataset, quantifying the impact of each feature (pixel) on the model's predictions. Unlike LIME, SHAP provides a global perspective on feature importance, considering all possible combinations of features and ensuring consistency in attributions (S. Lundberg & Lee, 2017b).

The selection of LIME and SHAP is motivated by their model-agnostic nature, making them suitable for various machine learning models, including CNNs. LIME's local interpretability aligns well with the need to understand individual predictions in diverse datasets, while SHAP's global interpretability complements this by providing an overarching view of feature importance.

3.4.1 Human Visual Inspection

In conjunction with quantitative interpretability techniques, human visual inspection plays a pivotal role in understanding and validating the interpretability results. This method involves

a qualitative analysis of the interpretability results, which promotes a holistic understanding of the decision-making rationale of the developed CNN models.

Human visual inspection entails examining the output of the model, such as visualizations, heatmaps, or saliency maps, to gain insights into which features or regions are influencing the model's predictions. This qualitative analysis allows researchers to intuitively grasp the patterns and relationships learned by the model, providing a complementary perspective to the quantitative results obtained through other interpretability techniques.

For instance, in image classification tasks like Dogs and Cats, visual inspection might involve scrutinizing the areas of an image that the model focuses on when classifying it as a dog or a cat. In digit recognition tasks like MNIST, researchers could examine which pixels are deemed most important by the model for making accurate predictions.

In scenarios where the analysis extends across multiple frequency bands, rendering us unaware of the most important frequency band (feature), human visual inspection, though a potent tool for comprehending model behavior, is subjected and may not be generally suitable. This challenge is particularly evident in intricate datasets like MEG COH images, where discerning neural connectivity patterns visually is demanding. The inherent limitations of human visual perception become apparent given the intricate nature and high dimensionality of the data, impeding the extraction of meaningful patterns without the aid of quantitative methods, which excel in handling intricate data structures. Additionally, techniques such as dimensionality reduction, clustering algorithms (e.g., K-means,

hierarchical clustering), feature selection methods (e.g., mutual information, recursive feature elimination), machine learning algorithms (e.g., random forests), and graph-based methods leveraging graph theory can be employed to uncover insights into neural connectivity patterns within the data.

Within this research, the MEG dataset, particularly the COH images, poses challenges for visual interpretation due to the abstract nature of neural connectivity patterns. The intricate visual details inherent in COH images make it impractical for human observers to extract meaningful insights through direct visual inspection alone. Addressing these challenges and achieving a more systematic and dependable understanding of the model's decision processes in the realm of MEG data necessitates the incorporation of quantitative interpretability techniques like LIME and SHAP. Specifically, explaining the feature importance, particularly in the case of frequency bands, is not straightforward through visual inspection alone.

In other words, while human visual inspection is a valuable tool in certain contexts, its limitations in handling the complexity of finding the frequency bands importance in MEG COH images underscore the necessity of complementary quantitative interpretability techniques for a more comprehensive understanding of the CNN model's decision-making processes in our specific case.

3.5 Integration of Interpretability in Discussion

The proposed methodology for this research endeavors to unravel the decision processes of CNN models applied to diverse datasets, emphasizing interpretability through a multi-faceted approach. The initial phase involves meticulous selection and description of datasets, namely Dogs and Cats for image classification, MNIST for digit recognition, and MEG for neuroscience applications. Each dataset, chosen for its unique characteristics, sets the stage for subsequent interpretability analyses.

Quantitative interpretability techniques, such as LIME and SHAP, play a central role in shedding light on the model's decision-making. LIME's local interpretability and SHAP's global perspective provide nuanced insights into individual predictions and overall feature importance. The incorporation of these techniques is informed by their model-agnostic nature and robust theoretical foundations.

Human visual inspection, though a powerful tool in certain contexts, is recognized for its limitations, particularly in handling the abstract and complex nature of MEG COH images. In response, LIME and SHAP act as crucial complements, overcoming challenges posed by presenting the frequency bands importance within the MEG dataset model.

The anticipated benefits of this methodology include a comprehensive understanding of CNN model decisions across varied datasets. By iteratively applying interpretability techniques to progressively complex datasets, we establish a foundation for transparent and reliable model

insights. The adaptability and versatility of LIME and SHAP contribute to a holistic interpretation, aligning human intuition with quantitative rigor.

Expected outcomes and results, to be presented in Chapter 4, involve clear and interpretable insights into the decision processes of CNN models. We anticipate visualizations, explanations, and feature importance analyses that enhance our understanding of model predictions. The methodology's structured progression from simpler datasets to the intricacies of MEG data ensures a robust foundation for unveiling the black-box nature of CNN models, fostering transparency and trust in their applications.

Chapter Four: Results and Discussion

4.1 Introduction

This chapter embarks on a journey into the heart of deep learning model interpretability. It delves into the complex process of developing and interpreting Convolutional Neural Network (CNN) models on three different datasets: the Dogs and Cats dataset, the MNIST digit recognition dataset, and the specialized MEG dataset for neuroscience applications.

By studying each stage of this process in detail, this research aims to unveil the opaque layers of these powerful models and shed light on their decision-making mechanisms. This in-depth exploration is crucial for fostering trust and understanding of deep learning models, allowing us to harness their strengths while mitigating potential biases and limitations.

This research begins with model development, where we delve into the specific details of each dataset, from data preprocessing to the construction of detailed CNN architectures. We examine the unique characteristics of each dataset, such as image size and distribution in the case of Dogs and Cats or the specific features of handwritten digits in MNIST. This tailored approach ensures that the constructed CNN models are optimized for each specific task.

After developing each model, the research delves into Interpretability Techniques. Two powerful interpretability methods are explored: Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). These methods provide

valuable insights into how CNN models arrive at their predictions, allowing us to understand the underlying logic and rationale behind their decisions.

Finally, the integration of interpretability in discussion seamlessly blends the quantitative and qualitative findings into a cohesive narrative. This culminates in a comprehensive analysis of the interpretability results, highlighting their implications for the analysis and application of deep learning models in domains ranging from image classification to neuroscience research.

This chapter lays the foundation for the subsequent exploration of interpretability results, presented in the next chapter. By delving into the exact process of model development and interpretability, we pave the path for a deeper understanding of how CNN models work, fostering trust and paving the way for responsible and impactful applications of deep learning technology.

4.2 Dogs and Cats Dataset

Differentiating between dogs' and cats' images is a popular use case for image classification models. This is because the results are easily interpretable by humans, allowing for straightforward evaluation of interpretation techniques like LIME and SHAP. In addition, the high accuracy achievable in this task makes it a valuable benchmark for developing and testing new algorithms.

4.2.1 About the Dataset

The Kaggle dataset (Sachin, 2019) "Dogs vs. Cats" is a widely used collection of images featuring two of the most popular pets worldwide. With a total of 25,000 images, evenly split between dogs and cats, this dataset serves as an excellent resource for training and testing machine learning models, particularly CNNs for image classification.

The dataset contains a variety of images showing various breeds, postures, and environments, reflecting the natural diversity of these pets. Each image is labeled as either a dog or a cat, providing a labeled dataset suitable for supervised learning tasks.

When working with this dataset, it's important to preprocess the images appropriately. Common preprocessing steps include resizing images to a standard size, normalizing pixel values, and augmenting the dataset using techniques like rotation and flipping. These steps help improve the model's generalization and performance on unseen data.

Data augmentation enhances generalization in machine learning by diversifying and enlarging the training dataset through transformations such as rotation, flipping, and adding noise. This process introduces variability through learning and fosters robustness against variations in real-world data. It mitigates overfitting by preventing the model from memorizing specific examples, implicitly regularizes the model by introducing noise, balances class distributions, and optimizes the use of labeled data, resulting in improved performance on unseen data.

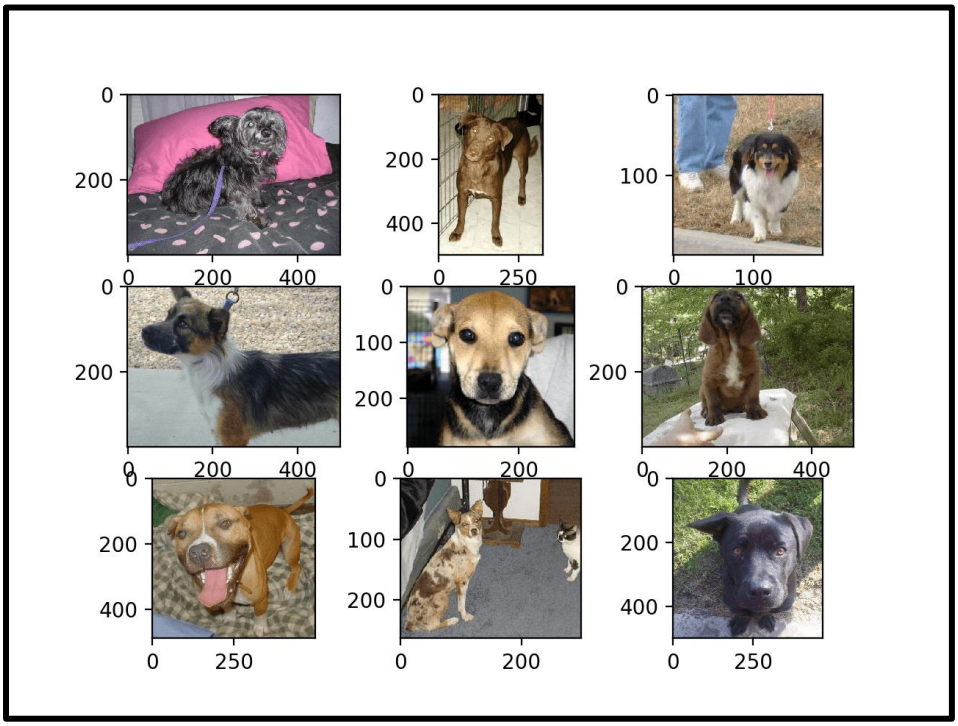


Figure 12 Samples of the Dogs and Cats Dataset with Varying Sizes, Shapes, and Different Breeds of Dogs and Cats.

4.2.2 Dataset Preparation

The first phase of building a CNN model involves data preprocessing. During this stage, we standardize the dimensions of all images to $200\ x\ 200$ pixels for first CNN model then use $224\ x\ 244$ pixels for VGG16 model, maintaining the color scheme in Red, Green, and Blue (RGB) (James et al., 2021). It is noteworthy that this research will present various models using both RGB and grayscale images, with the conversion being implemented directly in the code. Figure 13 below illustrates samples of dogs and cats after resizing all images to ensure uniform dimensions.

