**<Primary Thesis Title> EXPLAINABLE AI FOR DEEPFAKE DETECTION**

**<Secondary Title>Network Dissection Approach**

submitted as a requirement for the degree of

**Master of Computer Science focusing on Big Data and Artificial Intelligence**

Berlin, Germany, August 10, 2023



**by**

**<First Name(s)>Nazneen Mansoor**

Primary Thesis Supervisor: **Prof. Dr. Alexander Iliev**

Associate Thesis Supervisor: **Prof. Dr. Reiner Creutzburg**

# ABSTRACT

The rapid advancement of technology in multimedia has led to the emergence of malicious activities, such as DeepFake manipulation in politics and entertainment. As manipulated visuals become increasingly realistic, the need for a reliable and explainable DeepFake detection approach becomes essential. This research proposes a DeepFake detection technique employing the Network dissection algorithm to achieve interpretability.

The study involves two stages: (1) Detection of forged images using state-of-the-art CNNs (VGG-16, ResNet-50, and Inception V3) with a comparative analysis of their performance. (2) Implementation of the Network dissection algorithm to interpret and understand the internal representations of these models in classifying images as real or fake. The algorithm provides insights into the facial concepts learned by the models for decision-making.

Experimental results indicate that all CNN models employed in the study achieve good performance, with F1-scores ranging between 0.8 and 0.9. The proposed approach's explainability is demonstrated through interpretability reports generated for the DeepFake detection models. The interpretability aspect offers a deeper understanding of how models function and the features they learn.

While various DeepFake detection models exist, most lack an explanation for their classifications. This research addresses this gap by providing insights into how the models distinguish between real and manipulated images through interpretable neuronal representations. Overall, interpretability proves to be a crucial aspect of deep neural networks, shedding light on their hierarchical structure.

Keywords: Explainable Artificial Intelligence, Deep Learning, DeepFake Detection, Explainability, Convolutional Neural Network, VGG-16, ResNet-50, Inception V3, Network Dissection, Face Dictionary, Model Interpretability

# 

# ACKNOWLEDGEMENTS

I would like to use this opportunity to convey my heartfelt thanks and appreciation to everyone who helped and contributed to the successful completion of my thesis research project.

First and foremost, I am grateful to my supervisor, Prof. Dr. Alexander Iliev, whose guidance, expertise, and unwavering support have been instrumental in shaping the trajectory of my research. His valuable insights and constructive feedback have been invaluable assets in refining my work, and I am truly grateful for his mentorship throughout my journey at SRH Hochschule Berlin. Second, I would like to thank Prof. Dr. Reiner Creutzburg for his tremendous guidance and support throughout my studies and research at SRH. I am truly thankful for his contributions and the impact he has had on my education and research endeavors.

I would also like to express my sincere thanks to the university faculty and administration for providing me with a conducive academic environment and access to resources that were essential to carry out this research.

I would also like to express my sincere gratitude to my partner, Raslam Showkath, for his unwavering support, understanding, and encouragement. His constant encouragement has been a source of strength throughout the challenges of this research. My deepest appreciation goes to my family for their unconditional love, encouragement, and understanding. I’m truly grateful for their presence in my life.

Lastly, I extend my thanks to all my friends and colleagues who have provided valuable feedback, assistance, and encouragement throughout the research process.

I am humbled and grateful to all those mentioned above for their support, encouragement, and belief in me. This journey would not have been possible without their presence and contributions, and for that, I am deeply thankful.

# TABLE OF CONTENTS

[ABSTRACT ii](#_Toc144983749)

[ACKNOWLEDGEMENTS iii](#_Toc144983750)

[TABLE OF CONTENTS iv](#_Toc144983751)

[LIST OF FIGURES v](#_Toc144983752)

[LIST OF TABLES vi](#_Toc144983753)

[LIST OF EQUATIONS vii](#_Toc144983754)

[LIST OF ABBREVIATIONS viii](#_Toc144983755)

[1. INTRODUCTION 1](#_Toc144983756)

[1.1. Background 1](#_Toc144983757)

[1.2. Motivation 3](#_Toc144983758)

[1.3. Research Gap 5](#_Toc144983759)

[1.4. Research Objectives 6](#_Toc144983760)

[1.5. Research Methodologies 7](#_Toc144983761)

[1.6. Scope & Limitations 8](#_Toc144983762)

[2. LITERATURE REVIEW 10](#_Toc144983763)

[2.1 DeepFake Detection 11](#_Toc144983764)

[2.2 Explainable Artificial Intelligence 12](#_Toc144983765)

[2.2.1 Overview and Different XAI Methods 12](#_Toc144983766)

[2.2.2 XAI on Deepfake Detection 17](#_Toc144983767)

[2.2.3 Network Dissection 18](#_Toc144983768)

[3. DESIGN AND METHODOLOGY 22](#_Toc144983769)

[3.1 Methodology 22](#_Toc144983770)

[3.2 Design Specification 28](#_Toc144983771)

[3.2.1 DeepFake Detection 28](#_Toc144983772)

[3.2.2 XAI - Dissecting DeepFake Detection Models 34](#_Toc144983773)

[4. RESULTS AND EVALUATION 39](#_Toc144983774)

[4.1 DeepFake Detection 39](#_Toc144983775)

[4.2 XAI - Network Dissection Results 39](#_Toc144983776)

[5. DISCUSSION OF FINDINGS 48](#_Toc144983777)

[5.1 DeepFake Detection 48](#_Toc144983778)

[5.2 XAI - Network Dissection 49](#_Toc144983779)

[6. CONCLUSION 51](#_Toc144983780)

[7. RESEARCH CONTRIBUTION 52](#_Toc144983781)

[8. FUTURE WORK 53](#_Toc144983782)

[LIST OF REFERENCES - 1 -](#_Toc144983783)

[AFFIDAVIT - 1 -](#_Toc144983784)

# LIST OF FIGURES

[Figure 1: Deepfake Generation Model Using Encoder-Decoder. - 2 -](#_Toc42610001)

[Figure 2: Number of DeepFakes Papers in Years from 2014 to 2022. - 4 -](#_Toc42610002)

[Figure 3: The Number of Publications Related to XAI. - 4 -](#_Toc42610003)

[Figure 4: GAN Flow Diagram with Basic Function - 10 -](#_Toc42610004)

[Figure 5: Trend in XAI. - 13 -](#_Toc42610005)

[Figure 6: Explainable Artificial Intelligence (XAI) Approach. -14 -](#_Toc42610006)

[Figure 7: Feature Visualization. -19 -](#_Toc42610007)

[Figure 8: Network Dissection Method Illustration. - 20 -](#_Toc42610008)

[Figure 9: Results of Network Dissection. - 21 -](#_Toc42610009)

[Figure 10: Stages of KDD Process. -23 -](#_Toc42610010)

[Figure 11: Implementation Stages for DF Detection. - 30 -](#_Toc42610011)

[Figure 12: Implementation Stages for XAI. - 30 -](#_Toc42610012)

[Figure 13: CNN Architecture for DeepFake Detection Models - 31 -](#_Toc42610013)

[Figure 14: Real and Fake Images from Training Dataset - 33 -](#_Toc42610014)

[Figure 15: XAI Approach for Dissecting CNN Models - 35 -](#_Toc42610014)

[Figure 16: DeepFake Face Dictionary - 35 -](#_Toc42610014)

[Figure 17: Number of Images per Facial Concept in our DeepFake Face Dictionary - 36 -](#_Toc42610014)

[Figure 18: Overview of Network Dissection for DeepFake Detection Models - 36 -](#_Toc42610014)

[Figure 19: Learning Curves of DeepFake Detection Models - 40 -](#_Toc42610014)

[Figure 20: Distribution of Interpretable Units per Facial Region - 42 -](#_Toc42610014)

[Figure 21: Two Examples per Each Model of Different Concepts - 44 -](#_Toc42610014)

[Figure 22: Probabilistic Hierarchy Results - 44 -](#_Toc42610014)

[Figure 23: Dissection Histograms of The Three Models - 46 -](#_Toc42610014)

[Figure 24: Number of Units Within Multiple Layers for Three Models - 47 -](#_Toc42610014)

# LIST OF TABLES

[Table 1: XAI Methods. -15 -](#_Toc42608271)

[Table 2: PROS and CONS of Network Dissection. -19 -](#_Toc42608271)

[Table 3: Details of Dissected DeepFake Detection CNN models. - 37 -](#_Toc42608271)

[Table 4: Comparison of Performance of CNN Models. - 39 -](#_Toc42608271)

[Table 5: Details of Interpretations by CNN Models. - 42 -](#_Toc42608271)

# LIST OF EQUATIONS

[Equation 1: Model Accuracy Calculation. - 26 -](#_Toc42608277)

[Equation 2: Precision Calculation. - 26 -](#_Toc42608277)

[Equation 3: Recall Calculation. - 26 -](#_Toc42608277)

[Equation 4: F1-Score Calculation. - 27 -](#_Toc42608277)

[Equation 5: IoU Score Calculation.. - 37 -](#_Toc42608277)

# LIST OF ABBREVIATIONS

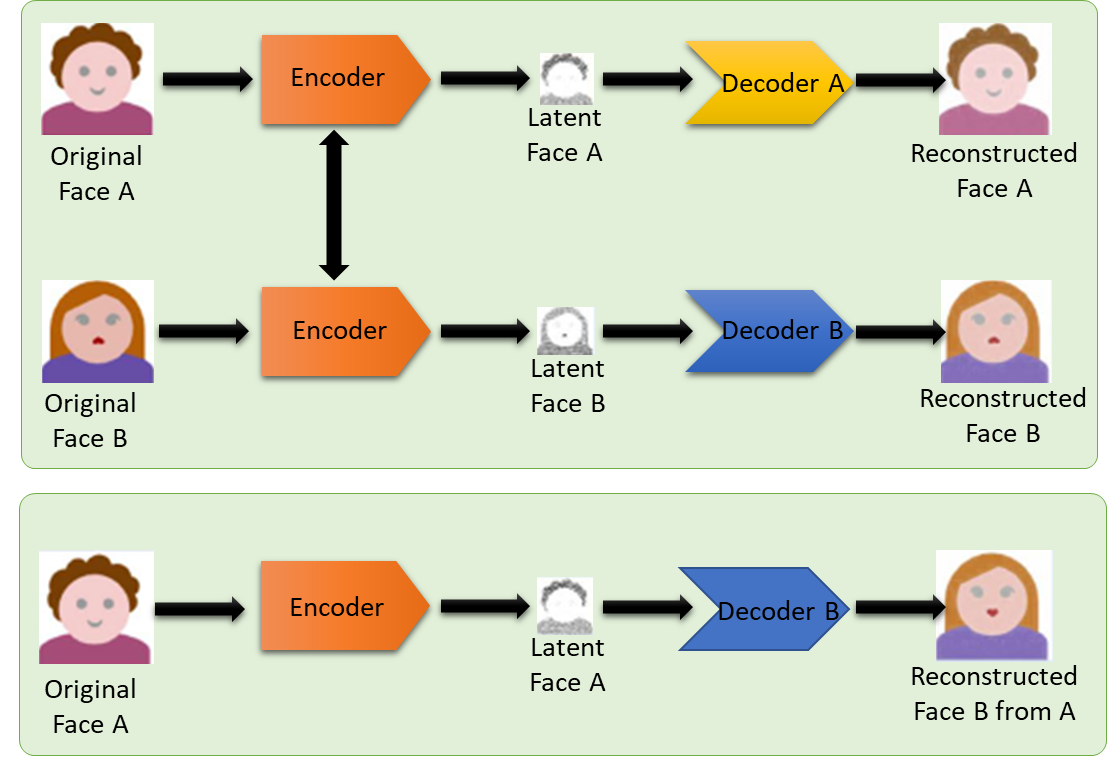
|  |  |
| --- | --- |
| Acronym | Explanation |
| AI | Artificial Intelligence |
| ALV | Activation Layer Visualization |
| CNN | Convolutional Neural Network |
| DeepLIFT | Deep Learning Important Features |
| DFDC | DeepFake Detection Challenge |
| DTD | Deep Taylor Decomposition |
| GAN | Gan Adversarial Network |
| GBP | Guided Backpropagation |
| HND | Hierarchical Network Dissection |
| IoU | Intersection over Union |
| KDD | Knowledge Discovery Database |
| LIME | Local Interpretable Model-Agnostic Explanations |
| LSTM | Long Short-Term Memory |
| ResNet | Residual Network |
| SHAP | Shapley Additive Explanations |
| VGG | Visual Geometry Group |
| XAI | Explainable Artificial Intelligence |

# INTRODUCTION

## Background

With the widespread growth in modern technology concerning mobile phones, and other editing tools, it has become convenient for the creation and distribution of digital content with ease to access social media platforms [9]. The immense research in the field of computer vision, and machine learning along with a considerable amount of data availability makes it harder to differentiate between altered and real media content. This has detonated social media platforms with forged videos, images, and fake content which appears realistic [9]. Lately, this has imposed a significant threat to society, business, democracy, and public discussion. All these years, the journalism industry has been struggling to refine the real content from fake news [9]. It harms the cybersecurity and national security of individuals and organizations.

Over the past years, deep learning technology has proven to be tremendously growing in computer vision. Deep learning techniques have been commonly used in all industries like healthcare, finance, automobile, etc. which provide solutions to complex problems. Nevertheless, the negative aspects of deep learning result in the creation of manipulated videos, photos, or other types of multimedia which can impose threats to democracy, privacy, and national security. These fake multimedia contents are termed ‘Deepfakes’ [15] [16]. The term “Deepfake” which is a combination of ‘deep learning’ and ‘fake’ initially appeared in November 2017 when a Reddit user anonymously published an algorithm to generate fake videos [13]. Deepfakes are the outcome of artificial intelligence algorithms that integrate, alter, and superimpose video clips, and images to generate fake videos that appear actual [14]. Deepfakes are generated using digital data which includes videos, images, or audio data to train the deep learning or machine learning models. The training process increases the chances of creating more deepfake content in the digital world. As technology advances, the accessibility and the potential to misuse them keep increasing. There are many techniques to create deepfakes including Gan adversarial networks (GAN), Variational Autoencoders, and diffusion models. The generation of Deepfakes includes training a neural network to explore a vast sample of data to replicate an individual’s movements, facial expressions, voice, and other features. With the help of AI tools, and deep learning technology face swapping and face mapping can be achieved. GANs are a class of generative algorithms with an architecture that consists of two neural networks, a Generator and a Discriminator [17]. Discriminator has a supervised approach that predicts the data as real or fake, trains on actual data, and feeds the outcome to the generator. The generator has an unsupervised learning approach which is also another neural network that has various hidden layers, activation, and loss functions [18]. The main purpose of the generator is to create fake images based on the discriminator’s feedback. GANs are termed adversarial networks because both neural networks are concurrently antagonistic to each other [9]. The backpropagation method is carried out to optimize the weight and bias of the networks until the discriminator is capable of executing the classification tasks. Some commonly used GANs to create fake content are WGAN, DCGAN, DRAGAN, LSGAN, and many more. Deepfake producers can be categorized into four different types: 1) activists or political players 2) fraudsters 3) deepfake hobbyist community 4) tenable actors like television companies [14]. Figure 1 illustrates a deepfake generation process by making use of encoder-decoder pairs.