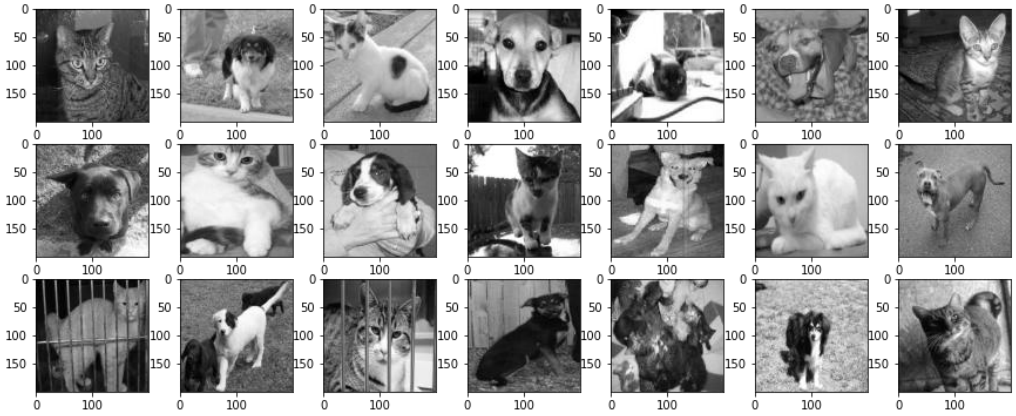


Figure 13 Samples of dogs and cats' images after standardizing the size and converting the color scheme to grayscale.

Following that, the data is divided into three distinct groups: the training set, the validation set, and the testing set. This partitioning is determined by the following percentages: 70% (17,500 images) for the training set, 20% (5,000 images) for the validation set, and 10% (2,500 images) for the testing set. This ensures that the model encounters previously unseen data during the testing phase. The distribution of images between dogs and cats is equal.

4.2.3 CNN Models Structures and Development for Dogs and Cat's Dataset

In this section, the exploration focuses on the CNN model, with the primary objective of achieving optimal accuracy. This research aim is to gain a comprehensive understanding of the model's complexities, particularly in the context of various XAI models. The focus is on obtaining the most effective interpretation results, which will play a key role in guiding subsequent interpretations of the magnetoencephalography (MEG) model.

Throughout this research, development, and implementation of three different CNN models. Initially, a CNN model was constructed. Subsequently, the model was fine-tuned this model

by adjusting key parameters such as the number of epochs, batch size, learning rate, and activation function. The first CNN architecture is processing grayscale images with $200x200$-pixel dimensions is characterized by a well-defined structure.

The network consists of multiple layers, including convolutional layers, max-pooling layers, dropout layers, an activation layer, and fully-connected layers. Specifically, it incorporates three sets of convolutional layers, with 32 filters in the first set, 64 filters in the second set, and 128 filters in the third set. Additionally, two convolutional layers with 512 filters each are included for added complexity.

The input images are grayscale and have dimensions of 200x200 pixels. The initial convolutional layer is designed with a kernel size of (3,3), and subsequent max-pooling layers use a pool size of (2,2) for down-sampling. Dropout layers with a dropout rate of 0.2 are strategically placed to mitigate overfitting.

The CNN architecture is finalized with a fully-connected layer and two dense layers. The first dense layer comprises 1024 units with the 'relu' activation function, and the final dense layer accommodates the number of classes specified by the length of the class subset with the 'sigmoid' activation function.

This CNN is tailored to effectively capture hierarchical features within grayscale images of dogs and cats, facilitating robust pattern recognition and classification. Figure 14 displays the accuracy and loss curves, revealing noticeable oscillations that require elimination. Consequently, the CNN model needs to be fine-tuned to address and rectify this issue. It is important to note that the model has been stopped at the 25th epoch due to these oscillations in accuracy, and further adjustments will be made to enhance the stability and performance of the model.

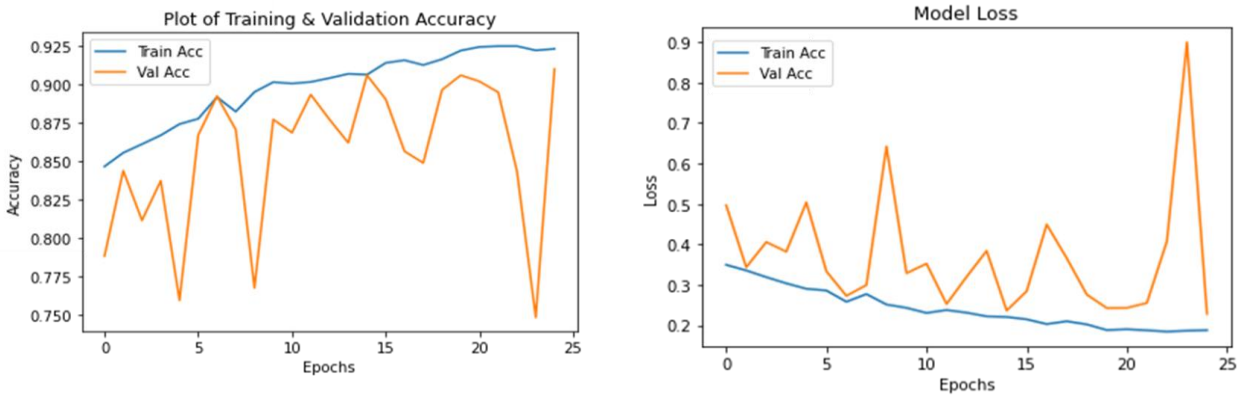


Figure 14 The accuracy and loss curves of first CNN model.

In response to the identified oscillations in the initial model's accuracy and loss curves (cf. Figure 14), the Convolutional Neural Network (CNN) underwent a tuning process. This tuning involved adjusting critical hyperparameters, including learning_rate, beta_1, and beta_2, within the Adam optimizer. The specific values used for tuning were learning_rate=0.001, beta_1=0.9, and beta_2=0.999.

Subsequent to this tuning process, Figure 15 is presented to visualize the updated accuracy and loss curves. Remarkably, these new curves exhibit a notable absence of oscillations,

indicating the success of the tuning efforts. The refined CNN model, with its adjusted hyperparameters, demonstrates improved stability during training, resulting in smoother convergence and enhanced overall performance.

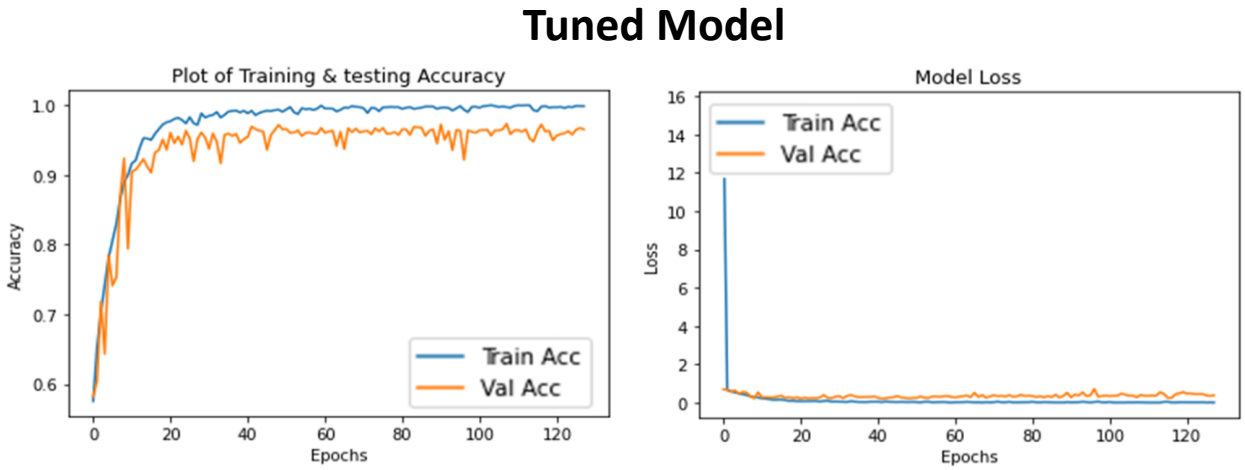


Figure 15 The results of these adjustments were significant improvements in the accuracy curve, with the elimination of oscillations and a consistent decrease in the average accuracy.

Table 1 displays the average training accuracy and validation accuracy metrics for both models throughout all epochs. The first model spans 25 epochs, while the tuned model extends over 125 epochs. It presents a comparative overview of the performance of these two models in terms of accuracy during both the training and validation phases. Notably, the validation accuracy is improved in the tuned model, indicating enhanced generalization capabilities.

Table 1 The Average Training and Validation Accuracy for two CNN models

Model	Training Accuracy	Validation Accuracy
Model #1	91.0%	82.0%
Tuned Model	90.0%	85.5%

To further refine our approach, we integrated a VGG16 (Mahardi et al., 2020) CNN model into our analysis. The VGG16 architecture is renowned for its depth and simplicity, comprising 16 layers with small (3x3) convolutional filters. It has demonstrated exceptional performance in image classification tasks. In our case, specific parameters of the VGG16 model were meticulously selected to positively influence the accuracy curve and overall accuracy. The inclusion of the VGG16 model significantly contributed to the refinement of our results, emphasizing the crucial role of parameter optimization in achieving superior model performance.

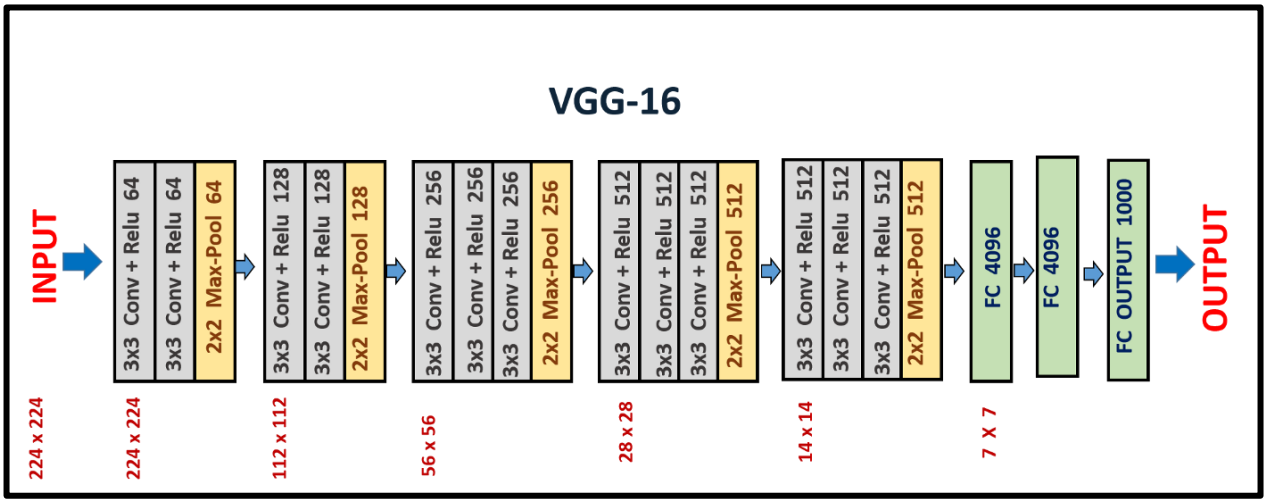


Figure 16 VGG16 model structure map (Khandelwal, 2020).

The decision to incorporate the VGG16 CNN model into our analysis stems from its well-established reputation for excellence in image classification tasks. VGG16 is recognized for its depth and simplicity, featuring a total of 16 layers with small (3x3) convolutional filters.

One of the key strengths of the VGG16 architecture lies in its ability to capture intricate features within images, thanks to its deep structure. The use of small convolutional filters allows the model to effectively learn and represent complex patterns, making it particularly suited for tasks that involve detailed visual recognition (Simonyan & Zisserman, 2014c).

In our specific implementation, we meticulously selected parameters of the VGG16 model to align with the requirements of our analysis. This careful parameter tuning was aimed at positively influencing the accuracy curve and enhancing overall accuracy. By leveraging the capabilities of the VGG16 architecture, we aimed to achieve superior performance in our image-related tasks.

The exceptional performance of VGG16 in image classification tasks has played a pivotal role in refining our results. The model's ability to learn hierarchical features, combined with our thoughtful parameter optimization, significantly contributed to the overall success of our analysis.

Furthermore, the decision to integrate VGG16 underscores the importance of parameter optimization in the context of deep learning models. The success of a neural network often hinges on fine-tuning parameters to match the specific characteristics of the dataset and task

at hand. In our case, the inclusion of VGG16 illustrates the impact of strategic parameter selection on achieving superior model performance, emphasizing the need for careful consideration in the design and implementation of convolutional neural networks.

VGG16 CNN model's input images have a shape of (224, 224, 3), signifying images with a height and width of 224 pixels and three-color channels (RGB). To tailor the model for the binary classification objective, modifications are introduced. The model output is flattened, followed by the addition of a dense layer with 128 units, employing the rectified linear unit (ReLU) activation function and He normal weight initialization. The model concludes with an output layer featuring a single unit and sigmoid activation, suitable for binary classification tasks. Stochastic Gradient Descent (SGD) is chosen as the optimization algorithm, with a learning rate of 0.001 and momentum of 0.9. The binary cross-entropy loss function is employed, well-suited for binary classification problems, and model training progress is evaluated based on accuracy. These parameter choices collectively shape the VGG16-based model to meet the specific requirements of binary classification, incorporating tailored adjustments to the architecture.

Following the training process, Figure 17 visually represents the training and loss accuracy for both the training and validation sets. The graphic depiction clearly indicates that this model improves the accuracy of distinguishing between images of dogs and cats. Furthermore, the accompanying table provides numerical details regarding the training and validation accuracy, offering a quantitative insight into the model's performance. It is worth

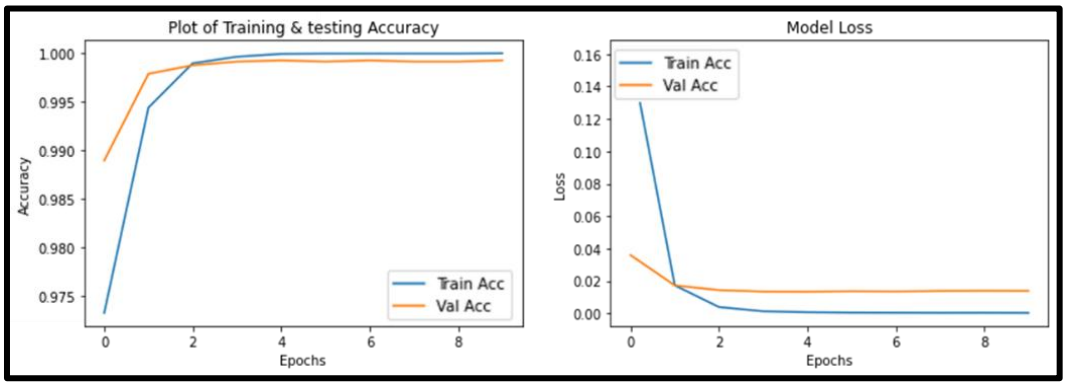


Figure 17 The accuracy and loss curves of VGG16 models using the Dogs and Cats Dataset.

noting that the number of epochs within this model is less than the previous two models, as the VGG16 model converges faster. This observation underscores the efficiency and quicker convergence achieved by the VGG16 architecture.

Table 2 The Average Training and Validation Accuracy for VGG16 Model

Training Accuracy	**Validation Accuracy**
99.8%	99.7%

This refined methodology, coupled with the incorporation of the VGG16 model, sets the stage for a more robust interpretation of the MEG model in subsequent analyses by finding the importance of each frequency band, reinforcing the significance of leveraging established architectures for enhanced model understanding and performance.

4.2.4 LIME Interpretation for Dogs and Cats Deep Learning Model

Having successfully created a robust CNN model utilizing VGG16, that achieves an impressive accuracy of over 99%, this section explores the results of applying LIME to previously unseen data. The ensuing discussion reveals the effectiveness of LIME in clarifying the model's predictions, highlighting instances where its interpretive power matches its accurate predictions.

Here are the LIME results, achieved through the strategic optimization of key parameters within the LIME function. The process involved fine-tuning each parameter using Python code:

- **num_features**: This parameter dictates the number of features (or variables) selected by LIME for interpretation. Adjusting this value allows for a balance between precision and simplicity in the interpretative model. Typically set to 10, providing a reasonable balance between precision and simplicity. Adjust based on the complexity of the model and interpretability requirements.
- **num_samples**: The num_samples parameter determines the number of perturbed samples generated by LIME to approximate the model's behavior. A higher value increases the robustness of the interpretation but may require more computational resources. Usually set to 500 or a similar moderate value. A higher value improves the robustness of interpretation at the cost of increased computational resources.
- **batch_size**: This parameter controls the size of the batches used during the LIME sampling process. It influences the efficiency of the computation and can be adjusted

based on the available computational capacity. The default batch size often ranges from 10 to 100, depending on the available computational resources. It can be adjusted to optimize the efficiency of the computation.

- **model_regressor**: Specifies the regression model used by LIME to approximate the complex model's behavior locally. Choosing an appropriate regressor enhances the fidelity of the interpretative results. The default choice is often a linear regression model.
- **segmentation_fn**: This parameter defines the function used for superpixel segmentation in image data. It plays a crucial role in determining the local regions considered by LIME for interpretation, contributing to the contextual understanding of predictions in image-based models. The default segmentation function varies based on the library or tool used. For image data, quickshift or felzenszwalb are common choices. The default parameters for these functions are often used, but they can be adjusted based on specific requirements. (Dieber & Kirrane, 2020; Garreau & Mardaoui, 2021; Ribeiro et al., 2016; Visani et al., 2020).

Furthermore, the values of these parameters in the table below are the ones employed in this research. The selection of these values was based on a trial-and-error technique, involving iterative adjustments and careful observation of the resulting LIME interpretations to ensure meaningful insights into the model's decision-making process.

Table 3 LIME Parameters and their Values

Parameter	**Value**
num_features	10
num_samples	500
batch_size	50
model_regressor	Linear
segmentation_fn	quickshift

When interpreting LIME results for images in the context of a dogs and cats' dataset, human visual inspection is essential. The color-coded superpixel regions generated by LIME and calculated using equation 4 provide valuable insights into the model's decision-making process. Red and green highlights are indicative of the influence of specific regions on the predicted class.

By examining the color-coded superpixel regions, one can recognize the impact of each region on the model's prediction for a specific image. Red regions indicate areas where an increase in pixel intensity contributes positively to the likelihood of the predicted class, while green regions signify the opposite. Essentially, red highlights highlight influential features that support the predicted class, and green highlights indicate features that may contradict the predicted class.

Through human visual inspection of these color-coded regions, interpreters can gain an intuitive understanding of which parts of the image are crucial for the model's decision, enhancing the transparency and interpretability of complex image classification models.

Figure 7 illustrates a random sample of LIME results, showcasing red superpixels to indicate the significance of cat-related values and green superpixels for dog-related values. A lighter shade of red or green signifies higher importance in relation to cats or dogs, respectively.

The LIME results provide insights into the model's decision-making process by highlighting superpixel regions and their corresponding importance, represented by the darkness of the red or green color. It is noteworthy that these superpixel regions outline the shape of dogs or cats, emphasizing specific body details like ears, eyes, and mouth. All these observations can be readily noticed through visual inspection, suggesting that the model relies on these features to make predictions. This visual analysis helps demystify the internal workings of the black box within the model, shedding light on its predictive mechanisms.

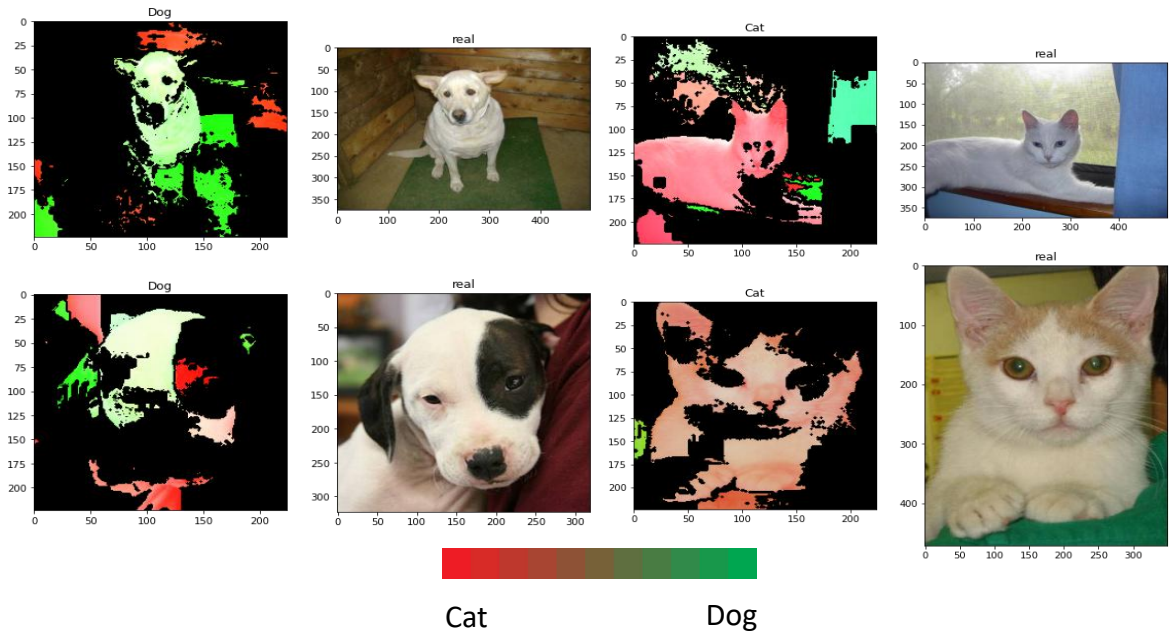


Figure 18 LIME Results for VGG16 Model

SHAP Interpretation for Dogs and Cats Deep Learning Model

The SHAP library is a powerful tool in the realm of machine learning interpretation, providing valuable insights into the decision-making processes of complex models. At the heart of SHAP lies the Explainer interface, a fundamental component facilitating the understanding of model predictions through Shapley values. Shapley values offer a principled way to allocate the contribution of each feature to the model's output, providing a nuanced understanding of feature importance.

In this context, the shap.Explainer serves as the primary gateway to unlocking the interpretability potential of SHAP. This interface accepts a machine learning model or a Python function as input, along with optional parameters, and returns a specialized object equipped with the chosen estimation algorithm. The utilization of Shapley values enables the explanation of not only traditional machine learning models but also arbitrary Python

functions, contributing to the versatility of the SHAP library (Abdollahi & Pradhan, 2023; Dardouillet et al., 2022).