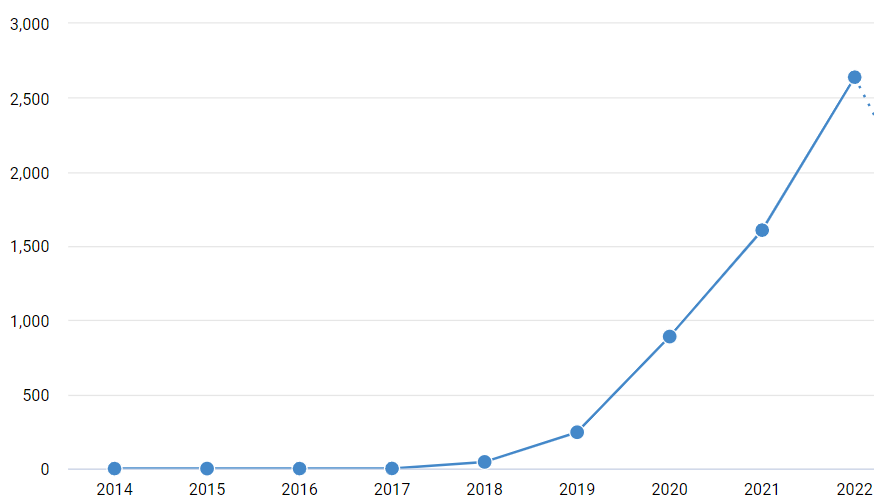


**Figure 1: Deepfake Generation Model Using Encoder-Decoder [1]**

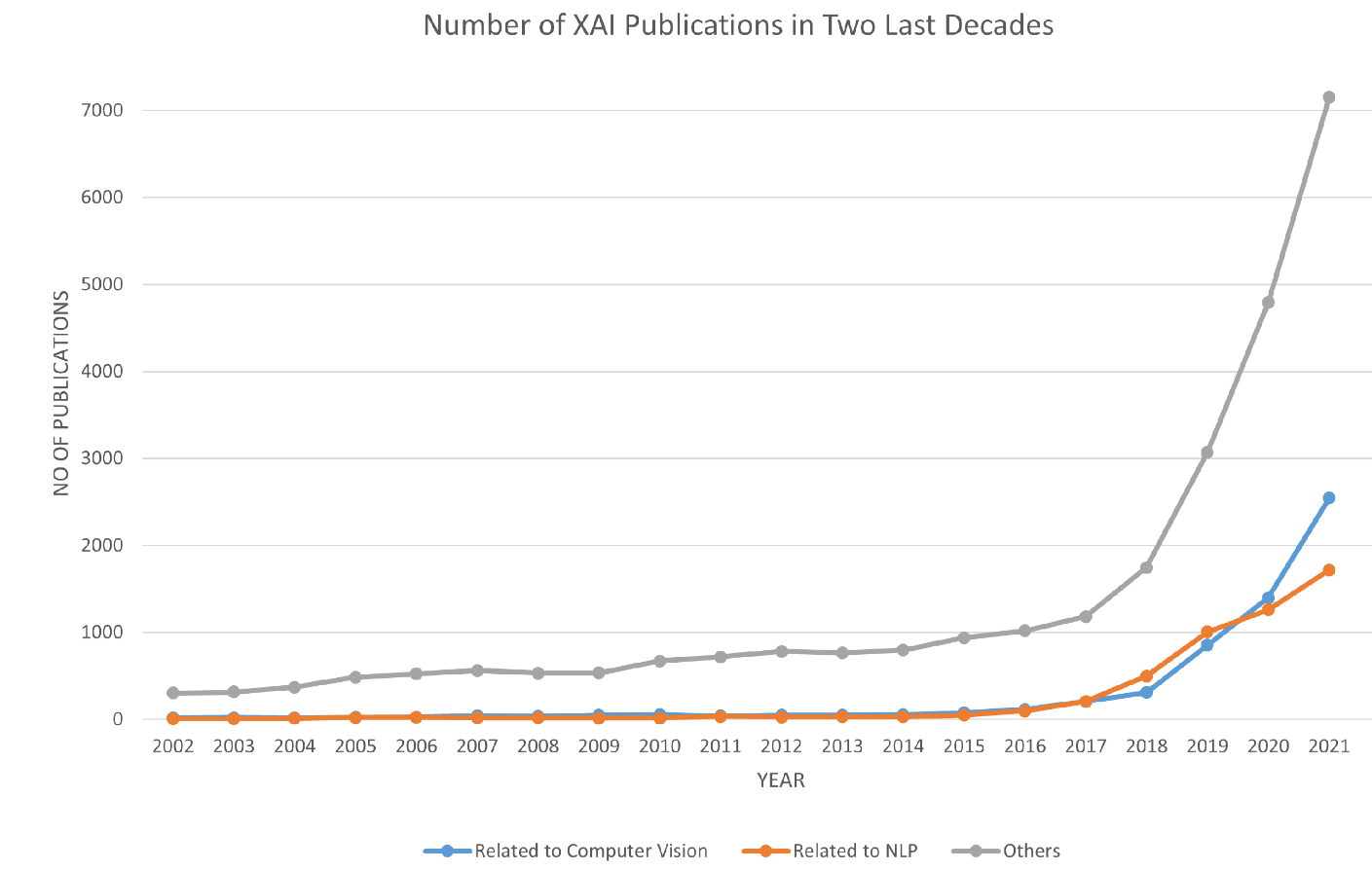
Images created using GANs have extensive applications ranging from text or image to image translation, blending and editing of photographs, face aging, generating new human poses, 3D images, and photo inpainting [9]. However, disagreeing on the benefits of a deepfake would be biased. The images generated using GAN can effectively deceive the eyes of humans. There are also several deepfake applications and software that are accessible to the public to create realistic content. Overall, deepfake generation algorithms are being developed at an alarming pace and their quality is consistently improving, thus making deepfake detection tools indispensable in today’s society. In addition to providing deep learning solutions to complex problems, it is important to understand the interpretation or internal representation of these deep learning models to trust that the AI has made the right decision and this can be achieved by using Explainable artificial intelligence (XAI). Explainable Artificial Intelligence is one of the interesting research topics which has been emerging in recent times.

## Motivation

In today’s world, the accumulation of fake images on social media platforms has become a great threat. There are different AI techniques and tools which help in creating these manipulated images. To deal with the negative aspects of deep learning algorithms, various studies have been conducted in this area of research. However, the creation of fake images is increasing to a higher level which makes the observers difficult to differentiate between fake and real images. With the advancements in the field of information and technology, these days the creation of deepfake content has become easier and it is an emerging concern in today’s digital environment. Deepfakes can be a real threat to society affecting the perceptions and sentiments of people [19]. Politicians, famous celebrities, and other known personalities are impacted by the unethical usage of these AI-based content such as the fabrication of fake news and adult material as well as online impersonation [7]. Deepfakes can also be involved in spreading false information from untrusted sources, money laundering, and data breaching activities. Given these negative implications, deepfake detection has become crucial in today’s society. In this research, a deep learning model is used to classify deepfakes into real and fake ones. To build the deep learning model, three different CNN architectures are used to compare the performance of these models for detecting DeepFakes. Additionally, this research deals with the reasoning behind the predictions made by the neural networks in classifying them as real or fake using explainable AI. The graph in Figure 2 represents the number of published papers on the topic of deepfakes until 2022. Although the entire list of papers is relatively lesser than compared to other domains, the increasing trend proves to be one of the resonant research areas in recent times. Figure 3 shows the number of publications in the field of explainable artificial intelligence for the last two decades and the plot shows how explainable artificial intelligence is blooming these days.



**Figure 2: Number of DeepFakes Papers in Years from 2014 to 2022, obtained from https://app.dimensions.ai/ at the end of 2022 with the search keyword “deepfake detection”**



**Figure 3: The Number of Publications Related to XAI [20]**

## Research Gap

In recent years, deep learning has ushered in a technological revolution, transforming the way we tackle complex problems across various domains. These advancements have empowered artificial intelligence (AI) systems to achieve unprecedented feats in image recognition, natural language processing, and more. However, as with any powerful tool, there are both positive and negative aspects to consider. One of the darker implications of deep learning technology is the proliferation of deepfake content, which poses a significant challenge to society's trust in digital media. Deepfakes are manipulated videos or images that have been generated using sophisticated AI techniques, particularly deep neural networks. These deceptive creations have the potential to incite chaos, propagate misinformation, and undermine the credibility of visual and audio evidence. Researchers and technology specialists have been working hard to build AI-based solutions for detecting deepfake content in response to this expanding threat. While progress has been made, there remains a critical issue to address: the lack of transparency and explainability in deepfake detection systems. As the research landscape currently stands, existing deepfake detection systems are proficient at distinguishing between genuine and fake media. They can flag a video or image as authentic or manipulated with remarkable accuracy. However, what these systems struggle to provide is a coherent and interpretable explanation for their classification decisions. In essence, they can tell us whether a piece of content is real or a deepfake but fall short in elucidating the underlying rationale behind their judgment. This critical gap in the ability to explain deepfake detection outcomes is a pressing concern for several reasons. Firstly, without a clear understanding of why a particular piece of content is labeled as a deepfake, it becomes challenging to trust and rely on these systems in real-world applications. Second, explanations are essential for the forensic and legal aspects of deepfake detection. Being able to provide evidence-based explanations for the classification of content can be crucial in legal proceedings or when assessing the credibility of media in journalism. Finally, interpretability plays a pivotal role in building trust with end-users and stakeholders, which is essential for the broader adoption of deepfake detection technology. In light of these challenges, it has become increasingly evident that there is a need for the development of interpretation methods specifically tailored for neural networks used in deepfake detection. With the help of explainable artificial intelligence, this research proposal aims at providing explainability for deepfake detection models, thereby improving their reliability.

## Research Objectives

The research work comprises two primary sections, each addressing a critical aspect of deepfake detection and the development of explainable AI models. In this study, the researchers employ popular pre-trained convolutional neural network (CNN) architectures, namely ResNet-50, Inception V3, and VGG16, to detect deepfake images. Additionally, they implement network dissection algorithms to enhance the interpretability and transparency of these deepfake detection models. The dataset employed for training and evaluation is Celeb-DF, a well-established dataset commonly used in the field of deepfake detection.

The first section of this research focuses on leveraging pre-trained CNN models, specifically ResNet-50, Inception V3, and VGG16, for the task of deepfake detection. Deepfake detection is a crucial application in the age of synthetic media, as it seeks to differentiate between real and manipulated content. The selection of these pre-trained models is significant because they are renowned for their effectiveness in image classification tasks. The researchers employ Celeb-DF, a dataset that presumably contains both real and deepfake images of celebrities, as the training and testing data. This dataset provides a diverse set of facial images, making it suitable for evaluating the robustness and generalization capability of the deepfake detection models. By utilizing CNN architectures, the researchers aim to demonstrate the effectiveness of convolutional neural networks in achieving high accuracy in identifying deepfake content. CNNs have proven to be highly proficient in capturing complex patterns and features within images, which is vital for detecting subtle manipulations commonly found in deepfake videos and images. The utilization of these pre-trained models further accelerates the development process and enhances the model's ability to discern real from fake.

The second section of the research introduces the concept of explainable AI (XAI) in the context of deepfake detection models. XAI is essential for building trust and understanding the decision-making processes of AI systems, especially in critical applications like deepfake detection. To achieve transparency and interpretability in the deepfake detection models, the researchers employ network dissection algorithms. Network dissection algorithms enable the dissection of CNN models into interpretable components, helping to understand the features and patterns that contribute to the model's decision. By applying these algorithms to the deepfake detection models, the researchers seek to provide insights into which specific elements or regions of the input images are crucial for identifying deepfakes. This not only enhances the model's transparency but also aids in identifying potential vulnerabilities or biases in the model's decision-making process.

## Research Methodologies

To explore the explainability potential of a deepfake detection model, a multi-step process is essential. This process involves creating a model that can classify images as either real or fake, and then incorporating explainable AI techniques to enhance the transparency of the model's decision-making process. The foundation of any deepfake detection model lies in the dataset used for training. To initiate the journey towards explainability, it is crucial to begin with a well-curated dataset that includes a diverse range of both real and fake images. This dataset serves as the training ground for the model, allowing it to learn the distinctive features and patterns that differentiate genuine from manipulated content. In selecting an appropriate deepfake dataset, researchers must consider factors such as the variety of deepfake techniques, image quality, and ethical considerations surrounding the use of synthetic content. Striking the right balance between realism and diversity within the dataset is vital to train a robust model capable of discerning intricate deepfake alterations.

Explainability is crucial for establishing trust in AI systems, especially when they are tasked with critical functions like deepfake detection. Once the model is trained to classify images, the next step is to enhance its transparency and interpretability. This is where explainable AI techniques come into play. The approach taken in this research draws inspiration from the Network Dissection algorithm proposed by previous researchers [49] [50] [52]. The fundamental idea behind Network Dissection is that certain units within convolutional neural network (CNN) architectures behave as object detectors. This insight forms the basis for understanding how the model perceives and interprets different elements within an image.

Network Dissection, as used in this research, involves dissecting various CNN architectures that have been trained on well-established facial datasets. These datasets, which are freely accessible, provide a rich source of real-world facial images, allowing for a deep exploration of the model's internal representations. The primary objective of Network Dissection is twofold: interpretation and comparison. Firstly, it aims to interpret the internal representations of the model that drive its decision-making process when classifying an image as fake or real. This involves identifying which units or neurons within the network are activated by specific features or concepts present in the image. Secondly, Network Dissection facilitates the comparison of representations learned by various deepfake detection models. By dissecting and analyzing different CNN architectures, researchers can gain insights into the similarities and differences in how these models perceive and classify deepfake content. To further enhance the interpretability of the model's internal representations, Network Dissection relies on the Intersection over Union (IoU) score. This score is used to pair specific units within the network with concepts or features present in the images. The IoU score measures the degree of overlap between the receptive field of a unit and the spatial extent of a concept. By leveraging the IoU score, researchers can precisely identify which units in the network are responsible for detecting certain elements within an image. For instance, a specific unit might be highly activated when a deepfake manipulation is detected, indicating its role as an effective detector of synthetic alterations.

## Scope & Limitations

In recent years, the field of deep learning has witnessed remarkable advancements, leading to the proliferation of manipulated images and videos. This proliferation raises significant concerns, as it can have detrimental consequences in various domains, including politics, entertainment, and journalism. To combat this growing issue, researchers have dedicated their efforts to developing systems capable of mitigating the negative impacts of manipulated multimedia content. Consequently, numerous studies have been conducted to counteract the rising trend of deepfakes. A predominant focus in previous research endeavors has been the classification of images or videos as either real or fake. While this approach is essential for detecting deepfakes, recent research has taken a step further by incorporating explainable AI techniques into deepfake detection models. By doing so, researchers aim not only to identify manipulated content but also to provide insights into why and how these models make their decisions. Among the various methods within explainable artificial intelligence, the network dissection algorithm emerges as a powerful tool for achieving transparency and reliability in deepfake detection models.

Network dissection, as an explainable AI method, is primarily concerned with unraveling the inner workings of convolutional neural network (CNN) models. It allows researchers to dissect the intricate layers of a CNN and comprehend the specific features and patterns that these models use for classification. This transparency is crucial in the context of deepfake detection because it enables experts to gain a deeper understanding of why a particular image or video has been classified as real or fake. However, it's essential to acknowledge that the application of the network dissection algorithm in the realm of deepfake detection is a relatively novel approach. This pioneering technique has been introduced and combined into the current study, marking a significant contribution to the field of deep learning and multimedia forensics. The network dissection algorithm's integration into deepfake detection models serves several essential purposes. Firstly, it enhances the interpretability and trustworthiness of these models by providing clear and concise explanations for their decisions. This transparency is invaluable, as it allows stakeholders, including law enforcement agencies, journalists, and the general public, to evaluate the credibility of deepfake detection results.

Secondly, the utilization of the network dissection algorithm allows for the identification of critical features and patterns within deepfake content. This knowledge can be used to develop more robust and effective countermeasures against deepfake generation techniques. By understanding the specific attributes that differentiate real and fake content, researchers can design better detection algorithms and preventive measures. Furthermore, this innovative approach also opens up avenues for ongoing research and development. As the network dissection algorithm gains traction in the field of deepfake detection, it is likely to spark further investigations and refinements. Researchers may explore ways to improve the algorithm's efficiency, adapt it to different CNN architectures, or integrate it with other explainable AI techniques for even greater insights.

In conclusion, the advancement of deep learning has led to the proliferation of manipulated multimedia content, necessitating robust solutions for deepfake detection. While much of the previous research has focused on classification, the integration of explainable AI techniques, such as the network dissection algorithm, represents a promising avenue for enhancing the reliability and transparency of deepfake detection models. This novel approach not only aids in the identification of manipulated content but also empowers researchers to better understand the underlying features and patterns that distinguish real from fake media. Consequently, the application of the network dissection algorithm in deepfake detection research holds significant potential in the ongoing fight against the negative impacts of deepfakes.