After attempting to identify the optimal SHAP parameter values for the dogs and cats dataset, it was found that the default parameters are the only ones effective. This explainer utilizes Shapley values to clarify any machine learning model or Python function. A comprehensive understanding of the parameters is essential for specifying the input data type and generating meaningful SHAP plots. The crucial parameters employed in the SHAP function in Python code are outlined below:

- **model:** User-supplied function or model object representing the machine learning model or algorithm under examination.
- **link:** A function used to map between the output units of the model and SHAP value units. The default is 'shap.links.identity,' but it can be explicitly specified as 'shap.links.logit' for probability units.
- **algorithm:** Determines the algorithm used to compute Shapley values, with options such as "auto," "permutation," "partition," "tree," or "linear." The 'auto' option attempts to make the best choice based on the passed model, but a specific algorithm can be provided.
- **output_names:** Optional names of the model outputs, useful for classifiers with multiple classes. If set to None, the Explanation object produced will not have any output_names.

- **seed:** An optional parameter for reproducibility, ensuring consistent random number generation within SHAP. Useful for comparing or reproducing results consistently and for debugging purposes (Abdollahi & Pradhan, 2023; Dardouillet et al., 2022; Visani et al., 2020).

When applying SHAP to interpret the VGG16 model trained on a dataset of dogs and cats, the resulting visual representations provide compelling insights into the interpretability of this advanced technique. The color scheme employed assigns red to positive SHAP values associated with dogs and blue to negative SHAP values associated with cats. In this context, the red hues signify an inclination towards a classification related to a dog, while the blue hues correspond to a cat-related classification.

Crucially, the magnitude of the SHAP value, as elucidated in Equation 5, serves as a metric of the significance of each pixel in influencing the model's decision between dogs and cats. Darker shades of red or blue within the visualizations indicate larger SHAP values, emphasizing the heightened impact of these specific pixels on the model's predictions.

Figure 19 presents a random sample from the dogs and cats' dataset to illustrate the interpretability gained through SHAP values. The image showcases the overlay of SHAP values, with red and blue regions highlighting the areas influencing the model's decision for dogs and cats, respectively. The varying intensity of color provides a visual guide to the importance of each pixel.

In addition to the numerical values derived from SHAP, human visual inspection plays a pivotal role in comprehending the intricacies of model predictions for this specific dataset. By scrutinizing the images overlaid with SHAP values, an analyst or domain expert can gain a deeper understanding of the model's focus on certain features or regions relevant to dogs or cats.

This interpretation not only serves as a powerful analytical tool for the dogs and cats' dataset but also contributes to making the model's black box more transparent and understandable in the context of distinguishing between these two classes. The combination of quantitative SHAP values and qualitative visual inspection allows stakeholders to gain insights into the model's inner workings, fostering trust and comprehension in the context of pet classification.

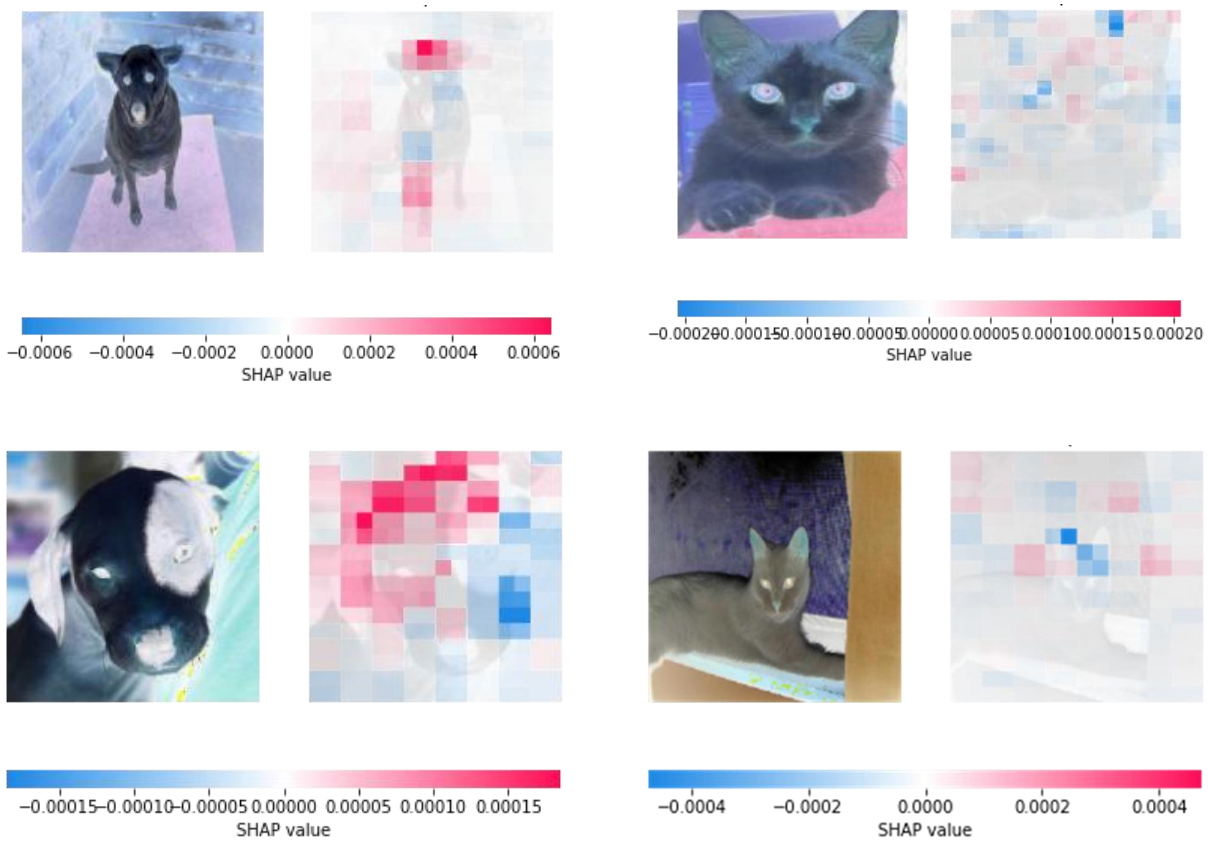


Figure 19 SHAP Results for VGG16 Model

4.3 MNIST Dataset

4.3.1 About MNIST Dataset

The MNIST (Modified National Institute of Standards and Technology) dataset is a cornerstone in the field of computer vision and machine learning, serving as a benchmark for evaluating and advancing various algorithms. Comprising a collection of 28x28 pixel grayscale images of handwritten digits (0 through 9), MNIST provides a diverse and well-labeled dataset for training and testing image classification models (Beohar & Rasool, 2021). Originally designed for the evaluation of traditional machine learning techniques, MNIST has endured as a fundamental resource, evolving to play a pivotal role in the realm of CNNs.

CNNs, with their ability to automatically learn hierarchical features from images, have demonstrated remarkable success in accurately classifying handwritten digits from the MNIST dataset (Cohen et al., 2017).

4.3.2 Dataset preparation

In this research, the MNIST dataset was modified by adding three additional digits as features, using distinct digital fonts for each class (digit). Each of them was labeled for easy access to each font. The fonts selected were: Californian FB as font 1, Verdana as font 2, and Times New Roman as font 3. This modification is particularly relevant since it may not be straightforward to address features using a dataset such as Dogs and Cats, where visual inspection might not capture font-related touches.

Figure 20 showcases sample images of these labeled fonts, providing a visual representation of the diverse font styles introduced to the MNIST dataset. It's essential to note that these images with the new fonts were created manually using drawing software, ensuring precision and consistency in their representation.

This modification serves as a crucial step in facilitating the research's objective, enabling the calculation of the significance of each font within the CNN model. Here, the different fonts are treated as features, and the research leverages the SHAP interpretation method to quantify the importance of each feature. The MNIST dataset, with its clear digit representations in

different fonts, provides a suitable environment for understanding the importance of font variations in the model's decision-making process.

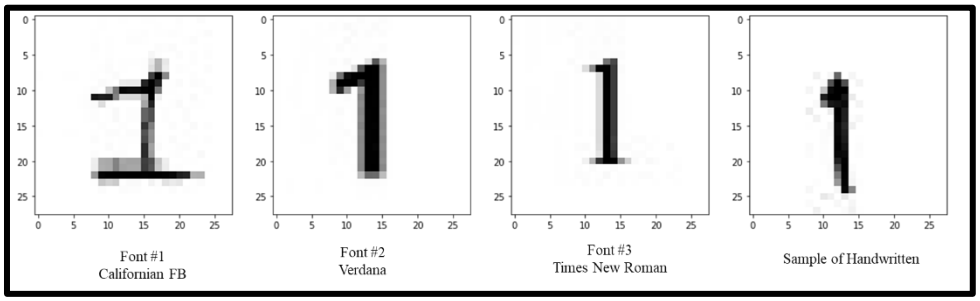


Figure 20 Sample of different fonts (Features)

In enhancing the robustness of the CNN model and avoiding image uniformity, an augmentation process was applied in addition to image generation for various fonts using python. To ensure diverse and realistic feature importance assessments within the CNN, images underwent augmentation techniques, including random zoom and random rotation. This approach not only contributes to a more reliable model but also introduces variability in the dataset, enabling CNN to discern genuine features and patterns across different orientations and scales. Figure 21 illustrates a random sample of modified MNIST augmented images, showcasing the impact of the augmentation process on the dataset.

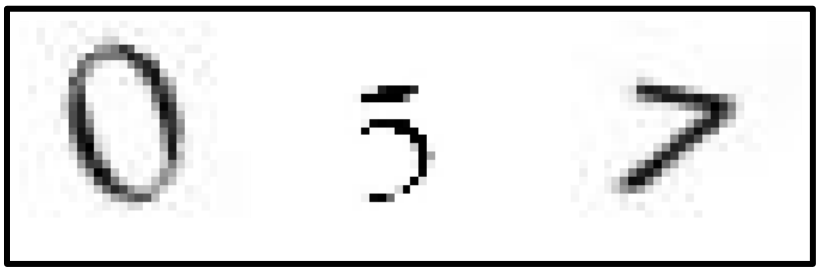


Figure 21 Sample of Augmented Images

4.3.3 CNN models structures and development

Similar to the approach taken with the dogs and cats dataset, a refined CNN model has been meticulously developed to align with the modified MNIST dataset. This tailored model has yielded exceptional accuracy, reaching **99.9%** for both training and validation sets. The architecture of the CNN model starts with an input layer designed for grayscale images with dimensions 28x28 pixels. It proceeds with a series of convolutional blocks, each consisting of two Conv2D layers with 64 filters, a 3x3 kernel size, and 'same' padding. BatchNormalization is applied after each convolutional layer, followed by LeakyReLU activation with a small alpha (0.1) to introduce non-linearity.

After each convolutional block, a MaxPooling2D layer with a pool size of (2, 2) is added to downsample spatial dimensions. To prevent overfitting, Dropout with a rate of 0.2 is employed, randomly dropping 20% of connections.

The model repeats similar convolutional blocks with increased filter size (128 and 256) and downsampling. BatchNormalization and LeakyReLU activation are consistently applied to enhance training stability and non-linearity.

After the convolutional layers, the model flattens the output and connects to a Dense layer with 256 units and LeakyReLU activation. BatchNormalization is once again applied before the final Dense layer with 10 units and a softmax activation, representing the 10 classes in a classification task.

In other words, the CNN structure involves the gradual extraction of hierarchical features through convolutional blocks, interspersed with max-pooling and dropout for downsampling and regularization. The model concludes with fully connected layers for classification, incorporating BatchNormalization and LeakyReLU for stability and non-linearity. Figure 22 illustrates the CNN model map for the modified MNIST dataset.