# LITERATURE REVIEW

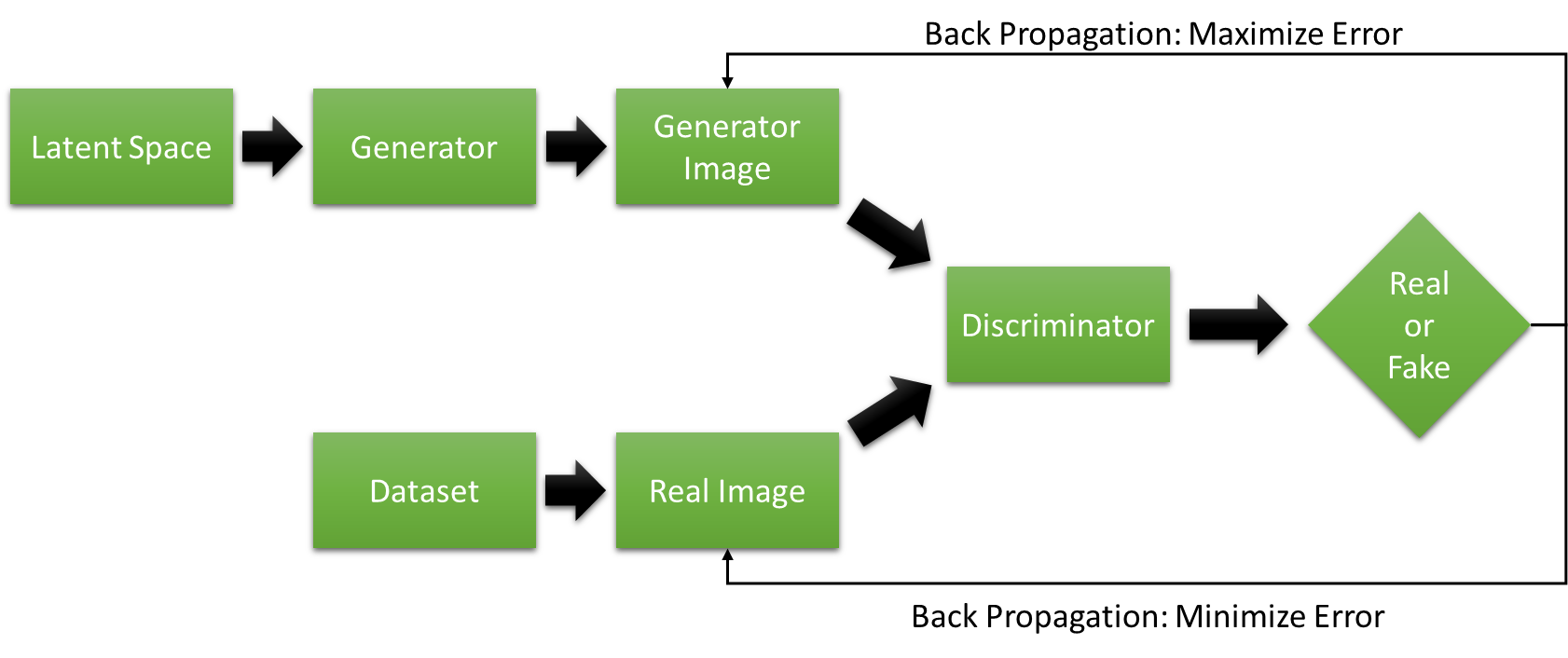
Literature review & subject research

With the advancement in the field of deep learning technology and the ease of generating fake content, media manipulation has become popular in recent years. These manipulated multimedia contents are created using the latest AI technologies and tools using deep learning algorithms. The rising number of deepfake content on various social media platforms demands a deepfake detection system which could be a solution to various threats in society. Over the past years, there are numerous studies which have been performed by various researchers in the area of deepfake detection.

It is significant to know the creation of these deepfakes to understand the different tools and solutions to detect them. The popular deep learning technique in the creation of deepfakes is using GANs. GANs are deep learning algorithms that include two neural networks that compete to create new synthetic data. GANs can be separated into three parts [22]:

* Generative – To explain how data is created visually
* Adversarial – Model training is carried out in an adversarial setting
* Networks – AI-based deep neural networks are used for training purposes

Figure 4 depicts a high-level flow diagram of GAN with the basic functions. The generator and discriminator play an antagonistic game during the training phase. The generator's goal is to create samples that deceive the discriminator, whereas the discriminator's goal is to enhance its capacity to differentiate between real and manipulated data. This competitive training forces both networks to develop over time.



**Figure 4: GAN Flow Diagram with Basic Function [21]**

## DeepFake Detection

Latest advancements in deep learning technology have emerged in creating authentic fake videos effortlessly which used to be done in Hollywood studios in older days. This phenomenon has been reckoned to the outbreak of the crisis in democracy, society, and constitutions and has proven to be a huge imprecation of technological development to human civilization [9]. Sohail [3] proposed a methodology to detect deepfake media using three different CNN models and a fusion of all three models to enhance generalization capability and robustness. The models were trained and tested on openly available DeepFake Detection Challenge (DFDC) data consisting of 400 videos and the fusion model attained 99% accuracy. Their proposed approach imposed different image augmentation techniques to make training data diverse which helps the model to learn different sets of features. Vurimi introduced a combination of ResNet CNN architecture and the LSTM (Long Short-Term Memory) approach [5] to detect deepfake videos. The model is trained in the Celeb-DF v2 dataset which consists of 590 actual YouTube videos and tested with an accuracy rate of 91%. Xueyu in their work [6] developed a novel framework known as ‘DeepFake disrupter’ to protect against deepfakes using Deepfake detection models mainly Xception, Resnet18, and Resnet50.

Features play a significant role in the classification task in the case of deep learning models. Mousa [7] implemented a CNN-based deep-learning model using mouth features to detect deepfake videos. They conducted the experiments on Celeb-DF videos and used MoviePy, an open-source application to edit and cut videos specifically to extract mouth features. Artem [8] put forward an approach to compare different EfficientNets models and their performance for deepfake detection. The authors did a comparison of models on DFDC training data and the study proves that no strong correlation is between the performance of the model and its size. The accuracy rate was higher for EfficientNet B4 and B5 models. Andreas implemented a model [10] to train the FaceForensics dataset which contains facial forgeries using a deep learning-based approach in a supervised fashion. Also, they created a large-scale dataset consisting of manipulated facial images which can be used for improving deepfake detection models. Ying Xu [11] introduced a different approach to generalize deepfake detection, unlike other deepfake detection techniques. They proposed a generalizable model for detecting unknown and novel Deepfakes using supervised contrastive learning. Additionally, the proposed approach aids in examining the learned features for explainability.

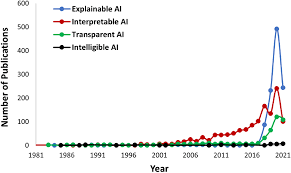
In the case of deep learning face-related tasks, the quality and quantity of the training data have a great impact on the outcome. The amount of time taken for collecting and labeling massive data quantity of data requires huge labor and remains expensive. Harshil [12] implemented a CNN approach for detecting Deepfakes on the CelebDF dataset by including various data augmentation techniques as part of pre-processing. Some of the augmentation techniques used by authors in [12] include GridMask, traditional augmentations, style transfer, and face morphing. The positive and negative effects of each augmentation technique were analyzed and compared for improving the performance. Ipshita [9] contributed to the detection and classification of deepfake images using pre-trained models such as InceptionV3 and VGG-16. Furthermore, as part of the research work fake images were generated using GAN, and localization and segmentation of fake facial images are done using a segment-based model.

Generally, we can apprehend that researches in the area of detecting fake images are still in scope as the generation of these manipulated images which look realistic is tremendously increasing in the current world. Although we have a walkthrough of the latest developments in this field of research, there are numerous study areas to address and explore.

## Explainable Artificial Intelligence

### Overview and Different XAI Methods

Explainable Artificial Intelligence (XAI) is an increasingly significant field within the broader domain of artificial intelligence (AI) and computer vision. It is characterized by the development of various techniques and approaches aimed at interpreting and elucidating the predictions generated by complex deep learning algorithms [49]. This burgeoning field addresses the need to make AI systems more transparent and understandable to human users, thereby promoting trust and effective management of these sophisticated "black-box" models. The relevance and growth of XAI can be attributed to the widespread adoption of deep learning solutions across diverse industries. Deep learning models have demonstrated remarkable capabilities in tasks such as image recognition, natural language processing, and autonomous decision-making. However, they often operate as opaque systems, making it challenging for users to grasp how and why specific predictions are made. This opacity can be particularly problematic in critical applications like healthcare, finance, and autonomous vehicles, where accountability and interpretability are paramount. XAI seeks to bridge this gap by developing a set of machine learning techniques that enable humans to comprehend, trust, and control AI systems effectively. These techniques serve as a means to shed light on the decision-making processes within deep learning models. By giving interpretable explanations for model predictions, XAI enables users to make educated decisions based on AI recommendations and enables developers to increase model robustness and fairness. The rising trend of XAI is evident in Figure 5, which illustrates the growing interest and adoption of XAI concepts and methodologies. This trend underscores the recognition of the importance of transparency and interpretability in AI systems. With the growth of deep learning solutions in all industries it has become relevant to understand the interpretation or internal representation of these ‘black-box’ models. Explainable artificial intelligence develops a set of machine-learning techniques that will enable human users to comprehend, appropriately trust, and manage the next generation of artificially intelligent partners [46] [47] [50]. Given the high complexity of deep learning models, it is difficult to understand and assess the internal decision process underlying predictions. In the case of images, explainability describes which parts of an image caused your model to predict specific classifications. Image explanations are essential for two different groups of people, mainly model stakeholders and model builders. For image models, each pixel is considered an individual feature and the explanation method assigns an attribution value to each pixel.



**Figure 5: Trend in XAI [52]**

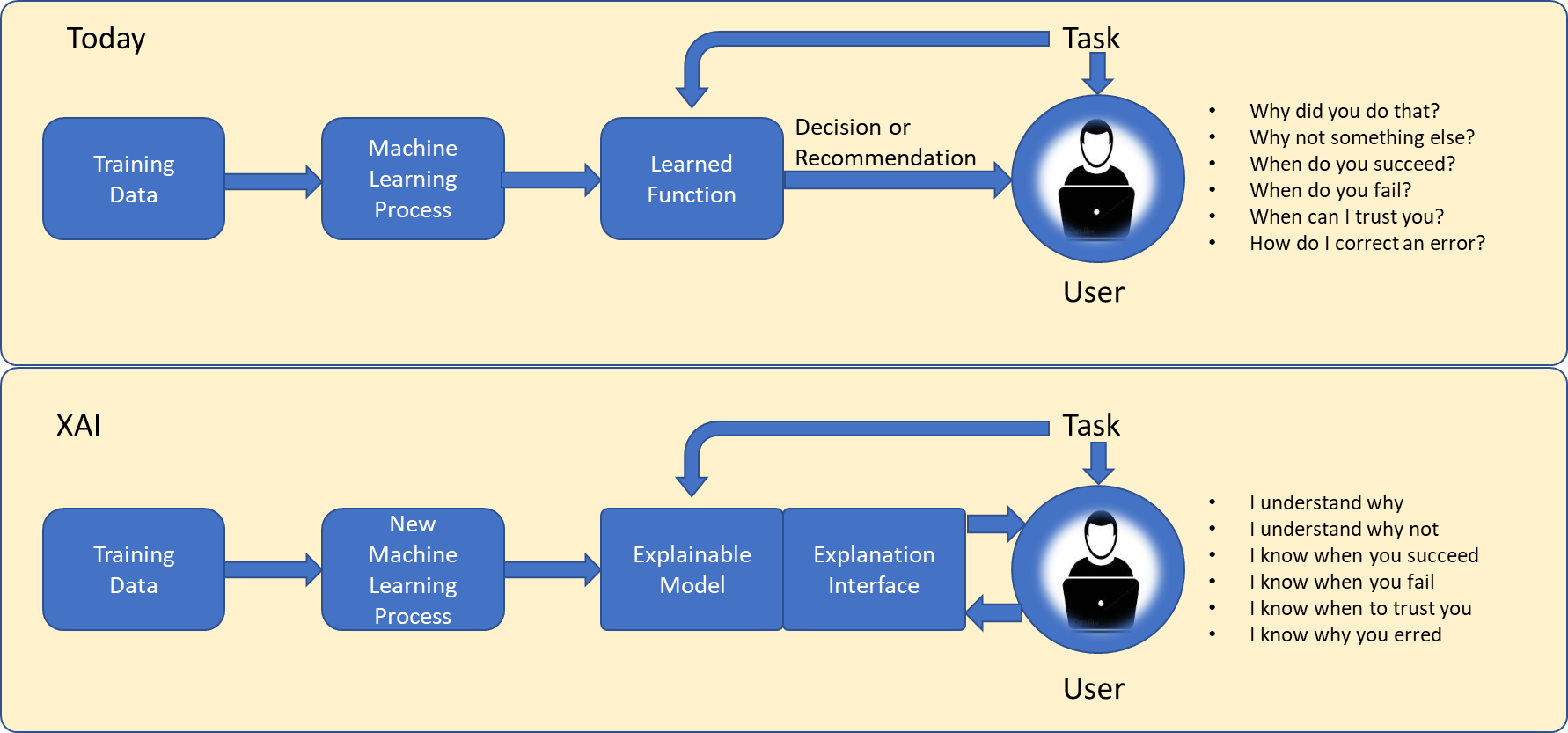
Figure 6 describes the scenario with and without explainable artificial intelligence. Unlike the traditional deep learning approaches which help in future predictions, explainable AI models give the reasoning behind the predictions. For explainability models, two elements draw attention along with the new machine-learning algorithm as shown in Figure 6 [46]. One of them is the explanatory model and the explanation interface is the other one.

An explainable model is a system with an adaptive rule-based reasoning approach that shows the cause-effect relationships between input data and outcomes attained by deep learning models [27]. An explainable model helps to explore the causes and form new strategies to overcome future situations in the case of image classifications.

There are three types of explainability:

* Pre-model explainability – understanding the training data
* Local post-model explainability – understanding one sample
* Global post-model explainability [41] – understanding the model’s overall behavior

Neural network predictions involve millions of mathematical calculations based on neural network architecture. It’s difficult to understand the predictions made by neural networks as we need to consider a lot of weights that are involved in building these models. This is why we need specific interpretation methods to interpret the predictions and behavior of neural networks. Firstly, neural network algorithms learn concepts and features from the hidden layers and special tools are required to uncover them [51]. Secondly, other model-agnostic methods like partial dependence plots or local models are mainly useful for tabular data, and text and image data require different techniques.



**Figure 6: Explainable Artificial intelligence (XAI) Approach [46]**

Table 1: XAI Methods

|  |  |
| --- | --- |
| Method name | Description |
| LIME & SHAP | -  Widely used especially for tabular data  -  Independent of the model  - Very slow with images and deep learning |
| Gradient-based methods | - Model specific  - Mainly designed for images  - Very fast but an understanding of model architecture is required |
| Saliency maps | -Oldest and most used explanation method for interpreting the predictions of CNN  - Simple, fast but not class-specific |
| Integrated gradient | -Optimal for images taken in non-natural environments  -Returns the individual pixels that notify a model’s prediction  -Time-consuming due to the number of samples to approximate the integral |
| Deconvnets | -Designed to work similarly to CNN but reverses the effect of convolution  -To reconstruct the activation on a specific layer, deconv layers are attached to corresponding CNN layers |
| Guided Backpropagation method | -Combination of saliency and deconvnet methods |
| Grad CAM | -Identify the parts of an image that most impact the classification score  -Improves class discrimination |
| Occlusion | -Replaces different contiguous rectangular patches of the input image with a given baseline and monitors the decrease of the predicted function |
| DeepLift | -Recursive backpropagation-based method |
| Smooth Grad | -This can be used on top of other attribution methods  -Generates multiple samples by adding Gaussian noise to the original input and averages the calculated attributions |
| XRAI | -Provides a heatmap of region-based attributions  -This method joins pixels into regions and shows the relative importance of different areas in an image and is more effective on natural images |

Table 1 shows some of the explainable methods that are currently used by researchers to provide the explainability behind neural network models [39] [47].