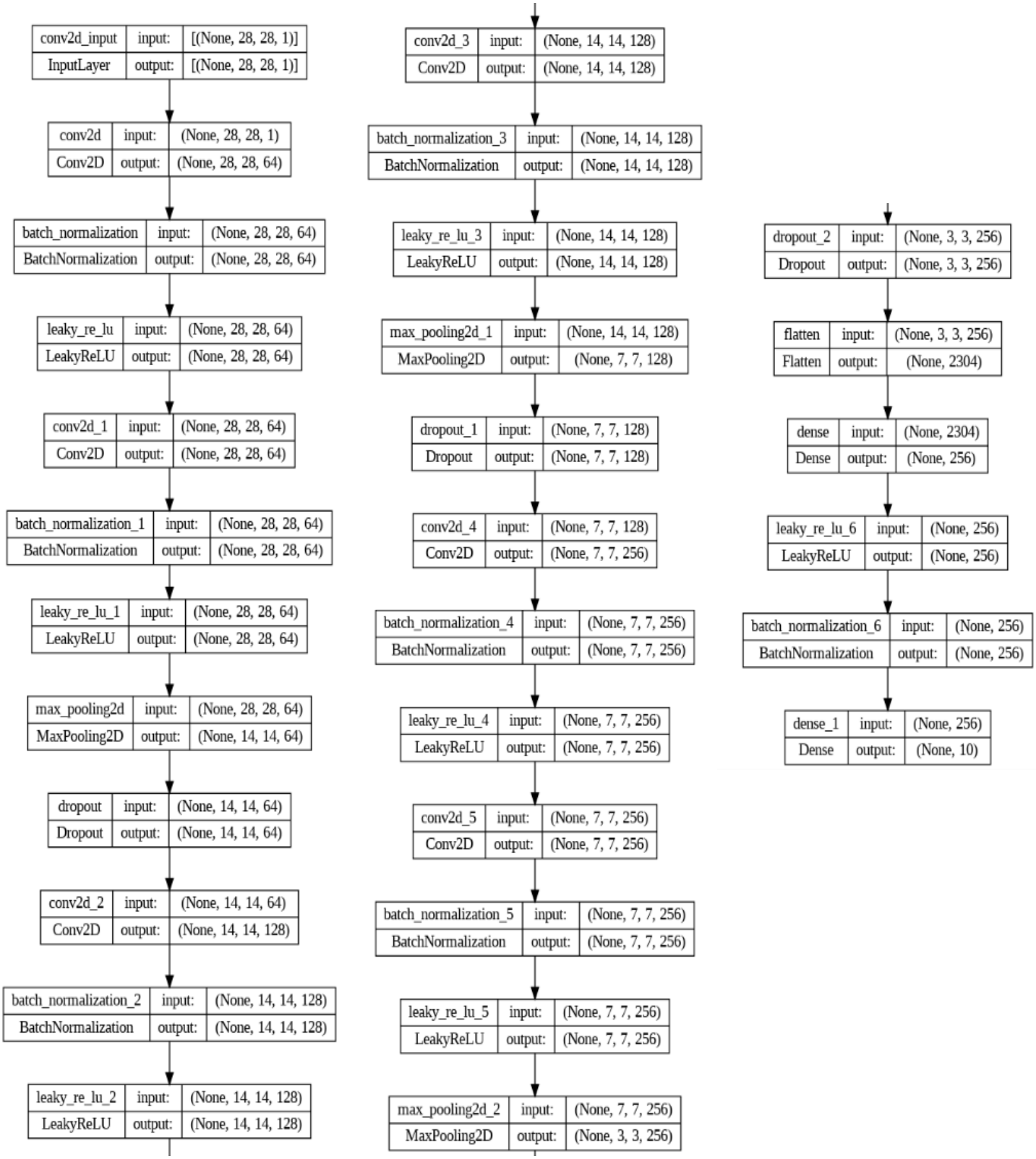


Figure 22 CNN model map that used for modified MNIST dataset.

The Modified MNIST dataset comprises 40,000 images evenly distributed among ten classes (numbers from 0 to 9). The data is divided into training and validation sets, with a split of 70% for training and 30% for testing. This division results in approximately 2,800 images for training and 1,200 images for validation for each class.

Below are the accuracy curves for this CNN, visually depicted in Figure 23.

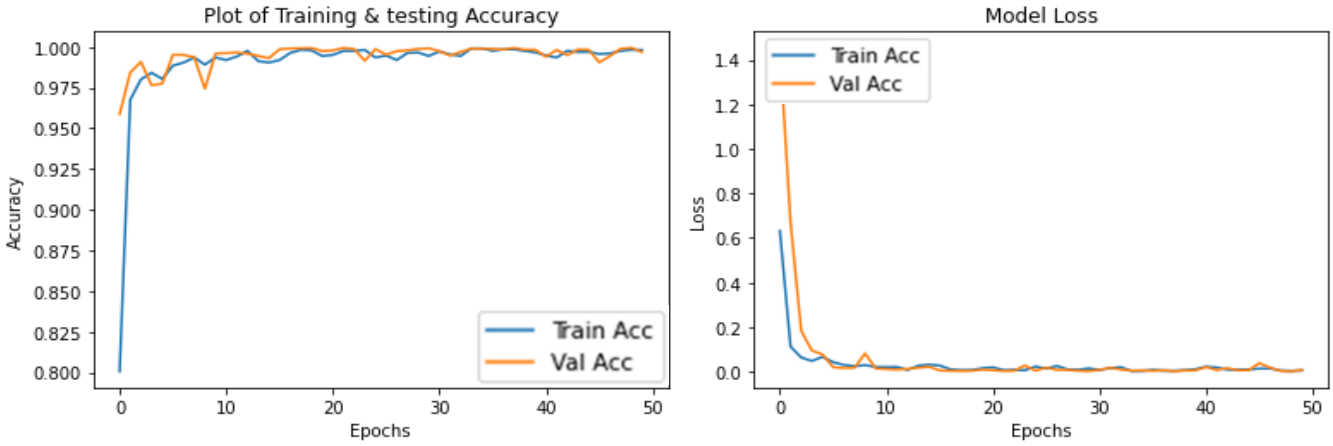


Figure 23 The accuracy and loss curves of models using the modified MNIST Dataset

Once again, the achieved accuracy was highly satisfactory, enabling the application of interpretation techniques to yield precise and reliable results.

4.3.4 SHAP Interpretation for MNIST Deep Learning Model

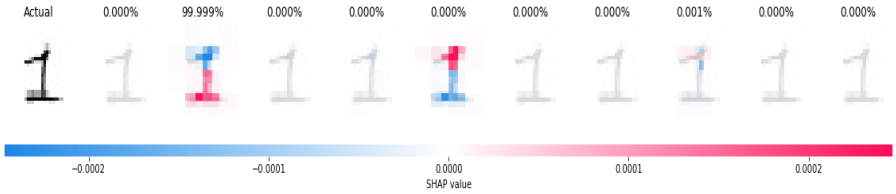
In this section, the research emphasizes interpretation using SHAP values, incorporating both visual inspection and mathematical calculations. It is essential to note that LIME is not utilized in this section, as the approach seeks to assess the importance of each pixel through SHAP values. Unlike LIME, which relies on perturbed regions, such an approach is not applicable in the current dataset context.

The primary objective is to validate the SHAP results, ensuring the CNN model's effective training. This validation process aims to establish the model's proficiency in making accurate predictions and serving as a reliable decision-maker. Additionally, the research strives to extend the applicability of this interpretative technique to MEG data, demonstrating its utility across diverse domains.

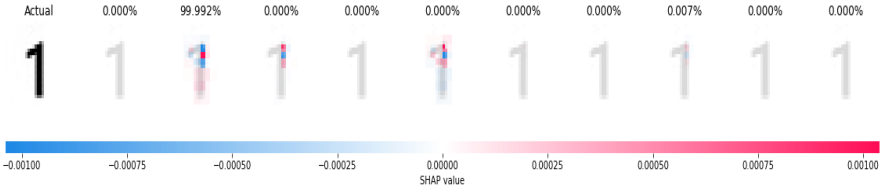
The subsequent images in Figure 24 depict the outcomes of applying SHAP to the CNN model for the digit 1 across four distinct fonts. In these visualizations, the red and blue regions correspond to SHAP values, where red signifies a positive impact on the prediction, and blue indicates a negative impact. Each column in the images represents the probability of the prediction for the respective classes 0-9. These visual representations offer insights into the factors influencing the model's predictions and provide a comprehensive understanding of the model's decision-making process across various fonts.

The SHAP parameters employed for the Modified MNIST dataset remain consistent with those used for analyzing dogs and cats. This uniformity in SHAP parameters ensures a standardized and comparable approach to interpreting the model's predictions across different datasets, fostering a cohesive analysis methodology.

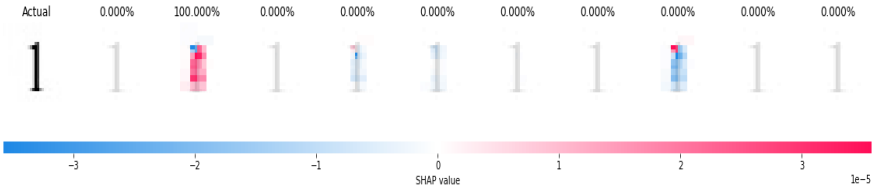
Font #1



Font #2



Font #3



Font #4

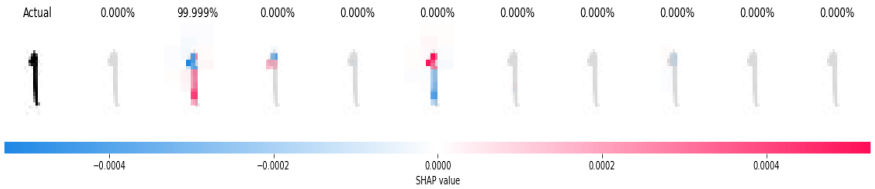


Figure 24 illustrates the outcomes of applying SHAP to the CNN model for the digit 1 across four distinct fonts.

Calculating the SHAP value for each pixel and considering the separation of SHAP values for each font (feature) yields 784 values for each image with dimensions of 28 x 28 pixels. Then employing the **equation 5** to compute the average of the absolute SHAP values, the importance of that particular font is derived as a numerical value. Aggregating all images in the testing dataset with the same font enables the determination of the overall importance of that font based on the CNN model developed. Subsequently, upon calculating the SHAP values for all testing images with different fonts, a comparison of font importance (feature importance) can be made, and the results are detailed in the Table 4.

Table 4 SHAP Values and Global Mean Predictions for each Font (Feature) within MNIST Dataset

Font	**SHAP Value x 10^{-3}**	**Global Mean Prediction**
Font #1	1.32	99.4 %
Font #2	1.39	100 %
Font #3	1.31	99.8 %
Font #4 (Handwritten)	1.23	91.4 %

SHAP values offer a quantitative measure of the impact each pixel or feature has on the model's predictions. Higher SHAP values indicate a more significant contribution to the output.

Utilizing the SHAP values presented in the above table to assess the importance of each font in the Modified MNIST dataset, a clear hierarchy emerges. Digital fonts, specifically fonts 2, 3, and 1, exhibit higher importance, while handwritten fonts demonstrate the least significance. This observation aligns with the intentional modification of the dataset, prioritizing the increased importance of digital fonts.