Avanti [25] used DeepLIFT (Deep Learning Important Features) method to decompose the prediction of a neural network using the backpropagation technique. The DeepLIFT technique has been applied to MNIST and the efficiency is better than other gradient-based methods. Andrei [26] proposed a region-based saliency method XRAI, that focuses on attributions. The XRAI method used in this approach produced better outcomes than other saliency techniques. Shawn implemented a new method namely Blur Integrated Gradients [28] to examine the attributions of deep neural networks. This technique was capable of scaling a neural network that identifies an object. Also, it avoids the requirement for a reference parameter ‘baseline’ like in Integrated Gradients.

Ioannis as part of the research compared ten different explainable artificial intelligence methods for understanding and interpreting predictions on remote sensing of various models using quantitative metrics to evaluate their performance [23]. Their study reveals that Grad-CAM, Lime, and Occlusion have proven to be the most reliable and interpretable XAI methods. Vidhi [35] put forward the Layer-wise Relevance Propagation method to understand the complex CNN classification model developed for classifying lung diseases from chest X-ray images. They did a performance comparison against other explainable methods like Guided Backpropagation (GB), Lime, and Deep Taylor Decomposition (DTD). Subrata [36] presented various methods like Activation Layer Visualization (ALV), LIME, SHAP, and GRAD-CAM to interpret the results from histopathology as it’s a challenging task to make trusty decisions. The performance of the neural network used for human fall detection is explained by the LIME technique by Jeyashree [37]. The outcome from the LIME method projects the features which are responsible for prediction by pointing the boundaries of input images and this helps in understanding the features responsible for the desired outcome. Kyle [38] proposed a strategy for explainability to visualize a CNN’s learned features using GAN for feature interpretation. Unlike other methods, this provides spatial localization which is relevant for classification. Kashif [44] used different XAI methods like Grad-CAM, Grad-CAM++Smooth, Vanilla saliency, and SmoothGrad to provide explainability on VGG-16 deep learning algorithm used for medical image classification. Christian proposed the XRAI approach [48] for providing a better understanding of decisions made by neural networks by keeping the target function explicit.

The above methods are the most widely used explainable artificial intelligence techniques. From the study, it can be summarized that some widely used saliency methods are independent of both the data and the model on which it was trained and hence it’s not a very good explanation of the model. Grad CAM, Guided backpropagation, Guided Grad-CAM, and XRAI are some of the methods which are currently used by researchers for explainability.

### XAI on Deepfake Detection

The significance of explainable artificial intelligence in today’s world and an overview of various XAI methods have been explained in the previous section. This section details the current research in the field of explainability which has been applied to different deepfake detection models. It explains the existing approaches or techniques used by researchers for deep learning algorithms developed for deepfake detection.

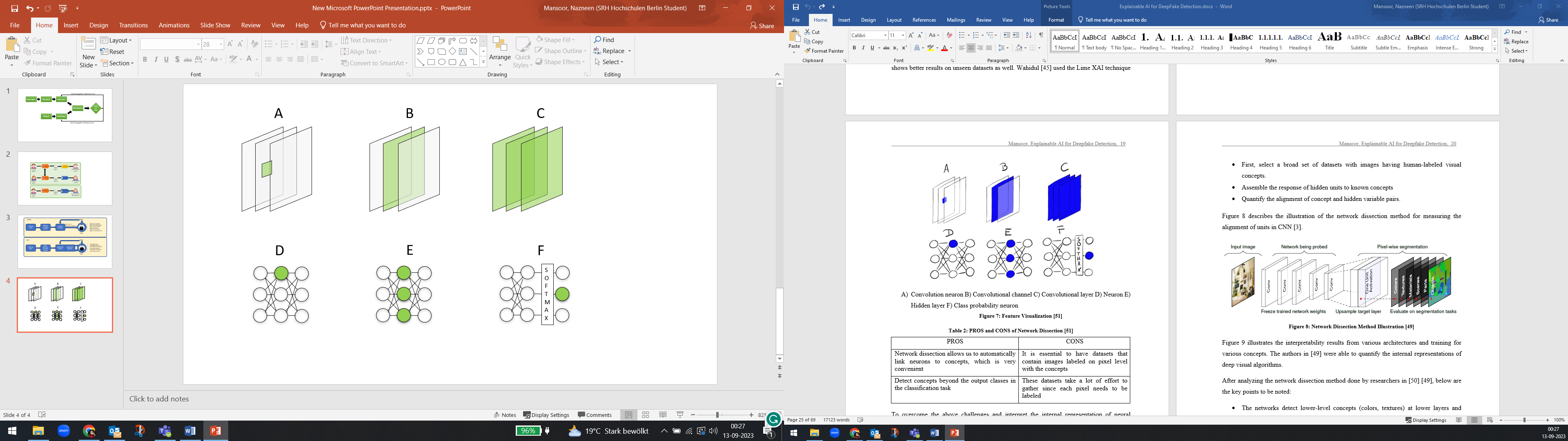
Ceyuan [24] introduced an approach for quantifying the layer-wise representations learned by GAN models which are used to generate scenes. The explainability module focuses on understanding the internal units of CNN models which act as object detectors to classify different types of GAN generated scenes. The quantitative and qualitative results from their research explain the features learned by training the model. Samuele [29] used XAI methods like GradCAM, and SHAP for explaining the predictions made by the EfficientNet model trained on the DFDC dataset consisting of manipulated videos. The results from the experiments show that human perceptions of deepfake explanation masks were not always in sync with objective metrics and other explainable artificial intelligence methods need to be considered for better results.

Federico [30] proposed a quantitative framework to estimate the informativeness and visual quality of explanations of Deepfake classifiers. The classification and explanation were done on the latest DFD and DFDC datasets. Shichao [32] proposed an FST-matching model (fake source target matching) for the interpretation of deepfake detection to learn features of images used to classify as real or fake ones. Sara [33] introduced an evaluation method for GANs by using the XAI method which is Logic Learning Machine and this has been used for improving data augmentation techniques. Loc [43] proposed an effective and interpretable method known as Dynamic Prototype Network for explaining deepfake temporal artifacts. Their experiments prove that the performance of the prototype network shows better results on unseen datasets as well. Wahidul [45] used the Lime XAI technique to interpret the CNN model used for identifying real and fake images and it also details the part of the image which caused to make the specific classification.

### Network Dissection

After analyzing and understanding the different explainable artificial intelligence methods for interpreting deep learning models, this research work will focus on the Network Dissection algorithm [49] [50]. Network Dissection is defined as an interpretability technique for convolutional neural networks (CNN) which provides relevant labels to their individual hidden units. This method has been used to quantify the interpretability of the Deepfake detection model to provide the explanation behind the classification of images as fake.

Neural networks learn concepts and high-level features in hidden layers from image pixels [51]. First convolutional layers learn features like simple textures and edges while later layers learn features that are complex textures and patterns. Furthermore, the last layers learn features like objects or parts of objects. Finally, the fully connected layers learn to map the activations from features to the individual predicted classes. Units can be referred to as individual neurons, feature maps, layers, or predicted class probability. These individual neurons are the atomic units of the network and most of the information can be captured by making feature visualizations of each neuron. Feature visualization is the technique of making these learned features explicit. For a unit in a neural network, feature visualization is done by selecting the input that maximizes the activation of such units [51]. Activation Maximization is a method to visualize neural networks and aims to maximize the activation of certain neurons. Feature visualization can be done for various different units [51] as depicted in Figure 7. But these visualizations do not prove that a neuron has learned a particular concept. Also, we do not have a measure to check how well a unit detects. Feature visualizations can convey what the neural network is doing but we cannot understand the neural network even if we look at thousands of feature visualizations [51].



1. Convolution neuron B) Convolutional channel C) Convolutional layer D) Neuron E) Hidden layer F) Class probability neuron

**Figure 7: Feature Visualization [51]**

Table 2: PROS and CONS of Network Dissection [51]

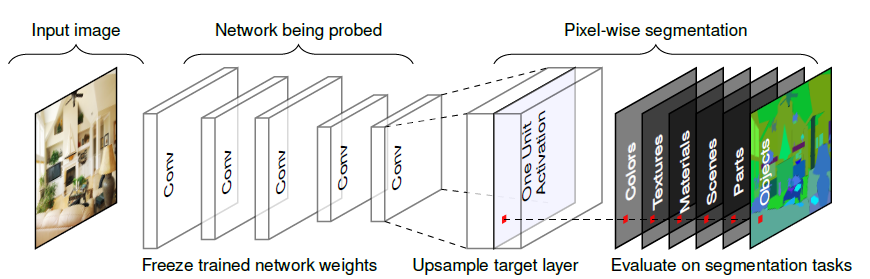
|  |  |
| --- | --- |
| PROS | CONS |
| Network dissection allows us to automatically link neurons to concepts, which is very convenient | It is essential to have datasets that contain images labeled on pixel level with the concepts |
| Detect concepts beyond the output classes in the classification task | These datasets take a lot of effort to gather since each pixel needs to be labeled |

To overcome the above challenges and interpret the internal representation of neural networks, the Network dissection method can be considered. Network Dissection is an interpretability method for CNNs that evaluate the alignment between individual hidden units and a set of visual semantic concepts [50]. Table 2 shows the advantages and disadvantages of network dissection.

In [49], the researchers proposed a general structure for network dissection for interpreting and quantifying the deep visual representations of convolutional neural networks by estimating the mapping between semantic concepts and individual hidden units. The authors have used Broden as the dataset which is referred to as a broadly and densely labeled dataset and tested the method on different CNNs like ResNet, AlexNet, VGG, and GoogLeNet which are trained on objects and scenic concepts. The proposed method is done in three different steps:

* First, select a broad set of datasets with images having human-labeled visual concepts.
* Assemble the response of hidden units to known concepts
* Quantify the alignment of concept and hidden variable pairs.

Figure 8 describes the illustration of the network dissection method for measuring the alignment of units in CNN [3].



**Figure 8: Network Dissection Method Illustration [49]**

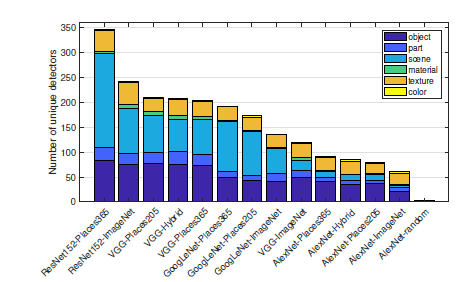
Figure 9 illustrates the interpretability results from various architectures and training for various concepts. The authors in [49] were able to quantify the internal representations of deep visual algorithms.

After analyzing the network dissection method done by researchers in [50] [49], below are the key points to be noted:

* The networks detect lower-level concepts (colors, textures) at lower layers and higher-level concepts (objects, parts) at higher layers.
* Batch normalization minimizes the number of unique concept detectors
* Increasing the number of channels within a layer increases the number of interpretable units
* Random initializations (training with different random seeds) result in slightly different numbers of interpretable units
* The number of unique concept detectors increases with the number of training iterations
* Networks trained on self-supervised tasks have fewer unique detectors compared to networks trained on supervised tasks

From the study, it’s understood that there are two important challenges that network dissection could not address:

* Concepts with a spatial overlap – there might be various facial concepts that occur concurrently in the same area of the facial image
* Global concepts – these concepts are not referring to particular locations of the face and are generally related to gender, age, skin tone, and beauty.



**Figure 9: Results of Network Dissection [3]**

In order to overcome the above challenges, an improved version of the Network Dissection algorithm called Hierarchical Network Dissection was done by researchers [52]. Hierarchical network dissection (HND) is used to interpret the internal representation of face-centric inference models [52]. It is used to dissect several face-centric models that are trained on commonly used facial datasets. From the analysis for HND in [52], it is clear that the results from HND depict models which are trained on different tasks learned varying internal representations. In [52], the researchers used HND to identify and quantify the bias in training data.

From the study, it’s understood that the quantitative and qualitative results as part of introducing XAI methods in different domains, could ensure better outcomes that help to trust decisions or predictions made by AI. Trust is an essential factor for AI decisions or predictions made by different models. Explainable AI is a significant step in evaluating the trustworthiness of these models.

# DESIGN AND METHODOLOGY

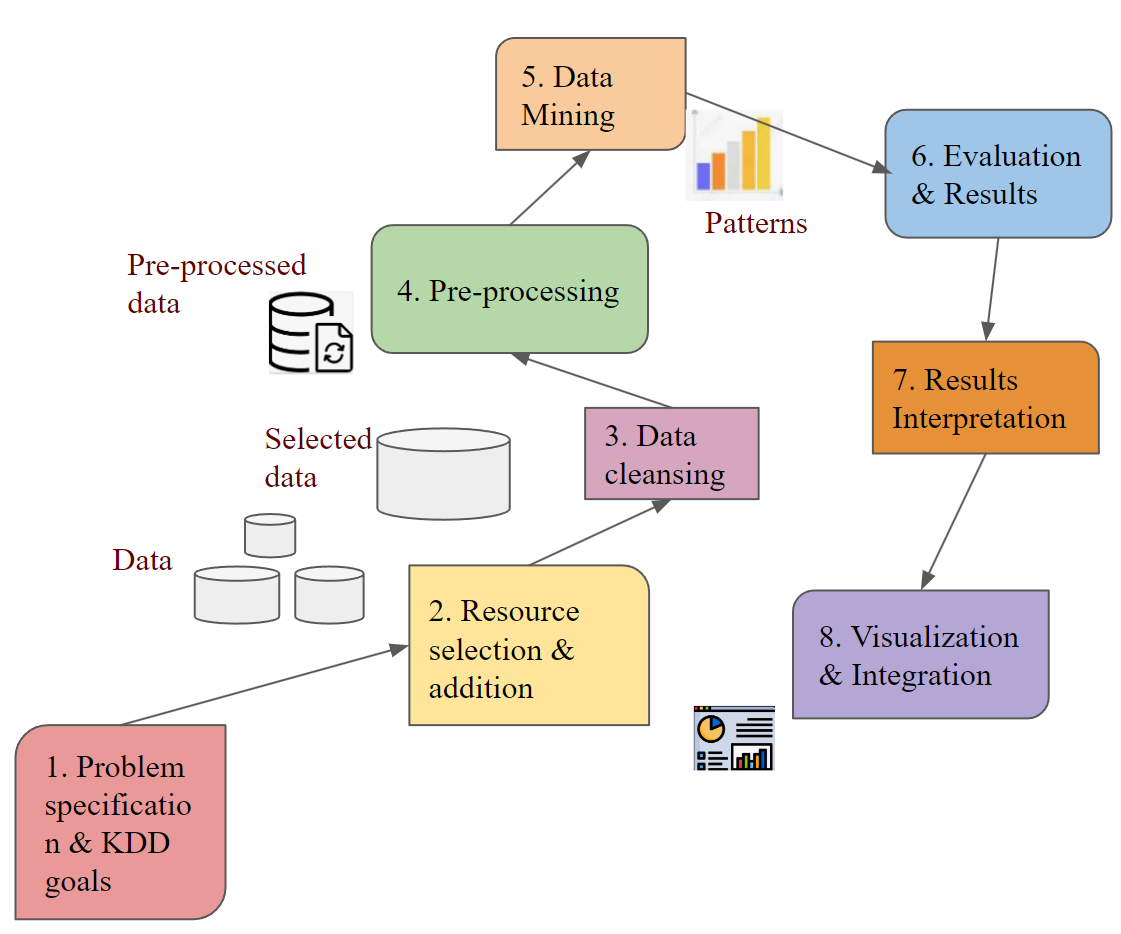
## Methodology

Big Data has been one of the significant concepts in the field of Data Science. Although there is no universal agreement on defining big data, it is widely used to describe large, diverse, and complicated data created from a variety of sources that require cloud applications for analysis and processing. In this aspect, videos and images provide data escalation at a higher scale due to the expansion of access to social media platforms. This has created a need for effective data processing in order to carry out the necessary implementation. A system capable of information abstraction and report generation automatically has become a decade-long requirement for assisting proper decision-making in AI or machine learning-based research efforts and organizational projects. Thus, the data mining concept has come up that can appropriately encapsulate data automatically, extract relevant information out of the stored data, and disclose and anticipate patterns out of them. Among the various commonly used data mining methodologies KDD process is selected for this research work.

KDD refers to Knowledge Discovery in Databases, the process of discovering data knowledge, and highlights the higher-level applications of certain Data mining techniques [54]. The primary goal of the KDD process is to retrieve information from enormous databases of data. KDD is defined as the systematic, exploratory investigation and modeling of large data sources [54]. KDD is the process of identifying valid, useful, and comprehensible patterns from vast and complex datasets. The foundation of the KDD technique is data mining, which includes the inference of algorithms that study the data, construct the model, and discover previously unknown patterns [54]. The availability and plethora of data in today’s world make data mining and KDD a matter of magnificent significance and need.

KDD is an interactive and iterative process consisting of nine steps. The procedure has numerous imaginative features in the sense that one cannot give a single formula or establish a comprehensive scientific classification for the correct judgments for each step and type of application. As a result, understanding the process and the various requirements and options at each stage is essential. The eight stages in a KDD process as in Figure 10 can be classified as stated below:

* Problem specification and KDD goals
* Resource selection and addition
* Data cleansing
* Pre-processing
* Data mining
* Evaluation and results
* Interpretation of results
* Visualization and Integration



**Figure 10: Stages of KDD Process**

1. Problem specification and KDD goals

This is the first preparatory stage and its purpose is to use the defined ideas which are related to the project proposal for developing an effective specification for the proposed work. This step includes stages like understanding the project tasks, identifying data availability, and the identification of hardware and software requirements. This stage helps in deciding the feasibility of the KDD goal. This phase includes a set of required resources such as cost, time, employees, software and hardware, and the high- and low-level tasks that will be completed during the project, the initial phase of a data dictionary, the feasibility measure determination, and document preparation for quality management [9].

DeepFake is one of the popularly discussed and apprehensive topics in recent times. Because of the negative impact on society, the economy, and politics, it demands immediate action. In this regard, the majority of computer vision research is focused on deepfake in order to establish advanced detection approaches in identifying hyper-realistic GAN-created images. The goal of the research is to develop a detection system. Understanding the significance of deepfake detection through a thorough analysis of the literature, an additional stage is appended to the proposed research work. The last stage involves explaining how to interpret the internal neuronal representation of deep detector models in classifying images as fake or real using the Network Dissection algorithm. The Celeb-DF dataset is selected for the research because most of the current publications in the domain have used it, making it easier to create a benchmark for our model. The proposed project includes training deep learning models which need high-end GPUs. The experiments are carried out in Jupyter Notebook and PyCharm with pre-installed libraries. All the experiments are done using the PyTorch application which is a machine learning framework used for computer vision problems.

1. Resource selection and addition

This phase is based on the data collected during the specification stage. The decisions taken in the previous step regarding gathering resources are used here. The data collection process in KDD mainly includes the gathering of data from various multiple data marts or data warehouses. Several data transformation processes are used to collect data in the appropriate format for the KDD project.

The Celeb-DF dataset is obtained from the paper [55] in which Yuezun Li developed a large-scale challenging dataset that contains 590 real videos gathered from YouTube with cases of different ethnic groups, genders, and ages, 5639 high-quality Deepfake videos of popular celebrities created using the synthetic approach. To test the explainability of the models, CelebAMask-HQ dataset is used [59]. The images were taken from the enormous CelebAMask-HQ collection, which consists of masked facial photographs [59]. A segmentation mask of facial concepts from the CelebA dataset is present in each image.

1. Data cleansing

The objective of this phase is to clean the collected data to make it appropriate for the KDD implementation. It entails numerous processes, such as outliers, dealing with missing values, errors, and untrustworthy data. This stage includes random sampling, handling outliers, missing values and inaccurate data processing, and database balance, including data deletion and duplication. Unlike numeric datasets, image datasets do not have the possibility of holding outlier or missing values that must be addressed in order to maintain data quality. Although, the presence of corrupt images is verified in the Celeb-DF dataset.

1. Pre-processing

This stage makes the data appropriate for implementation by extracting patterns from the data by doing multiple iterations. It makes it easier to retrieve relevant information from the data for enhancing the efficiency of the mining algorithms. The steps involved during this phase are feature selection, feature construction, and discretization. It is not required to complete all of the stages in this stage; which actions must be completed depends on what makes the data eligible for model implementation. When a color image is represented as numeric data, the color values vary from 0 to 255. Hence, data normalization is done to ensure the pixels and distribution of the input parameter are similar.

1. Data mining

Among the KDD stages, this is the most significant step where the knowledge and patterns are abstracted and obtained from the data. The data mining algorithms are diverse, and each has its own performance parameters like precision, recall, accuracy, etc. Multiple tasks can be executed on the same dataset using different algorithms. Parameters are classified into two types: those responsible for enforcing algorithms' search space, which impacts average performance, and those responsible for managing alternatives relevant to a given situation or problem. The mining is performed in the database's training section and then evaluated based on the performance of the algorithms and the parameters used. If the results are not up to the mark, model parameters are altered, and run the model again. If the results are still not satisfactory then the parameters are retrograded and the model runs again to examine the result. This is sometimes accomplished by compromising accuracy or by employing more sophisticated rules in the algorithms.

The research work has been done in two phases: detection of deepfake images followed by explaining the internal parts of the deepfake detection model. In the detection stage, the classification of fake and real images is executed using the ResNet-50, Inception V3, and VGG-16 CNN architectures. Here, the input layers of the CNN architectures are modified in terms of the classification task. Further, the internal parts of the detection models are explained using the Network dissection algorithm.

1. Evaluation and results

The evaluation of results is achieved using several ways by the models. The selected method is determined in the data-mining phase, and performance is determined in this step by choosing the evaluation parameters, and outcomes are compared based on that. Evaluation against the test dataset is one of the most extensive ways of evaluation. The training of the model is done using a training dataset and tested using a test dataset. The next level is reached when satisfactory findings are obtained with the testing dataset. There are several reasons for the depletion of the obtained result quality. Sometimes, it can be the result of underfitting or overfitting of the model. This can be rectified by modifying the parameters.

The performance evaluation metrics used for deepfake detection are accuracy, recall, F1-score, and precision of the models with CNN architectures (ResNet-50, Inception V3, and VGG-16).

Cross-Entropy loss is the loss function used for the classification of images. Classification accuracy also known as model accuracy is the ratio of the number of true predictions and the total number of classifications for the input images: Mathematically it is calculated as in Equation 1.

Equation 1: Model Accuracy Calculation.

Mathematical calculations for other performance metrics used in this work are given in Equations 2, 3, and 4.

Equation 2: Precision Calculation [62]

Equation 3: Recall Calculation [62]

Equation 4: F1-Score Calculation [63]

Model accuracy is used as one of the evaluation metrics because it is the most commonly used one for performance evaluation. But accuracy cannot be considered as it works better on balanced data. In practical scenarios, training, validation, and test datasets are unbalanced, and hence in this experiment, other performance metrics are also used as mentioned before.

1. Interpretation of results

Analysis and interpretation of results is a stage that requires considerable examination by domain specialists. The majority of work is done by KDD engineers using different software tools and their experience and knowledge of the project. The primary requirement for discovered knowledge is to persuade domain specialists who can describe the findings utilizing their vast expertise in the problem area. To comprehend anomalies, behaviors that differ dramatically from the experience of domain experts should be thoroughly examined. The most likely cause of these disparities is an error during the KDD process, which occurs frequently during the steps of data pre-processing and cleaning. Several large modifications occur in the databases throughout these stages, which are time intensive and tough to verify by eye. In certain circumstances, however, the disparities can indicate crucial information about the application areas that the domain specialists do not fully comprehend. The main result of the interpretation step is a group of patterns representing knowledge that has been confirmed and warranted by experts in the domain. Frequently, domain specialists can describe patterns that predict certain outcomes when given to them, but they would not normally be able to look at that specific pattern due to the huge variety from which to choose. The discovered knowledge can then be summarized and depicted in an appropriate form that can be provided to higher management with the goal of leveraging this information.

In the interpretation stage, the evaluation metrics are explained by the resulting values for corresponding models. The loss and accuracy plots are used to evaluate the efficiency of the deepfake detection models used in the proposed work. In the early phase of training, the expectation for loss values is relatively high as the model is learning on its own. However, with the rise in the number of training epochs, the loss functions show a substantial decreasing trend, indicating that the models are improving and successfully classifying real and fake images.

1. Visualization and Integration

The last stage of the KDD process involves utilizing the knowledge gained in previous phases. Until now, the results of the data mining stage have been thoroughly evaluated and validated by domain experts. As a result, the knowledge is now thought to be of high quality, true, and applicable to the field of application. Exploitation combines results and outcomes into the work environment that, if used properly, will benefit the organization. Comparing the accuracy of the pre-trained CNN models to detect fake images with the suggested explainable model to explain the internal parts of the model, it may be expanded for performing explainability in other scenarios thereby improving the reliability of predictions made by AI. There are different ways of visualizing and interpreting the results. In this experiment, there are different plots that describe and compare the efficiency of the models. These plots are done using various Python libraries.

## Design Specification

The proposed work has been categorized into two stages – detection and explainability. In the detection stage, the classification task is done with two CNN models VGG-16, ResNet-50, and Inception V3 from where a comparative analysis was made in regards to the model accuracy to evaluate how well the model can differentiate between real and fake images. The models are trained using the Celeb-DF dataset which contains deepfake images. The explainability model is tested by using a DeepFake Face dictionary formed from the CelebAMask-HQ dataset. In the following Network dissection algorithm, the proposed explainability model is used to understand the individual units of the pre-trained CNNs in classifying as real or fake images. The overall process has been executed using the Jupyter Notebook platform and PyCharm community edition on which deep learning algorithms can be trained at increased speed. The implementation stages of DeepFake detection and explainable artificial intelligence are portrayed in Figures 11 and 12.

### DeepFake Detection

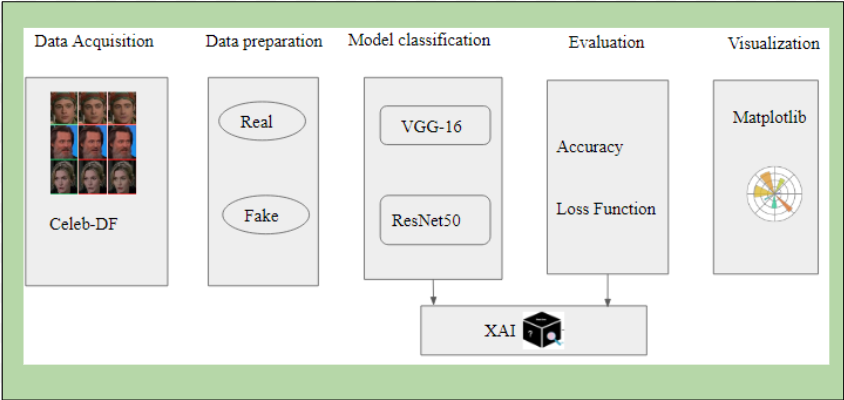
In recent years, the field of computer vision has witnessed remarkable advancements, particularly in the domain of image manipulation and detection. With the proliferation of manipulated or fake images and videos, detecting such alterations has become a critical challenge. Convolutional Neural Networks (CNNs) have emerged as powerful tools for addressing this challenge, demonstrating superior performance in distinguishing between real and manipulated images. In this elaboration, we will delve into the key aspects of this research, including the architecture of DeepFake detection models using CNNs and the utilization of three prominent CNN architectures: Inception V3, ResNet-50, and VGG-16. Moreover, we will explore the adaptations made to these architectures to suit the Celeb-DF dataset, which differs from the ImageFolder dataset, and the implications of these developments in the broader context of computer vision.

The advent of DeepFake technology has raised serious concerns regarding the authenticity of visual content in various applications, such as news reporting, entertainment, and cybersecurity. In response to this challenge, researchers have turned to CNNs as a robust solution for detecting manipulated images and videos. Figure 13 provides a visualization of the CNN architecture designed specifically for DeepFake detection models. One of the significant accomplishments in this domain is the detection of manipulated images using the Celeb-DF dataset, a benchmark dataset commonly employed for evaluating DeepFake detection algorithms. To tackle this task effectively, three well-established CNN architectures—Inception V3, ResNet-50, and VGG-16—have been harnessed.

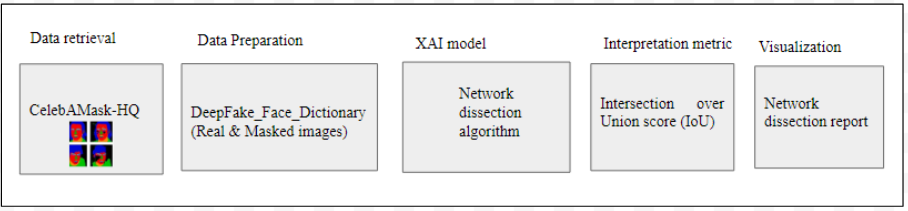
A critical consideration in using the Celeb-DF dataset lies in the variance of image dimensions compared to datasets like ImageFolder. To address this discrepancy, certain modifications have been introduced to the conventional CNN architectures. These adaptations ensure that the models are capable of handling the unique characteristics of the Celeb-DF dataset. Additionally, the output layer of the network has undergone alterations. Unlike the ImageFolder data, where multiple classes are present, this research focuses on binary classification—distinguishing between real and fake images. As a result, the updated architecture is optimized for processing colored images with a resolution of 224 x 224 pixels. To ensure uniformity and optimal data preprocessing, the image data are normalized to a scale ranging from 0 to 1 through a rescaling layer. Following this preprocessing step, the rescaled images are passed through the base models: ResNet-50, Inception V3, and VGG-16. The output from these base models is then subjected to an adaptive average pooling layer, followed by a dropout layer with a dropout rate of 50%. Finally, a sigmoid function is applied in the output layer, which is tailored for binary classification tasks, serving as the model's decision-making mechanism to classify images as either real or fake.

* VGG-16

VGG-16, standing for Visual Geometry Group 16, represents a significant milestone in the field of computer vision. This CNN architecture is constructed with 16 weight layers and gained widespread recognition when it participated in the ILSVRC ImageNet Challenge, securing the best model award in 2014. Even today, VGG-16 is considered one of the premier vision model architectures. A distinctive feature of VGG-16 is its simplicity, which contrasts with the trend of introducing numerous hyperparameters in contemporary models. Instead of complicating the architecture, VGG-16 primarily employs 3x3 filter convolution layers, often with a stride of 1, along with the use of stride 2 padding and 2x2 max-pooling layers. This consistent arrangement of convolution and max-pooling layers is one of the keys to its success. VGG-16's architectural consistency extends to the layout of its layers throughout the model. Notably, it encompasses two fully connected layers, followed by a softmax layer in the output. However, it's crucial to acknowledge that VGG-16's simplicity comes at a computational cost, as it is a large network with approximately 138 million parameters.