The SHAP interpretation results (SHAP Values) validate and confirm this trend. The model's attribution of higher importance to digital fonts, as evidenced by the SHAP values, reinforces the consistency between the dataset modification strategy and the model's decision-making process. This alignment adds credibility to the interpretability of the model's predictions and

underscores the effectiveness of the chosen approach in emphasizing the desired characteristics within the dataset.

4.4 Magnetoencephalography (MEG) Images Dataset

4.4.1 Introduction

Magnetoencephalography (MEG) (Kim & Davis, 2021) is a non-invasive method of measuring the magnetic fields produced by neuronal activity in the brain. MEG has high temporal resolution on the order of milliseconds and provides a more direct assessment of brain activity than hemodynamic-based neuroimaging techniques like functional magnetic resonance imaging (fMRI) and positron emission tomography (PET).

Coherence (COH) is a metric that indicates the degree of synchronization between two signals. In the context of MEG, COH is used to investigate the functional connectivity between different brain regions. COH is calculated by taking the cross-spectral density of two signals and normalizing it by the power spectral density of each signal. COH values range from 0 to 1, with 0 representing no synchronization and 1 representing perfect synchronization between two brain areas. COH analysis has been used to study the connectivity of various cognitive processes such as attention, perception, and memory. It has also been instrumental in investigating the effects of neurological disorders such as Alzheimer's disease and epilepsy on brain connectivity (Kim & Davis, 2021).

4.4.2 The MEG Dataset

The MEG dataset comprises data from two research labs, FZ Jülich and the Human Connectome Project (HCP) (Kampel et al., 2023), and its is available only through them.

The MEG dataset utilized the coherence (COH) connectivity method, incorporating six frequency bands as features. The dataset comprises information from 126 subjects, each contributing an average of three data files, resulting in a total of 372 files. Within each file, the number of COH images varies, ranging from four to nine, corresponding to the six frequency bands. The connectivity data matrix is structured as 68×68, representing a symmetric arrangement. With a total of 2278 unique connectivity values, derived from the formula $68 \times (68 - 1)/2$, the dataset provides comprehensive insights into the neural connectivity patterns across the specified frequency bands for each subject.

The MEG dataset further includes a connectivity data matrix, structured as $68 \times 68 \times 6$, representing six connectivity "images" corresponding to different frequency bands. These frequency bands serve as features in the dataset. Additionally, the frequency bands array provides information about the frequency ranges associated with each band.

The designated names for the frequency bands are as follows:

- Delta: 1 – 4 Hz
- Theta: 4 – 8 Hz
- Alpha: 8 – 13 Hz
- Beta1: 13 – 20 Hz
- Beta2: 20 – 30 Hz
- Gamma1: 30 – 45 Hz

This information enhances the understanding of the MEG dataset by specifying the frequency bands associated with each connectivity "image" in the connectivity data matrix, facilitating a more comprehensive analysis of neural connectivity patterns across distinct frequency ranges.

The following images in Figure 25 illustrate samples of the six frequency bands.

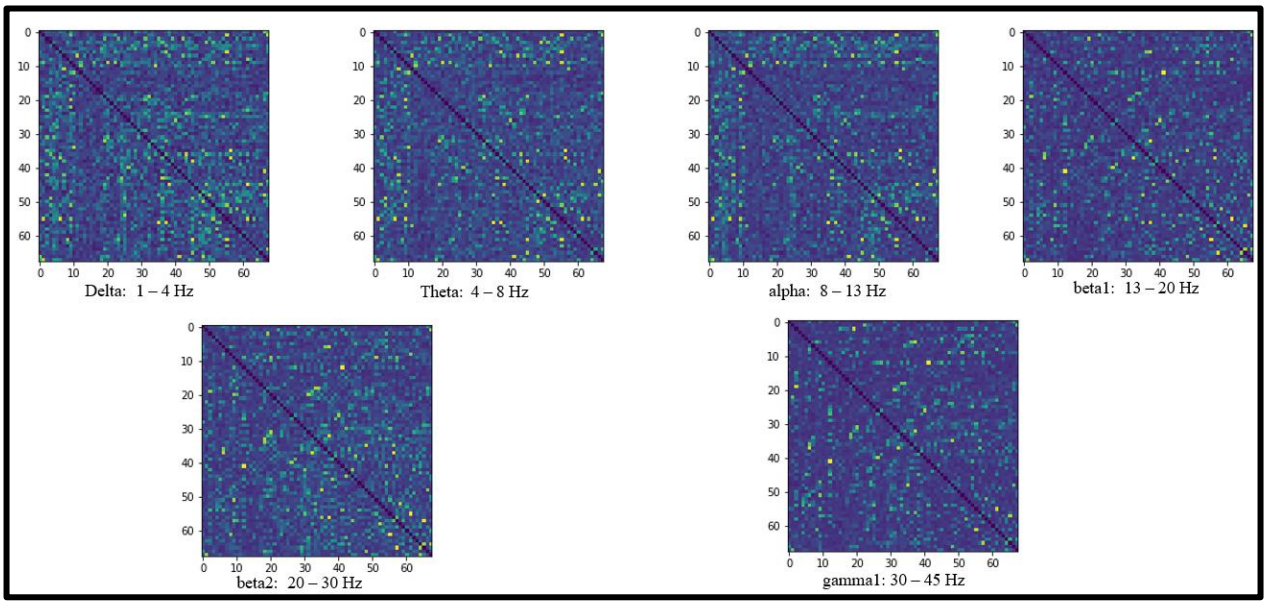


Figure 25 Samples of the six frequency bands within one class.

4.4.3 MEG Dataset Preprocessing

Files are stored in a compressed NumPy format (npz) and named with all the necessary parameters for extracting images across all frequencies. For instance, the file name "100307_unprocessed_MEG_3_Restin_4D,rest1,COH,dur60_05,ovl2.5,nfreqs06,ncon09.npz" follows a specific structure. In this example, "100307" represents the class name, "nfreqs06" indicates the inclusion of 6 frequency bands, and "ncon09" specifies that the file contains 9 COH samples.

In Python, the images representing the connectivity matrices were extracted, and subsequently saved as image files. These files are organized and stored within folders named according to their respective class names.

The MEG dataset has been strategically divided into distinct subsets to facilitate comprehensive training, validation, and analysis. The training data, constituting 50% of the dataset with 5198 images, serves as the foundational set for training the CNN model. The validation data, comprising 20% of the dataset and consisting of 2011 images, plays a critical role in evaluating model performance during training to prevent overfitting. Furthermore, 30% of the dataset, totaling 3187 images, has been earmarked as the analysis data, remaining unseen by CNN during training and validation phases. This subset serves as an independent pool for rigorous testing, and notably, these images will also form the test dataset for subsequent interpretation using SHAP. This inclusion enhances the interpretability of the model's predictions and sheds light on the contribution of each feature in the analysis process.

4.4.4 CNN Model Structure and Development

Utilizing an identical CNN model structure as employed for the MNIST dataset, as depicted in Figure 9, and adhering to the same technique and approach applied to the MNIST dataset, we maintain a configuration comprising a total of 27 layers. These layers encompass a mix of convolutional layers, pooling layers, dropout layers, a flatten layer, and dense layers. Each specific layer type assumes a pivotal role in the comprehensive process of feature extraction, spatial down-sampling, and classification within the model.

The average validation accuracy achieved by this CNN stands at an impressive 92%, which positions it favorably for the subsequent SHAP analysis and interpretation processes. Figure 28 below provides a visual representation of the accuracy curves for both training and loss which shows also that the model converge after the 8th epoch for the training and after the 13th epoch in loss curve, that offering insights into the model's learning dynamics.

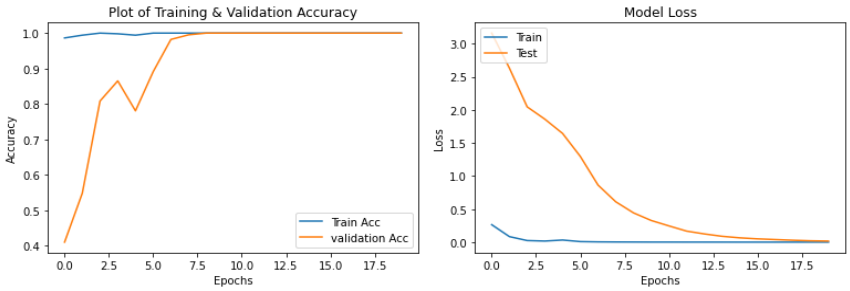


Figure 26 The accuracy and loss curves of CNN model using the MEG Dataset

4.4.5 SHAP Interpretation for MEG Model

Within this section, the CNN model utilizing the MEG dataset undergoes SHAP analysis to interpret the model results and ascertain the importance of features, specifically the six frequency bands. The application of equation 5 is employed to quantify these features numerically, as visual inspection proves impractical at this stage. Furthermore, upon the application of SHAP to previously unseen data, the ensuing results unfold to provide valuable insights. Table 5 provides the significance of each frequency that calculated as SHAP values, it should be emphasized that the model achieves a prediction accuracy of nearly 100%.

Table 5 The SHAP values for each frequency (Feature) according to the CNN Model

Frequancy	**SHAP Value x 10^{-5}**
Delta	7.310
Theta	7.266

Alpha	7.396
Beta1	7.329
Beta2	7.375
Gamma1	7.468

The interpretation of the CNN model utilizing the six frequency bands as features is elucidated through the analysis of SHAP values. These values quantify the impact of each frequency band on the model's predictions. Notably, the SHAP values indicate that all six frequency bands—Delta, Theta, Alpha, Beta1, Beta2, and Gamma1—contribute significantly to the model's output, with subtle variations in their individual influences. The positive nature of these SHAP values suggests that higher values in each frequency band correspond to an increased likelihood of the predicted outcome. Of particular interest, the SHAP values for Gamma1 stand out as the highest among the frequency bands, implying that Gamma1 may possess a slightly stronger influence on the CNN model's predictions compared to the other bands. This nuanced understanding facilitates a more informed interpretation of the model's decision-making process, aiding in the identification of the most influential frequency band in the context of the given dataset and prediction task.

Chapter Five: Conclusion and Future Work

5.1 Introduction

This chapter serves as the synthesis of interpretation into dogs and cats, the MNIST dataset, and MEG models. It systematically presents the results obtained at each stage and thoroughly examines the strengths and weaknesses inherent in our approach. The central focus lies in unraveling the implications of these results.

Rather than a mere enumeration of validation and prediction accuracy, the chapter engages in an in-depth analysis of the findings. The examination extends beyond the surface, incorporating tools like SHAP and LIME for both visual inspection and mathematical scrutiny. These tools provide a nuanced understanding of the models, offering insights that transcend mere numerical representation.

As a conclusion of this thesis, the objective is to distill the key insights. The closing remarks do not confine themselves to the specifics of dogs, cats, MNIST, and MEG models; instead, they unfold broader implications that transcend the boundaries of each individual stage. This concluding chapter invites readers to traverse the interconnected pathways of our exploration, unveiling the significance of our findings.