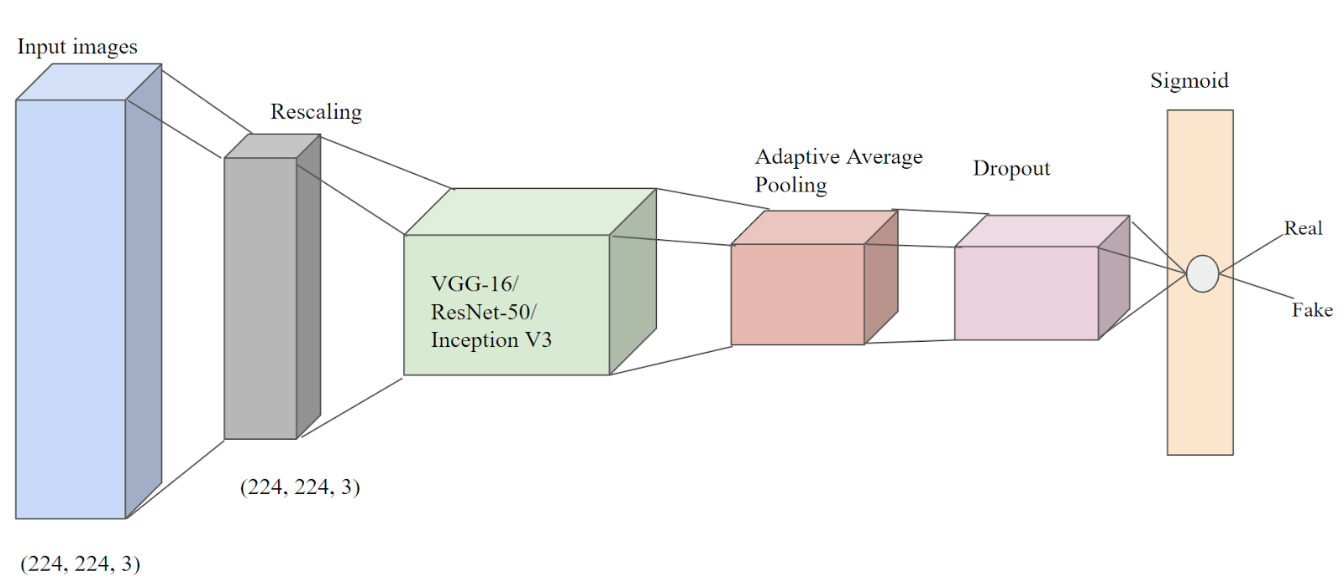
**Figure 11: Implementation Stages for DF Detection**



**Figure 12: Implementation Stages for XAI**

* ResNet-50

ResNet-50 represents an evolution of the ResNet architecture and is a CNN with 48 convolution layers, one average pooling layer, and one max pooling layer, totaling 50 layers in depth. After the introduction of AlexNet, ResNet has been one of the most influential architectures in the realm of deep learning. ResNet's innovation lies in its establishment of residual connections between different layers, an approach that has led to reduced loss, enhanced knowledge acquisition, and improved training performance. When a layer features a residual connection, it indicates that the layer's output is a convolutional combination of both its input and the input from an earlier layer. This architectural breakthrough has had a profound impact, enabling the creation of exceptionally deep networks while addressing challenges like vanishing gradients. ResNet-50, in particular, has gained popularity for its ability to capture complex image features effectively, making it an ideal candidate for DeepFake detection.



**Figure 13:** **CNN Architecture for DeepFake detection Models**

* Inception V3

Inception V3, developed by Google, is the third iteration of the Inception architecture. Initially trained on the ImageNet dataset, which comprises over one million training images and spans 1,000 classes, Inception V3 has demonstrated significant prowess in image analysis and object detection. In the experiment under consideration, the PyTorch version of Inception V3 is utilized. The architecture boasts a range of features designed to enhance its performance in image analysis and object detection tasks. These features include Factorized 7x7 convolutions, label smoothing, and auxiliary classifiers. Factorized 7x7 convolutions allow the network to efficiently capture spatial information in images. Label smoothing contributes to improved model generalization by preventing overconfidence in predictions. Auxiliary classifiers provide intermediate supervision during training, facilitating better feature extraction and learning.

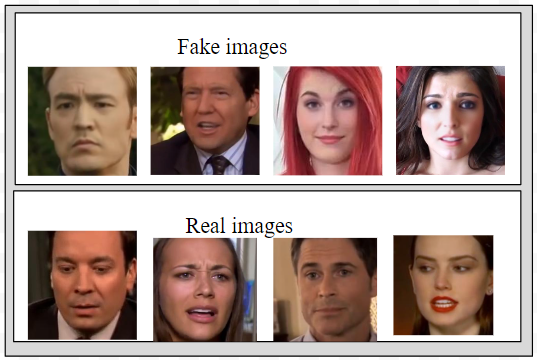
#### Dataset Description

The training dataset used to train the two Convolutional Neural Network (CNN) architectures plays a crucial role in the development of deep learning models for image classification tasks. In this case, the dataset consists of a mix of fake and real images, sourced from the Celeb-DF dataset [58]. This dataset is instrumental in advancing the field of deep learning and computer vision, particularly in addressing the challenges posed by DeepFake videos, which have gained notoriety for their potential to spread disinformation and manipulate digital content.

The Celeb-DF dataset is a valuable resource for researchers in the field of deep learning and computer vision, containing a substantial collection of both DeepFake and real videos. Specifically, the dataset comprises 5,639 DeepFake videos and 590 real videos, providing a comprehensive representation of both synthetic and authentic content. Each video in the collection lasts about 13 seconds on average and has a standard frame rate of 30 frames per second. The authenticity of the real videos in the Celeb-DF dataset is notable, as they are curated from publicly accessible YouTube videos. These videos encompass 59 celebrity interviews, which include individuals from diverse ethnic backgrounds, age groups, and genders. This diversity is essential for ensuring that the CNN models are capable of generalizing their learning across various demographics and characteristics.

In the context of the DeepFake videos within the Celeb-DF dataset, they are the result of a face-swapping technique applied to 59 distinct subjects. This technique involves digitally manipulating the facial features of one individual to make them appear as another person, which is a common method for creating DeepFake content. These synthetic videos pose a significant challenge in the detection of manipulated content, making them an ideal test case for evaluating the effectiveness of the CNN architectures. The overall dataset contains a substantial number of images, with a total of 19,457 individual frames extracted from the videos. To ensure proper training, validation, and testing of the CNN models, a balanced split of the dataset is employed. Specifically, 80% of the data, amounting to 15,565 images, is allocated for training the models. This large training set enables the models to learn the underlying patterns and features necessary for distinguishing between real and DeepFake content. Furthermore, 10% of the dataset is set aside for validation purposes. This validation set is instrumental during the training process as it helps monitor the model's performance and adjust hyperparameters to prevent overfitting, ensuring that the model generalizes well to new, unseen data. Finally, the remaining 10% of the dataset is reserved for testing the models. This test set serves as the ultimate evaluation, assessing the CNN architectures' ability to accurately classify real and fake images.

To provide a visual representation of the data, Figure 14 is included, depicting a selection of real and fake images from the Celeb-DF dataset that were used for training. These images showcase the diversity of the dataset, with various celebrity subjects and different levels of image quality, posing a range of challenges for the CNN models. In addition to the dataset itself, data augmentation techniques are employed to enhance the model's ability to learn and generalize effectively. These techniques introduce variations to the training data, effectively increasing the dataset's size without collecting additional samples. Some of the data augmentation techniques used for this work include random rotation, horizontal flip, center cropping, and resizing. These data augmentation techniques are essential for improving the model's ability to handle variations and uncertainties in real-world data. By exposing the model to a wider range of image transformations during training, it becomes more robust and capable of making accurate predictions on unseen, real-world data.



**Figure 14:** **Real and Fake Images from Training Dataset [58]**

#### Implementation

For evaluating CNN-based classifiers (ResNet50, Inception V3, and VGG-16) detection efficiency, a test dataset is formed which includes images from the Celeb-DF dataset. The performance metric used to compare the deepfake detection models are accuracy, recall, F1-score, and precision. Cross-Entropy loss is selected as the loss function and Adam optimizer of learning rate 0.00001. Also, for ResNet-50 and Inception V3, the batch size is opted as 32 for training, validation, and test datasets. The batch size is set to 16 with a learning rate of 0.001 for VGG-16. After conducting multiple experiments with different sets of epochs, the epoch value is set to 15. Hyperparameter tuning has been done to choose the parameters after doing multiple experiments with various loss functions, optimizers, and batch size combinations. The experimented batch size values are 16, 32, and 64 with Adam, and Stochastic gradient descent optimizers and different sets of learning rates ranging from 0.01 to 0.00001.

### XAI - Dissecting DeepFake Detection Models

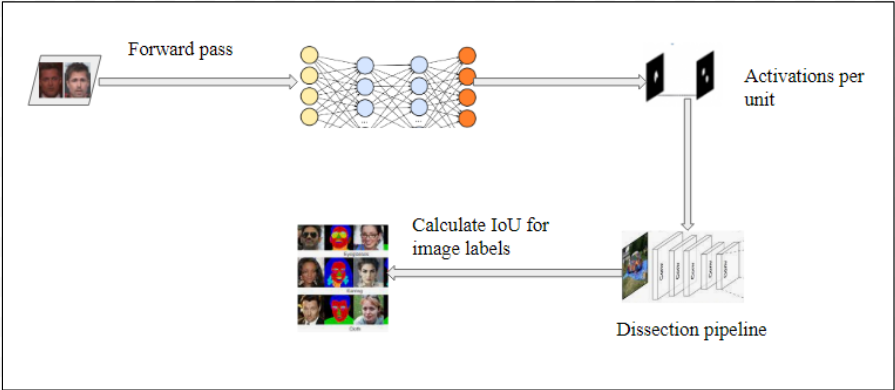
#### Network Dissection algorithm

The unit-concept pairing of Network dissection is based on our DeepFake Face Dictionary which consists of various facial concepts. In the proposed work, the Network dissection algorithm uses the DeepFake Face Dictionary which includes concepts of different facial features.

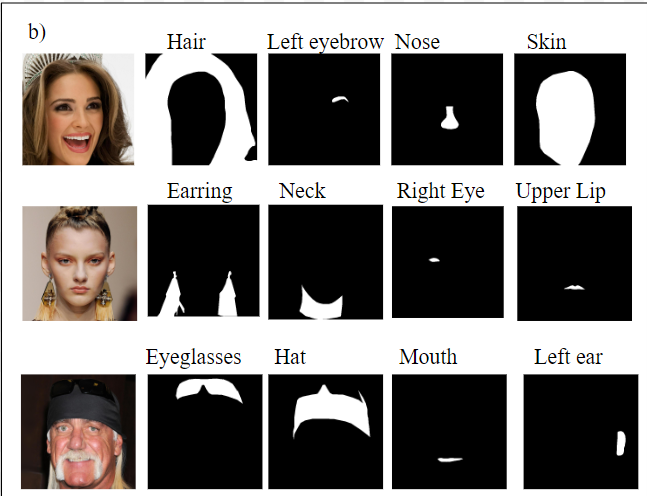
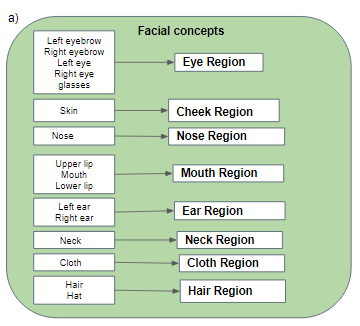
The proposed method, the explainability model based on the network dissection algorithm for the trained CNN models, detects units that could detect corresponding facial concepts using unit-concept pairing. Figure 15 describes an overview of the probabilistic approach to determine what units of the CNN models act as facial concept detectors which helps in the classification of images as real or fake ones. In the proposed work, a forward pass is taken across all the images in the DeepFake Face Dictionary to store the activation maps of the dissected layer and run the network dissection algorithm to produce IoU scores for each concept per unit.

#### Dataset Description

Our DeepFake Face Dictionary consists of 17 facial concepts and a group of corresponding images for each facial concept. The images are obtained from the CelebAMask-HQ dataset which is a large-scale dataset consisting of masked face images [59]. Each image contains a segmentation mask of facial features corresponding to the CelebA dataset. The masks of the images were manually annotated and included facial concepts with 19 classes and among them, the labels which have been used in the proposed work are nose, skin, ears, eyes, lip, hair, mouth, hat, neck, earring, eyeglasses, and cloth. Figure 16 depicts some examples of facial concepts and their respective masks from the DeepFake Face Dictionary. Overall, there are 100 individual images and 1238 masked images corresponding to various facial concepts. The graph in Figure 17 shows the histogram of the number of labeled occurrences for each facial concept present in the DeepFake Face Dictionary.



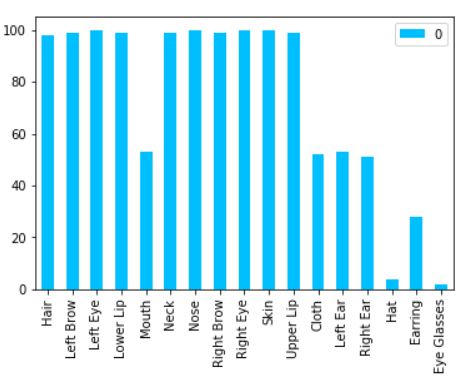
**Figure 15:** **XAI Approach for Dissecting CNN Models**



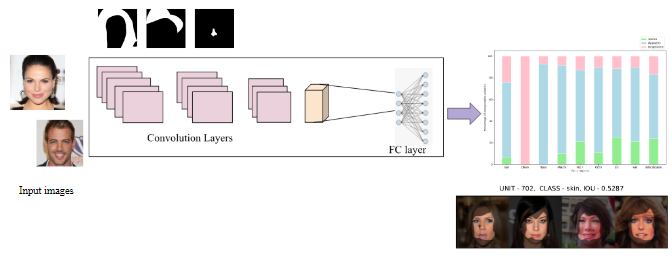
**Figure 16:** **DeepFake Face Dictionary a) Facial Concepts and corresponding regions from dictionary b) Samples of images and respective segmentation masks for facial concepts**

#### Implementation

Network dissection algorithm is used to dissect the CNN models trained to classify the images as real or fake. To test the explainability potential of the models, the network dissection algorithm is performed to dissect different layers of the CNN models. This aimed to assess the feasibility of implementation of the deepfake detection models. Figure 18 represents an overview of Network dissection for deepfake detection CNN models.



**Figure 17:** **Number of Images per Facial Concept in our DeepFake Face Dictionary**



**Figure 18:** **Overview of Network Dissection for DeepFake Detection Models**

The workflow can be explained in different steps as follows:

⦁ Extract the activation maps of all the input images (say N) present in the input dataset which is ‘DeepFake Face dictionary’. The face dictionary contained labeled facial concepts for the given layer in the selected model loader script depending on different CNN architectures. For all the neurons or units in the specific layer, N number of maps are generated using the main script which initially stored the activations in a NumPy memory map. For deeper layers, the size of the memory map is small due to the lower resolution levels of the activation maps.

⦁ Evaluate the threshold values for each unit using spatial features and evaluate a value in which the probability of any spatial region with a value greater than the threshold value is set as a quantile threshold. The experiments are carried out with different quantile threshold values for segmenting the activation maps for each unit. These activation maps represent the salient areas of the images which help in predicting them as real or fake. In the proposed work, the selected threshold values are 0.4, 0.04, and 0.005. Based on the threshold values there are variations in the amount of interpretable units to detect facial concepts.

⦁ Segment the activation maps using threshold values per unit and estimate their intersections over unions (IoU) scores. For each pair of unit-concept, a list of IoU values is generated. The concept with the maximum IoU score is noted and the neuron is considered interpretable if the concept IoU exceeds a certain threshold value.

Table 3 summarizes the details of the models that were dissected, given the layer that is dissected during the experiments, and depicts the results for each model. The focus of these interpretability experiments was on the deeper layers of the models because based on previous studies high-level concepts are mainly found in the layers that are closer to the output and the face dictionary consists of higher-level concepts. The performance metric used to test the explainability of these CNN models is Intersection over Union score (IoU) which is given in Equation 5.

Equation 5: IoU Score Calculation.

Table 3: Details of Dissected DeepFake Detection CNN models

|  |  |  |
| --- | --- | --- |
| CNN Architecture | Dissected Layer | Performance (metric) |
| VGG-16 | Block 5 - Layer13(conv3) | 86.2% (Accuracy) |
| ResNet-50 | Block 5 - Layer4(conv3) | 82.8% (Accuracy) |
| Inception V3 | Block 5 - branch6(conv3) | 0.824 (F1-score) |

For each unit, the concepts with the highest IoU are noted in the output Net dissect report, and the unit is considered interpretable for those top concepts having values greater than 0.04. Although we can get the dominant concepts for each unit based on IoU scores, sometimes there is more than one concept that has a high IoU and it lies in a similar area of the face. In these situations, it’s better to find a hierarchy of concepts that lie in the same facial region as that of the concept with high IoU. This is done by identifying the facial region and generating probabilities for each concept that lies within that facial region.

# RESULTS AND EVALUATION

The results of the developed deepfake detection models are shown in the first part of this section and the second part describes the results from the interpretations of these models. Also, in this chapter, the comparison of the performance and interpretations of proposed CNN models are also discussed.

## DeepFake Detection

In the proposed work, deepfake detection is performed using three commonly used CNN architectures: Resnet50, Inception V3, and VGG-16. The performance metrics used to compare all three models include accuracy, precision, recall, and F1-score. The results from the test data are represented in Table 4.

Table 4: Comparison of Performance of CNN models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy (%) | Precision | Recall | F1-score | Trainable Params |
| ResNet-50 | 82.8 | 0.853 | 0.856 | 0.854 | 5.3M |
| VGG-16 | 86.2 | 0.978 | 0.785 | 0.871 | 10.3M |
| InceptionV3 | 76.1 | 0.729 | 0.947 | 0.824 | 5.3M |

Figure 19 shows the learning curves obtained by training the CNN models with different hyperparameters and includes both accuracy and loss plots with an increasing number of epochs. The learning curve plots depict the results obtained for 15 epochs. The optimizer used for all three models is the same (Adam optimizer) with a learning rate of 0.00001 and batch size of 32 for ResNet-50 and Inception V3. On the other hand, for VGG-16, the selected batch size and learning rate are 16 and 0.0001 respectively. All the hyperparameters are chosen after conducting multiple experiments on the models.

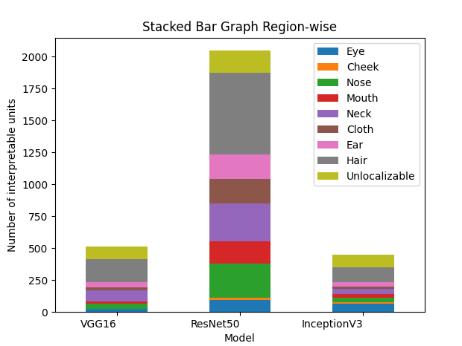
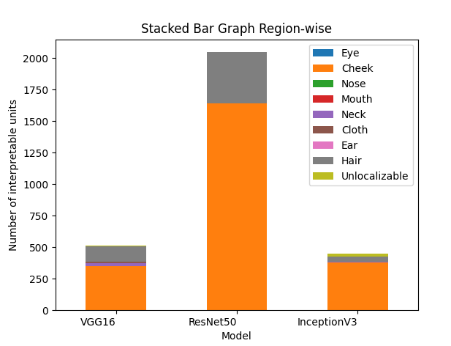
## XAI - Network Dissection Results

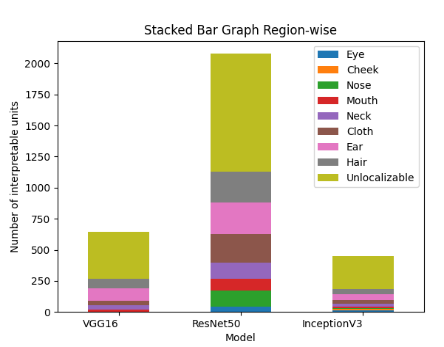
This section of the proposed work aims at explaining the underlying decision process of the DeepFake detection models. After doing multiple experiments with DeepFake Face Dictionary obtained from the CelebAMask-HQ dataset, a network dissection report is generated for each model. The report provides information about the interpretable facial concepts by each model with different IoU scores.

|  |  |
| --- | --- |
| **ResNet-50** | |
| https://lh4.googleusercontent.com/WIqIjOdRmab6ebf-rIX1Nhd5PdJ-aku8spDb0t3sCAejuGJJG5boeUK0bgPZsP8_30cPP2IlEjrs_p2Nm28B8gNE91zDEi58SnDe2D_9PoiY3nvETwLMv7QweDvE-F5wFT76aqG0b_Jj | https://lh6.googleusercontent.com/GgjZqS8yOnPqPQbSCuJgCeQwlk_14jZgbtGGenwVYfgCTsFHiDEErzbiQQqFUu39LlfM69zQ12B__85TP9GcRr4dpp1JU45NUScX2uyLZJGrHLr4Fkyq4St6n0DzEo70PpHii8T6IvUw |
| **VGG-16** | |
| https://lh4.googleusercontent.com/SXWfq4T6ZfC-AY44vLgSsBaJ_ZIPnVszFNmvMqG6PnmPojjw75_Uw0AVMw9mkQwBTQ07JXFrgoWG_sulyBa0e-l-PfE7scOvhSIaySPcWwq-qtLZeRq73D06XPqqNBFCEp8ZSZ2b3TaI | https://lh3.googleusercontent.com/KVuyknS_BV46fWp_KQXu-luDWTUQwqaFGM4MmGrp8ayRBbK-9KxlPBVsmOgdYdH-OGd8VUUSbOlBLxDm7KXi_NTXZFxtzHKAF8MDxgKi35tpj9VIpRHtJ7ROYDGlmEck-QCfQPwulVyA |
| **Inception V3** | |
| https://lh6.googleusercontent.com/_NWqfcjL-b8CxuNlT7G-aDniqEizyxOCdid11oSCPsIeVkkf-e4RLEhajcSD5H-hxUB_slBJs3oxNnRzmjNwylcFARpFE93Jpw3oF4cVG3liHBoBfpTr3Z2-tUpqzC0ZWCaUnsL_xTwX | https://lh6.googleusercontent.com/qQjZWhoZvzGdZvLbxr9zih9Yy9LFA3ZyKgl5jAXB6gldZLqpaDmN3rZQwMnjs5cHHXsjlcXaa2PbQHNfdRdelVMoMT_QzCMenM7-3aWYeUeCDlfi14FaNrhSl-lmdOTpmJMVHfTZJ0_C |

**Figure 19:** **Learning Curves of DeepFake Detection Models**

To test the explainability of the models, different threshold values are set to compare the features learned by different architectures. This gives an overview of how models learn different facial concepts based on segmentation. Figure 20 depicts the number of interpretable units detected by each model with regard to the various facial regions for different threshold values. Figure 21 represents a sample of a few of the facial concepts with their IoU scores detected by the models. In Table 5, the information about the amount of interpretable and uninterpretable units along with the top IoU scores and their corresponding facial concepts are listed.

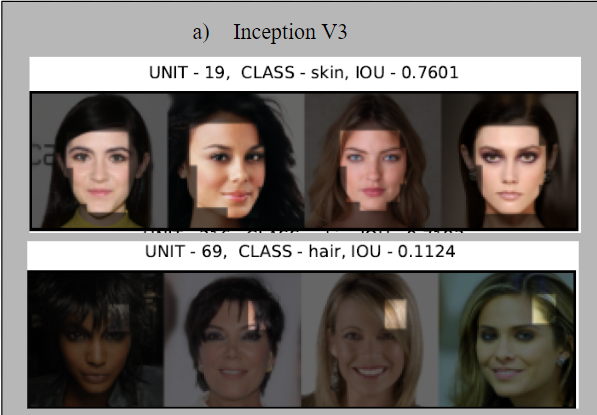


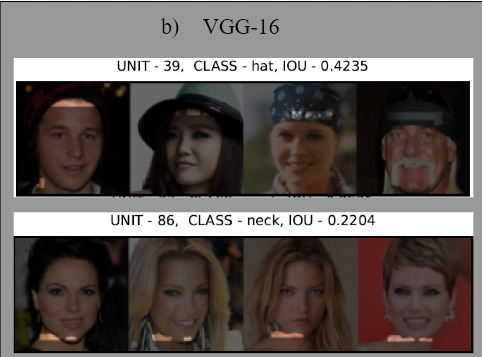


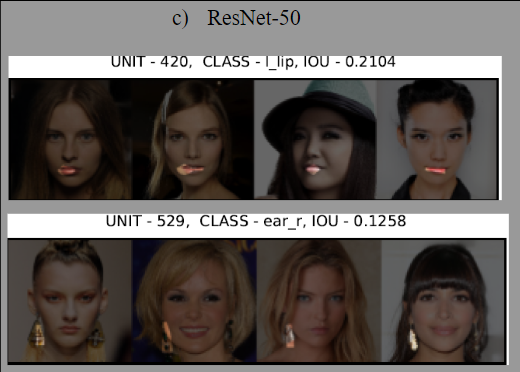
**Figure 20:** **Distribution of Interpretable Units per Facial Region a) Threshold value - 0.4 b) Threshold value - 0.04 c) Threshold value - 0.005**

Table 5: Details of Interpretations by CNN Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Models** | **Threshold values** | **Interpretable units** | **Uninterpretable units** | **Top IoU score** | **Facial concept** |
| **VGG-16** | 0.4 | 509/512 | 3/512 | 0.5918 | skin |
| 0.04 | 417/512 | 95/512 | 0.4235 | hat |
| 0.005 | 133/512 | 379/512 | 0.3187 | cloth |
| **ResNet-50** | 0.4 | 2048/2048 | 0 | 0.5287 | skin |
| 0.04 | 1875/2048 | 173/2048 | 0.3524 | hat |
| 0.005 | 1128/2048 | 950/2048 | 0.3510 | earring |
| **Inception V3** | 0.4 | 423/448 | 25/448 | 0.7601 | skin |
| 0.04 | 350/448 | 98/448 | 0.4286 | cloth |
| 0.005 | 182/448 | 266/448 | 0.4229 | earring |

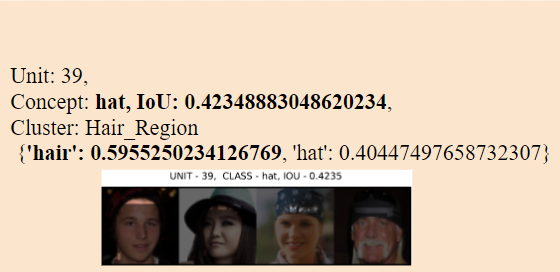


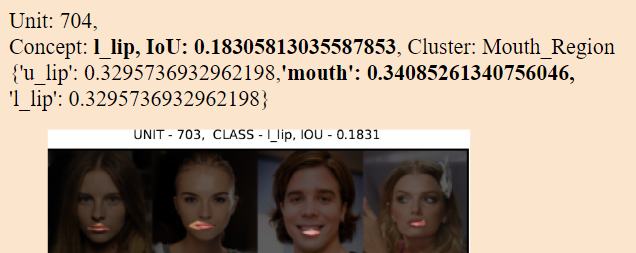




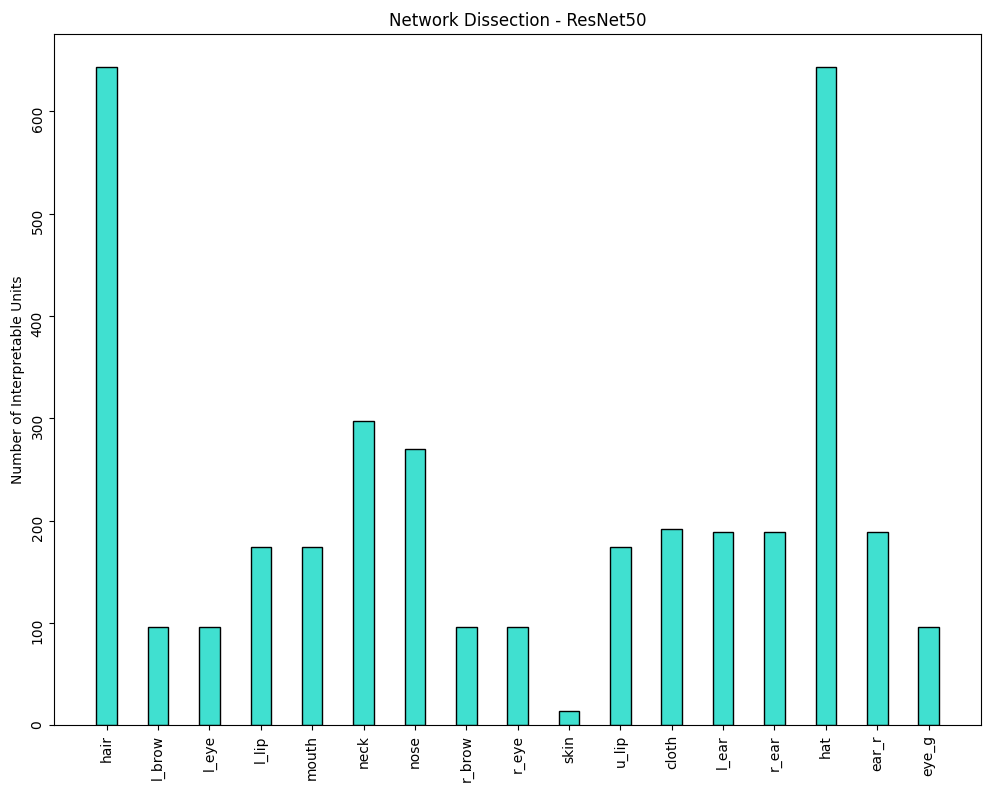
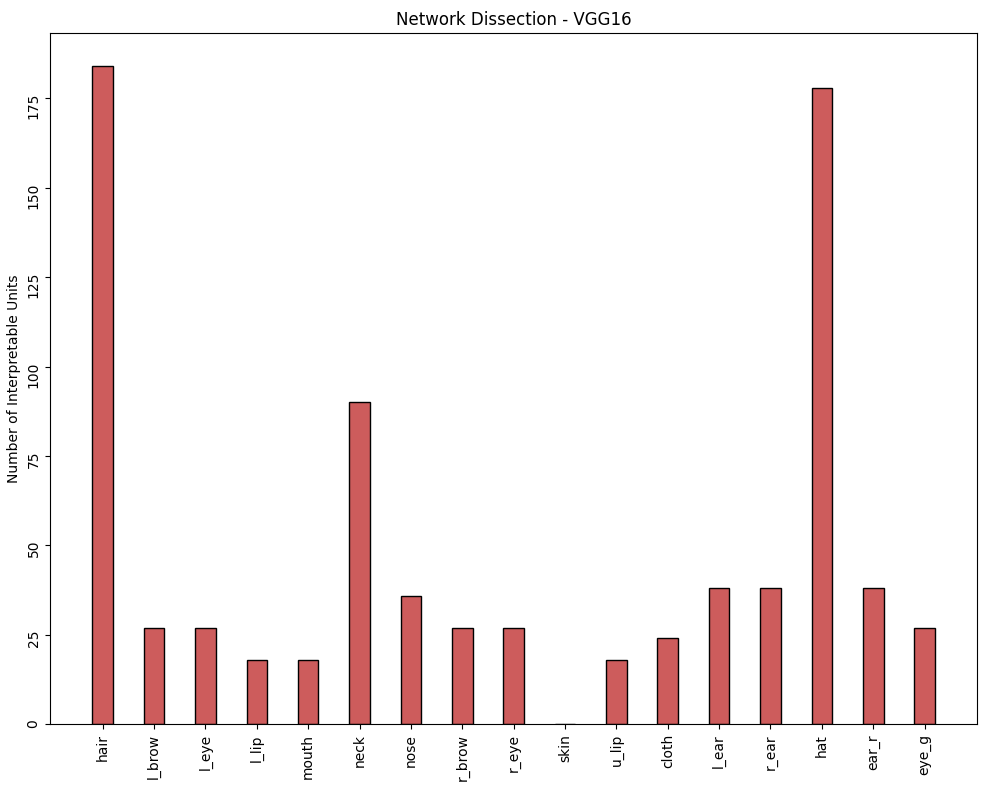
**Figure 21:** **Two Examples per Each Model of Different Concepts: a) Inception V3 b) VGG-16 c) ResNet-50**

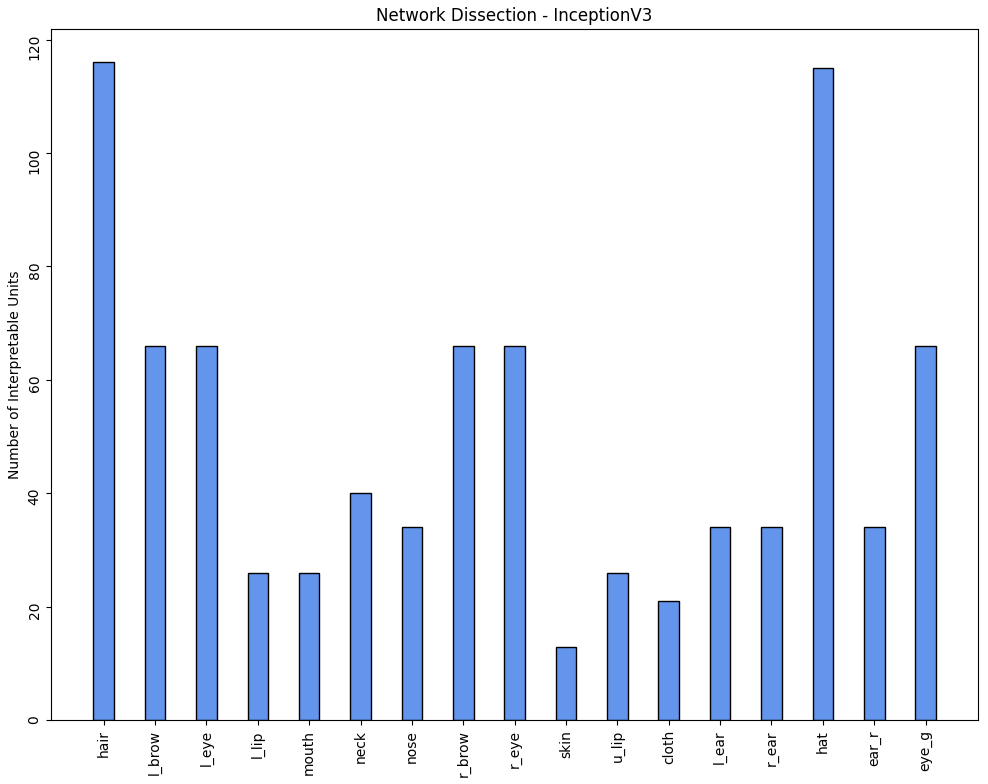
As stated before, Figure 22 shows samples of two facial concepts which describe the formulation of probabilistic hierarchy.