5.2 Stage 1: Dogs and Cats Model Interpretation

The initial phase of this study involved constructing a robust CNN as a powerful model for distinguishing between images of dogs and cats. Subsequently, LIME and SHAP

technologies were applied to interpret the model's predictions. The objective was to assess if the model's predictions relied on meaningful features within the dogs and cat’s dataset. Encouragingly, both LIME and SHAP yielded impressive results.

In the case of LIME, the technology highlighted superpixel regions using two colors, aiding analysts in understanding whether these regions influenced the model's prediction towards a dog or a cat. Similarly, SHAP identified the most crucial pixels in the images, also utilizing two colors to facilitate interpretation of whether the depicted features were indicative of a dog or a cat.

Analysis of the results from both SHAP and LIME demonstrated that they successfully identified important regions, commonly referred to as features. These highlighted regions included the body, ears, eyes, and overall shape, aligning with expectations and providing robust interpretations.

Moreover, the outcomes of both LIME and SHAP at this stage demonstrated their impact on facilitating human visual inspection to comprehend the interpretation results. The combined use of these technologies and visual inspection enhances the understanding of the model's decision-making process in distinguishing between dogs and cats.

5.3 Stage 2: MNIST Model Interpretation

The modification of the MNIST dataset involved incorporating specific fonts to mirror the characteristics of the MEG dataset, which is distinguished by clear features, namely the six

frequency bands. This adjustment aimed to align the MNIST dataset with the MEG dataset, enabling the application of consistent interpretation techniques.

For the mathematical approach, SHAP was chosen over LIME. This decision stems from LIME's computation of the importance of superpixel regions, which lacks symmetry and systematic consistency. In contrast, SHAP calculates importance through SHAP values for each pixel, highlighting them using two colors to denote positive and negative values importance for the final interpretation result.

Visual inspection affirmed the correctness of the CNN predictions, emphasizing the accurate delineation of digit borders and shapes. The mathematical approach, employing equation 5 to assess the importance of different fonts, corroborated this accuracy by revealing higher importance for the added digital fonts compared to handwritten ones.

Notably, the model's exceptional performance resulted in minimal disparities across features. The font images introduced subtle changes to digit shapes, resulting in correspondingly slight alterations in predictions and SHAP analysis. This underscores a challenge — distinguishing whether variations in SHAP values are substantial, particularly for high-performing models.

Considering this, the study contemplated the scenario of poor to moderately performing models, suggesting a test using the model at smaller number of epochs. However, the narrative doesn't explore further into the specific outcomes or analyses for these epochs, leaving room for exploration in subsequent discussions or investigations.

5.4 Stage 3: MEG Model Interpretation

In the third stage, subsequent to confirming the alignment of visual inspection and the mathematical approach in terms of interpretation results and judgment, we exclusively apply the mathematical technique of SHAP on the MEG data.

An important observation surfaces from this analysis, particularly highlighting Gamma1, which stands out by displaying the highest SHAP values among the various frequency bands. This observation suggests that Gamma1 may exert a slightly more pronounced influence on the predictions made by the CNN model compared to the other frequency bands. This nuanced understanding significantly contributes to the interpretation of the model's decision-making process, offering valuable insights into the hierarchical influence among the frequency bands within the specific context of the given dataset and prediction task.

One of the primary advantages of incorporating XAI into neuroscience research is its ability to enhance transparency and trust in AI-driven analyses. Unlike traditional AI models, which provide outputs without explanations, XAI techniques such as attention mechanisms and feature attribution methods offer insights into the factors influencing model predictions. This transparency not only fosters confidence in the reliability of AI-driven analyses but also enables researchers to validate and refine their models based on the interpretability of their outputs.

Moreover, XAI techniques can aid in identifying biomarkers and patterns indicative of neurological disorders, thereby facilitating early detection and intervention. By elucidating

the features contributing to model predictions, XAI enables researchers to pinpoint subtle changes in brain activity or structure associated with specific disorders. This knowledge can inform the development of diagnostic tools and personalized treatment strategies tailored to individual patients, ultimately improving clinical outcomes.

Furthermore, XAI can support interdisciplinary collaborations between neuroscientists, data scientists, and clinicians by bridging the gap between complex AI algorithms and domain-specific knowledge. By providing interpretable insights into AI models' decision-making processes, XAI fosters communication and collaboration across diverse fields, facilitating the translation of research findings into clinical practice.

However, despite its potential benefits, the feasibility of implementing XAI in neuroscience faces several challenges. These include the computational complexity of XAI techniques, the need for large and diverse datasets for training interpretable models, and the integration of XAI into existing neuroimaging and clinical workflows. Addressing these challenges requires interdisciplinary efforts and innovative approaches to ensure the effective integration of XAI into neuroscientific research and clinical practice.

When evaluating model performance, especially in scenarios where computational efficiency is paramount, considering time complexity becomes crucial. Time complexity assesses how computational time scales with the size of input data, directly impacting the speed at which predictions are made. Models with lower time complexity are generally preferred as they can handle larger datasets and provide quicker predictions. However, achieving a balance

between time complexity and predictive accuracy is essential. It's important to note that allocating time for explained predictions is necessary to ensure reliable results. Comparing different models based on their time complexity allows for informed decisions regarding which model best fits the computational constraints of a given task. It enables practitioners to select models that meet both efficiency and accuracy requirements, ensuring optimal performance in various applications.

5.5 Future Work

The current study has laid a foundation for understanding the interpretability of the CNN model using Shapley values. Future research can build upon this groundwork, exploring several avenues to enhance our comprehension and refine methodologies.

Firstly, there is a need for a comprehensive comparative analysis between Shapley values and alternative interpretability metrics, such as permutation importance or tree-based feature importance. This comparative investigation would shed light on the relative strengths and weaknesses of different techniques, contributing to a more holistic understanding of model interpretability.

Another area of exploration is the dynamic interaction between features during different stages of model training. Future research could delve into assessing how Shapley values evolve throughout the training process, investigating whether certain features become more influential over time.

Finally, future studies should consider extending the analysis to models with poor to moderate performance. Understanding how Shapley values behave in different performance scenarios is crucial for a more comprehensive evaluation of their utility.

In essence, future work should aim to refine and expand upon the current study, addressing specific aspects of model interpretability, exploring additional methodologies, and applying insights gained to diverse contexts and model scenarios. This iterative process will contribute to the ongoing evolution of interpretability techniques in the realm of machine learning.

References

Abdollahi, A., & Pradhan, B. (2023). Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. *Science of The Total Environment*, *879*, 163004. https://doi.org/10.1016/j.scitotenv.2023.163004

Alelyani, S. (2021). Detection and Evaluation of Machine Learning Bias. *Applied Sciences*, *11*(14), 6271. https://doi.org/10.3390/app11146271

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021a). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021b). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Beohar, D., & Rasool, A. (2021). Handwritten Digit Recognition of MNIST dataset using Deep Learning state-of-the-art Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 542–548. https://doi.org/10.1109/ESCI50559.2021.9396870

Bhandari, M., Yogarajah, P., Kavitha, M. S., & Condell, J. (2023). Exploring the Capabilities of a Lightweight CNN Model in Accurately Identifying Renal Abnormalities: Cysts, Stones, and Tumors, Using LIME and SHAP. *Applied Sciences*, *13*(5), 3125. https://doi.org/10.3390/app13053125

Brandt, R., Raatjens, D., & Gaydadjiev, G. (2023). *Precise Benchmarking of Explainable AI Attribution Methods*.

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey. *Machine Learning and Knowledge Extraction*, *3*(4), 966–989. https://doi.org/10.3390/make3040048

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs*.

Chen, S. (2021). *Interpretation of multi-label classification models using shapley values*.

Chen, Y. (2022). The Bias Detection in Deep Learning Models. *Machine Learning Journal*, *17*(3), 211–229.

Chu, J., Cai, J., Song, H., Zhang, Y., & Wei, L. (2020). A Novel Bilinear Feature and Multi-Layer Fused Convolutional Neural Network for Tactile Shape Recognition. *Sensors*, *20*, 5822. https://doi.org/10.3390/s20205822

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). *EMNIST: an extension of MNIST to handwritten letters*.

da Silva Castanheira, J., Orozco Perez, H. D., Misic, B., & Baillet, S. (2021). Brief segments of neurophysiological activity enable individual differentiation. *Nature Communications*, *12*(1), 5713. https://doi.org/10.1038/s41467-021-25895-8

Daniele, A., van Krieken, E., Serafini, L., & van Harmelen, F. (2023). Refining neural network predictions using background knowledge. *Machine Learning*, *112*(9), 3293–3331. https://doi.org/10.1007/s10994-023-06310-3

Dardouillet, P., Benoit, A., Amri, E., Bolon, P., Dubucq, D., & Crédoz, A. (2022, August). Explainability of Image Semantic Segmentation Through SHAP Values. *ICPR-XAIE -26TH International Conference on Pattern Recognition 2-Nd Workshop on Explainable and Ethical AI*. https://hal.science/hal-03719597

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/10.1109/SP.2016.42

Demuru, M., & Fraschini, M. (2020). EEG fingerprinting: Subject-specific signature based on the aperiodic component of power spectrum. *Computers in Biology and Medicine*, *120*, 103748. https://doi.org/10.1016/j.compbiomed.2020.103748

Demuru, M., Gouw, A. A., Hillebrand, A., Stam, C. J., van Dijk, B. W., Scheltens, P., Tijms, B. M., Konijnenberg, E., ten Kate, M., den Braber, A., Smit, D. J. A., Boomsma, D. I., & Visser, P. J. (2017). Functional and effective whole brain connectivity using magnetoencephalography to identify monozygotic twin pairs. *Scientific Reports*, *7*(1), 9685. https://doi.org/10.1038/s41598-017-10235-y

Dhillon, A., & Verma, G. K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence*, *9*(2), 85–112. https://doi.org/10.1007/s13748-019-00203-0

Dieber, J., & Kirrane, S. (2020). *Why model why? Assessing the strengths and limitations of LIME*.

Doe, J., & Brown, A. (2023). Feature Significance in Convolutional Neural Networks. *Neural Computing & Applications*, *31*(2), 59–71.

Fryer, D., Strümke, I., & Nguyen, H. (2021). *Shapley values for feature selection: The good, the bad, and the axioms*.

Garreau, D., & Mardaoui, D. (2021, July). What does LIME really see in images? *ICML 2021 - 38th International Conference on Machine Learning*. https://hal.science/hal-03233014

Gashi, M., Vuković, M., Jekic, N., Thalmann, S., Holzinger, A., Jean-Quartier, C., & Jeanquartier, F. (2022). State-of-the-Art Explainability Methods with Focus on Visual Analytics Showcased by Glioma Classification. *BioMedInformatics*, *2*(1), 139–158. https://doi.org/10.3390/biomedinformatics2010009

Giovannetti, A., Susi, G., Casti, P., Mencattini, A., Pusil, S., López, M. E., Di Natale, C., & Martinelli, E. (2021). Deep-MEG: spatiotemporal CNN features and multiband ensemble classification for predicting the early signs of Alzheimer's disease with magnetoencephalography. *Neural Computing and Applications*, *33*(21), 14651–14667. https://doi.org/10.1007/s00521-021-06105-4

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Guo, Y., Chen, J., Du, Q., Hengel, A. Van Den, Shi, Q., & Tan, M. (2016). *The Shallow End: Empowering Shallower Deep-Convolutional Networks through Auxiliary Outputs*.

Hasasneh, A., Kampel, N., Sripad, P., Shah, N. J., & Dammers, J. (2018). Deep Learning Approach for Automatic Classification of Ocular and Cardiac Artifacts in MEG Data. *Journal of Engineering*, *2018*, 1–10. https://doi.org/10.1155/2018/1350692

Huang, A., Li, J., & Shankar, N. (n.d.). *Interpretability*. Retrieved December 25, 2023, from https://blog.ml.cmu.edu/2020/08/31/6-interpretability/

Huang, C., Kong, C., & Lucey, S. (2017). *CNNs are Globally Optimal Given Multi-Layer Support*.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). *Densely Connected Convolutional Networks*.

James, T., Schillaci, C., & Lipani, A. (2021). Convolutional neural networks for water segmentation using sentinel-2 red, green, blue (RGB) composites and derived spectral indices. *International Journal of Remote Sensing*, *42*(14), 5338–5365. https://doi.org/10.1080/01431161.2021.1913298

Jang, R. (2023). *Learning representations by forward-propagating errors*.

Kampel, N., Kiefer, C. M., Shah, N. J., Neuner, I., & Dammers, J. (2023). Neural fingerprinting on MEG time series using MiniRocket. *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1229371

Kim, J. A., & Davis, K. D. (2021). Magnetoencephalography: physics, techniques, and applications in the basic and clinical neurosciences. *Journal of Neurophysiology*, *125*(3), 938–956. https://doi.org/10.1152/jn.00530.2020

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5686–5697. https://doi.org/10.1145/2858036.2858529

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kun, Y., Chunqing, G., & Yuehui, G. (2022). An Optimized LIME Scheme for Medical Low Light Level Image Enhancement. *Computational Intelligence and Neuroscience*, *2022*, 1–8. https://doi.org/10.1155/2022/9613936

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, Y.-G., Oh, J.-Y., Kim, D., & Kim, G. (2023). SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. *Journal of Electrical Engineering & Technology*, *18*(1), 579–588. https://doi.org/10.1007/s42835-022-01161-9

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Lu, J., Feng, J., Fan, Z., Huang, L., Zheng, C., & Li, W. (2018). *Automated Strabismus Detection based on Deep neural networks for Telemedicine Applications*.

Lundberg, S., & Lee, S.-I. (2017a). *A Unified Approach to Interpreting Model Predictions*.

Lundberg, S., & Lee, S.-I. (2017b). *A Unified Approach to Interpreting Model Predictions*.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Mahardi, Wang, I.-H., Lee, K.-C., & Chang, S.-L. (2020). Images Classification of Dogs and Cats using Fine-Tuned VGG Models. *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, 230–233. https://doi.org/10.1109/ECICE50847.2020.9301918

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*.

Marcilio, W. E., & Eler, D. M. (2020). From explanations to feature selection: assessing SHAP values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. https://doi.org/10.1109/SIBGRAPI51738.2020.00053

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences*.

Mishkin, D., Sergievskiy, N., & Matas, J. (2016). *Systematic evaluation of CNN advances on the ImageNet*. https://doi.org/10.1016/j.cviu.2017.05.007

Mokhtari, K. El, Higdon, B. P., & Başar, A. (2019). Interpreting financial time series with SHAP values. *Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, 166–172.

Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Molnar, C., Casalicchio, G., & Bischl, B. (2020). *Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges*. https://doi.org/10.1007/978-3-030-65965-3_28

Mouches, P., Dejean, T., Jung, J., Bouet, R., Lartizien, C., & Quentin, R. (2023). *Time CNN and Graph Convolution Network for Epileptic Spike Detection in MEG Data*.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Nishani, E., & Çiço, B. (2017). A Systematic Mapping Study of Computer Vision Approaches based on Deep Learning and Neural Network. *Proceedings of the 8th Balkan Conference in Informatics*, 1–8. https://doi.org/10.1145/3136273.3136301

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2021). *Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital*. https://doi.org/10.1016/j.cmpb.2021.106584

Olisah, C. C., Trewhella, B., Li, B., Smith, M. L., Winstone, B., Whitfield, E. C., Fernández, F. F., & Duncalfe, H. (2024). *Convolutional Neural Network Ensemble Learning for Hyperspectral Imaging-based Blackberry Fruit Ripeness Detection in Uncontrolled Farm Environment*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Russell, S. J. , & N. (2021). *Artificial Intelligence A Modern Approach* (4th ed).

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*(2), 272. https://doi.org/10.1007/s42452-021-04148-9

Sabour, S., Frosst, N., & Hinton, G. E. (2017). *Dynamic Routing Between Capsules*.

Sachin. (2019). *Kaggle: Cats-vs-Dogs image dataset for binary classification.* Https://Www.Kaggle.Com/Datasets/Shaunthesheep/Microsoft-Catsvsdogs-Dataset.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2021). *Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods*. http://arxiv.org/abs/2111.02398

Sareen, E., Zahar, S., Ville, D. Van De, Gupta, A., Griffa, A., & Amico, E. (2021). Exploring MEG brain fingerprints: Evaluation, pitfalls, and interpretations. *NeuroImage*, *240*, 118331. https://doi.org/10.1016/j.neuroimage.2021.118331

Sarker, I. H. (2021a). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Sarker, I. H. (2021b). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Schallner, L., Rabold, J., Scholz, O., & Schmid, U. (2020). *Effect of Superpixel Aggregation on Explanations in LIME – A Case Study with Biological Data* (pp. 147–158). https://doi.org/10.1007/978-3-030-43823-4_13

Scott Lundberg. (2018). *SHAP documentation*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. https://doi.org/10.1007/s11263-019-01228-7

Shah, S. S., & Sheppard, J. W. (2020a). Evaluating Explanations of Convolutional Neural Network Image Classifications. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207129

Shah, S. S., & Sheppard, J. W. (2020b). Evaluating Explanations of Convolutional Neural Network Image Classifications. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207129

sherena johnson. (2010). *NIST Special Database 19*. https://www.nist.gov/srd/nist-special-database-19

Simonyan, K., & Zisserman, A. (2014a). *Very Deep Convolutional Networks for Large-Scale Image Recognition*.

Simonyan, K., & Zisserman, A. (2014b). *Very Deep Convolutional Networks for Large-Scale Image Recognition*.

Simonyan, K., & Zisserman, A. (2014c). *Very Deep Convolutional Networks for Large-Scale Image Recognition*.

Singh, M. K., & Kumar, B. (2023). *Fine Tuning the Pre-trained Convolutional Neural Network Models for Hyperspectral Image Classification Using Transfer Learning* (pp. 271–283). https://doi.org/10.1007/978-981-19-7892-0_21

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). *Striving for Simplicity: The All Convolutional Net*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). *Going Deeper with Convolutions*.

Tirumala, S. S. (2021). Exploring Neural Network Layers for Knowledge Discovery. *Procedia Computer Science*, *193*, 173–182. https://doi.org/10.1016/j.procs.2021.10.017

van der Velden, B. H. M. ., Ragusi, M. A. A., Janse, M. H. A., Loo, C. E., & Gilhuijs, K. G. A. (2020). Interpretable deep learning regression for breast density estimation on MRI. In H. K. Hahn & M. A. Mazurowski (Eds.), *Medical Imaging 2020: Computer-Aided Diagnosis* (p. 69). SPIE. https://doi.org/10.1117/12.2549003

Ventura, F., Cerquitelli, T., & Giacalone, F. (2018). *Black-Box Model Explained Through an Assessment of Its Interpretable Features* (pp. 138–149). https://doi.org/10.1007/978-3-030-00063-9_15

Visani, G., Bagli, E., & Chesani, F. (2020). *OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms*.

Wang, M., Liu, X., Gao, Y., Ma, X., & Soomro, N. Q. (2017). Superpixel segmentation: A benchmark. *Signal Processing: Image Communication*, *56*, 28–39. https://doi.org/10.1016/j.image.2017.04.007

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yang, F., Sun, Q., Jin, H., & Zhou, Z. (2020). *Superpixel Segmentation with Fully Convolutional Networks*.

Zafar, M. R., & Khan, N. (2021). Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction*, *3*(3), 525–541. https://doi.org/10.3390/make3030027

Zeiler, M. D., & Fergus, R. (2013). *Visualizing and Understanding Convolutional Networks*.

Zhang, X., & Gao, J. (2020). Measuring Feature Importance of Convolutional Neural Networks. *IEEE Access*, *8*, 196062–196074. https://doi.org/10.1109/ACCESS.2020.3034625

Zhao, X., Huang, W., Huang, X., Robu, V., & Flynn, D. (2020). *BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations*.

ملخص الرسالة

في السنوات الأخيرة، قدمت نماذج التعلم العميق تطبيقات مختلفة في شتى المجالات وخصوصاً في المجالات الطبية، ومن ضمنها مجال العلوم العصبية، حيث إن التأكد من تفسير النتائج التي تتنبأ بها نماذج التعلم العميق كانت ولا زالت تحدٍ مهم وخصوصاً استكشاف الطبقات الخفية في هذه النماذج وهو ما يسمى بتفسير الصندوق الأسود Black box. يقوم هذا البحث، المعنون ب 'فحص السمات في الطبقات الخفية لنماذج التعلم العميق المطبقة على بيانات علم الأعصاب'، باستخدام تقنيات التعلم العميق المُفسِرة XAI لإلقاء الضوء على عمليات اتخاذ القرارات لشبكات الأعصاب التكاملية المُطبَّقة على البيانات العصبية.

مع التركيز على بيانات MEG والتي تحتوي على ست حزمات من الترددات كميزات Features، يتعثر التفسير البصري لنتائج LIME وSHAP بسبب تعقيد صورMEG COH . ومع ذلك، يفسر هذا البحث نماذج التعلم العميق للتميز بين الكلاب والقطط ومجموعة الأرقام MNIST بدقة عالية، مُظهرًا فعالية LIME وSHAP جنبًا إلى جنب مع التفسير البصري والمعادلات الرياضية للقيم النموذج وبشكل حاسم، يتوسع البحث لكشف الطبقات الخفية التي كانت تُعْتَبَر تاريخيا صناديق سوداء في نماذج التعلم العميق، حيث يقوم هذا البحث بدراسة أهمية المميزات باستخدام حساب قيم SHAP لكل بكسل في صور MEG والتي تقوم بدرها بحساب أهمية الميزات التي يستند عليها نموذج التعلم العميق، مما يوفر رؤى في عمليات اتخاذ القرارات وتعزيز شفافية النموذج بشكل عام.

فيما يخص بيانات علم الأعصاب وتحديدا صور MEG، يحدد البحث تردد Gamma1 بأعلى قيم SHAP، مشيرًا إلى تأثيره البارز على توقعات التعلم العميق. يسهم هذا الفهم المعقد في تفسير عملية اتخاذ القرارات الخاصة بالنموذج، مقدمًا رؤى قيمة حول التأثير التسلسلي بين حزم التردد داخل مجالات معينة ضمن البيانات المعطاة ومهمة التنبؤ. في الختام، يعالج هذا البحث تحدي التفسير في علم الأعصاب من خلال الكشف عن الصندوق الأسود في الطبقات الخفية، مما يعزز عمليات اتخاذ القرارات المستنيرة في تطبيقات العلوم العصبية.