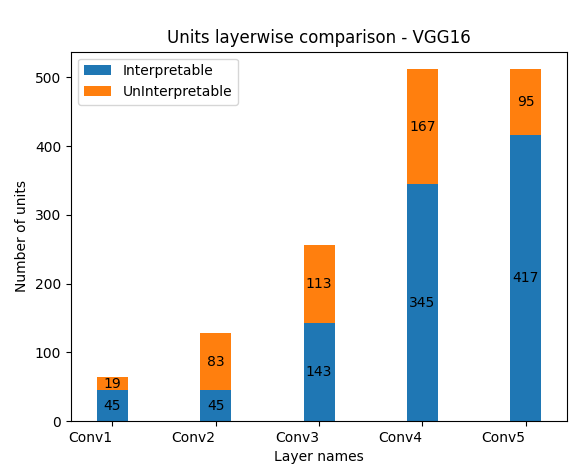


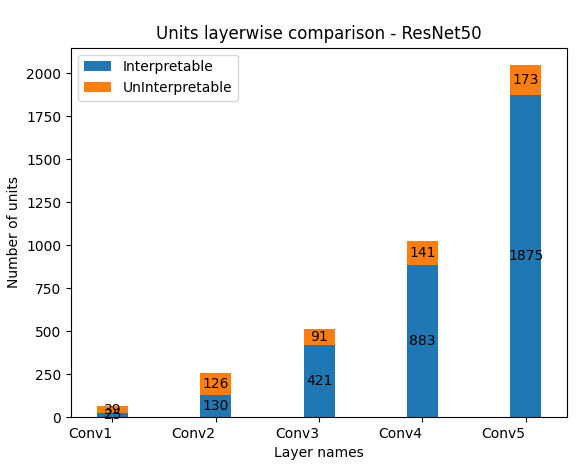
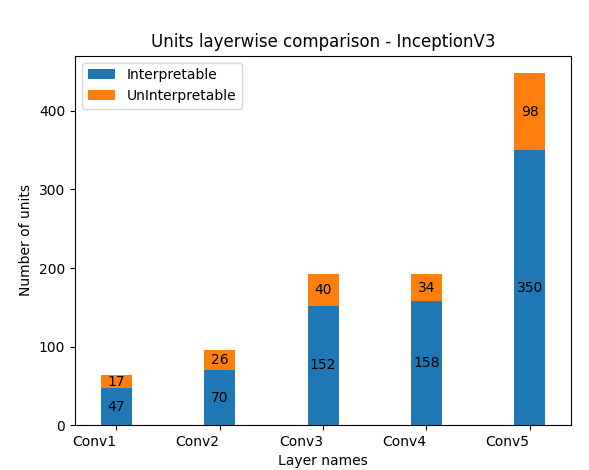
**Figure 22. Probabilistic Hierarchy Results**





**Figure 23. Dissection Histograms of the Three Models: a) VGG-16 b) ResNet-50 c) Inception V3**



**Figure 24. Number of Units Within Multiple Layers for Three Models: a) VGG-16 b) ResNet-50 c) Inception V3**

Figure 23 illustrates the interpretability reports of the dissected models and represents the number of units learned by each model for each facial concept for threshold value set as 0.04. This representation provides information about the top facial concepts detected by different CNN architectures. All the results are obtained for dissecting the final output layers of the models. Figure 24 shows the count of interpretable units per convolution layer for various trained models.

# DISCUSSION OF FINDINGS

## DeepFake Detection

Plenty of experiments have been conducted on the subject of deepfake detection, but we are yet in the transformative stages to find solutions for this problem. When models are evaluated on the same datasets they were trained on, nearly all of the studies exhibit outstanding performance scores, but if the models are tested on unseen data, they appear to struggle.

The results obtained from the proposed work detail the comparison of the efficiency of different CNN architectures for detecting DeepFake images.

The first part of the results chapter describes the comparison of three CNN models with various performance metrics and learning curves. When working with real-time data which is unbalanced, the training, validation, and testing accuracies are not effective performance indicators of models due to the vast difference between false negatives and false positives. Hence, it is relevant to include other metrics to evaluate the efficiency of the models other than accuracy. This is evident from the values of the F1-score, as it can be seen that all three models exhibit a good performance with respect to F1-score. F1-score ranges between 0.8 and 0.9 for the three models. On the other hand, the evaluation based on accuracy shows varied differences.

Among all the models, VGG-16 outperformed the other two models with an accuracy of 86% and an F1-score of 0.87. The study revealed that the increase in the number of parameters doesn’t imply an increase in performance metrics. As you can see in the case of ResNet-50 and Inception V3, both the models have nearly 5.3M trainable parameters but the efficiency of the models differ. This shows that efficiency depends on various factors.

To achieve the above results various experiments have been conducted with different hyperparameters. Initially, the experiments are executed using different values for batch sizes ranging from 16 to 64 for all the datasets and epochs ranging from 10 to 25. Similarly, several learning rates like 0.01, 0.001, and 0.00001 are used for selecting the optimal value. For ResNet-50 and Inception V3, batch sizes 16 and 64 displayed fluctuations in the validation curves. However, VGG-16 provided better results with a batch size of 16 and 0.0001 as the learning rate. Although there are fluctuations in the validation curve of VGG-16 results, both training and validation accuracies and loss functions converge to the same point after 15 epochs. For choosing the number of epochs the learning loss curve pattern is studied and by using the early stopping technique, overfitting issues are controlled to a great extent.

From the experiments, it is clear that the validation data has lots of fluctuations compared to the training data which can be practically possible with real-time data. Ultimately, by performing hyperparameter tuning the results were improved for all three models. The fluctuations in the validation data can be improved by implementing different cross-validation techniques for data split which is not performed during the study. In terms of training time for 15 epochs, ResNet50 has less training time compared to Inception V3 and VGG-16.

## XAI - Network Dissection

In addition to the detection of DeepFakes, this proposed work conducted various studies in understanding the internal parts of the developed CNN models by interpreting the facial features learned by these models. We know that neural networks have the ability to learn features by their own hidden layers and it’s considered as black-box models. The results from the explainable AI section help in understanding the internal neuronal representation of three models used for deepfake detection.

In general, the network dissection results provide an overview that the units of CNN models have learned a variety of facial concepts. Although, there are nearly fewer uninterpretable units for threshold 0.4 the types of different facial concepts learned by the models are comparatively less. Mostly all three models detected cheek region as the major concept followed by hair region. With the decreasing value for threshold, say to 0.005, the number of uninterpretable units increased which leads to fewer units detecting concepts. In comparison to all three threshold values, 0.04 is the better threshold value which identifies a variety of facial concepts for all CNN architectures used in the proposed study.

It is observed from Figure 20, that the hair region is one of the top facial regions detected by all three models for all threshold values. The cheek region contributes the most for threshold 0.4 for all three models, whereas for 0.04 and 0.005 threshold values neck and eye regions are detected by VGG-16 and ResNet-50 respectively. On the other hand, Inception V3 detects eye and ear regions for threshold values 0.04 and 0.005 respectively.

From Table 5, it is noticed that the top facial concepts learned by models include skin, hat, cloth, and earrings. In spite of the fact that higher threshold values lead to better IoU scores the types of facial features learned by the models are reduced in number. Hence, the representation of individual facial concepts learned by models is shown for threshold value 0.04 as given in Figure 23 which covers all the concepts.

All the experiments for the above results are carried out for dissecting the final layers of the neural networks. This is because dissecting the earlier layers of the models are avoided due to the model’s inability to learn higher-level concepts. This statement can be proven by analyzing the results from Figure 24. From the figure, it is observed that with the initial layers, the proportion of the number of interpretable units is less compared to the final layers of the models.

As previously stated, using the IoU approach inspired by the Network Dissection algorithm on the images from our DeepFake Face Dictionary, it is not feasible to say that a concept that produces an IoU that is similar to the top concept cannot be considered interpretable without doing an extensive analysis of the activations created for that concept with regard to other concepts that are located in the same facial regions. Consequently, we learn that one unit can be linked with more than one concept by creating a hierarchy among concepts from that facial region. This characteristic is explained in Figure 22 and clearly shows how challenging it is for a neuron to differentiate between concepts that look similar but have different features.

# CONCLUSION

With the advent of technology in the multimedia sphere, several malicious activities, such as the misuse of forged videos and images in different domains like politics and entertainment, have evolved as a negative element of it. This development in the field of multimedia has made manipulated and real images nearly identical, thus leading to several types of research in this domain. The most commonly used phrase, 'DeepFake,' has the capability of producing hyper-realistic visuals. To deal with the latest manipulation methods in visual media, an explainable DeepFake detection approach is critical. This research work aims in building a DeepFake detection technique to provide an explainable model.

The proposed research study involves the detection and interpretation of the DeepFake detection models using the Network dissection algorithm. This is implemented in two stages: 1) To detect forged images three state-of-the-art CNNs (VGG-16, ResNet-50, and Inception V3) are used, and a comparative analysis of these models. 2) Network dissection algorithm to interpret and understand the internal parts of the models for the explainable model. The proposed network dissection algorithm has the ability to understand the internal representations of these models and thereby interpret the different facial concepts learned by them in classifying the images as real or fake. It is compulsive to note that the technology used to generate manipulated images is also utilized to detect them, demonstrating that scientific and technical breakthroughs are both a benefit and a curse.

Experiments have proven that all the CNN models which are used in the proposed work has achieved a good performance with F1-scores ranging between 0.8 and 0.9. The explainability part of the proposed study is persuaded from Network Dissection algorithm which involves in generating interpretability reports for the three DeepFake detection models that have been trained on well-known DeepFake dataset. This also provides an extensive discussion on how model interpretability can be utilized to have a deeper understanding of the working of models and the representation of information or features learned by these models.

In spite of the fact that there are various DeepFake detection tools or models to identify the forged images or videos, it has been noticed that the current research doesn’t provide the reasoning or explanation for the classification of images using the Network dissection algorithm. This research work provides an overview of how the models differentiate the images from real or manipulated ones by understanding the internal neuronal representation of the models. It can be concluded that interpretability is a key aspect of deep neural networks that provide an understanding of their hierarchical structure.

# research cONTRIBUTION

Overall, the planned contribution of this research is to advance the knowledge and techniques in DeepFake detection and enhance the interpretability of deep learning models, thereby benefiting both the scientific community and society at large.

* Development of an explainable DeepFake detection technique: The proposed research aims to contribute to the field of multimedia technology by developing a DeepFake detection approach that not only identifies manipulated visuals but also provides an explanation for the classification. This could potentially enhance the reliability and trustworthiness of DeepFake detection systems.
* Utilization of the Network dissection algorithm for interpretability: By implementing the Network dissection algorithm, this research seeks to offer a novel approach to understanding the internal representations of deep learning models used for DeepFake detection. This could lead to better insights into how these models differentiate between real and fake images, thereby contributing to the interpretability of deep neural networks.
* Comparative analysis of state-of-the-art CNNs: Through the investigation of three popular CNN architectures (VGG-16, ResNet-50, and Inception V3) for detecting forged images, this research can provide valuable insights into the performance of different CNN models in tackling DeepFake manipulation.
* Insights into facial concepts learned by models: The research aims to shed light on the features and facial concepts learned by deep learning models in classifying images as real or fake. This understanding could pave the way for more accurate and robust DeepFake detection methods.
* Advancement in the field of explainable artificial intelligence: The interpretability aspect of the proposed approach could offer a deeper understanding of the working of deep neural networks. This could contribute to the broader field of explainable artificial intelligence, helping researchers and practitioners better understand the decision-making processes of complex deep learning models which are known as black-box models.

# FUTURE WORK

This research work inspires future work to develop more explainable and interpretable AI systems. However, it is significant to note that the interpretability of Network Dissection results is constrained because many facial concepts are not incorporated in the DeepFake Face Dictionary used for the proposed work. In the future, it can be suggested to include more facial concepts in the face dictionary with an increased count in the number of images. This helps to understand the interpretability of units on other facial features as well. After understanding the internal parts of the models, furthermore, the efficiency can be improved by adjusting the weights or bias corresponding to the activation of each neuron. Also, future works can include performing other XAI methods to compare and understand the benefits of the interpretability method used in this research work.

Regarding improving the performance of DeepFake detection models’ other techniques like cross-validation methods or building ensemble models could be done. Hence, in the future, there is extensive scope to explore and utilize more techniques and architectures that are relevant for efficient and modern ways of understanding the interpretations of models with respect to explainable artificial intelligence.

# LIST OF REFERENCES

1. Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. 2022. Deep learning for deepfakes creation and detection: A survey. Comput. Vis. Image Underst. 223, C (Oct 2022). <https://doi.org/10.1016/j.cviu.2022.103525>
2. K. Lin, W. Han, Z. Gu and S. Li, "A Survey of DeepFakes Generation and Detection," 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 2021, pp. 474-478, doi: 10.1109/DSC53577.2021.00076.
3. Khan, Sohail Ahmed & Artusi, Alessandro & Dai, Hang. (2021). Adversarially robust deepfake media detection using fused convolutional neural network predictions.
4. Ferreira, Sara & Antunes, Mario & Correia, Manuel Eduardo. (2021). A Dataset of Photos and Videos for Digital Forensics Analysis Using Machine Learning Processing. Data. 6. 87. 10.3390/data6080087.
5. Vamsi, V. V. V. N. S., “Deepfake detection in digital media forensics”, <i>Global Transitions Proceedings</i>, vol. 3, no. 1, pp. 74–79, 2022. doi:10.1016/j.gltp.2022.04.017.
6. X. Wang, J. Huang, S. Ma, S. Nepal and C. Xu, "DeepFake Disrupter: The Detector of DeepFake Is My Friend," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 14900-14909, doi: 10.1109/CVPR52688.2022.01450.
7. M. T. Jafar, M. Ababneh, M. Al-Zoube and A. Elhassan, "Forensics and Analysis of Deepfake Videos," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 053-058, doi: 10.1109/ICICS49469.2020.239493.
8. A. A. Pokroy and A. D. Egorov, "EfficientNets for DeepFake Detection: Comparison of Pretrained Models," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 598-600, doi: 10.1109/ElConRus51938.2021.9396092.
9. Rong, I. (2020, January 1). Detection and segmentation of deepfake face images generated by GANs using segmented based CNN. Detection and Segmentation of Deepfake Face Images Generated by GANs Using Segmented Based CNN. https://esource.dbs.ie/handle/10788/4244
10. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009.
11. Y. Xu, K. Raja and M. Pedersen, "Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection," 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2022, pp. 379-389, doi: 10.1109/WACVW54805.2022.00044.
12. Improving Deepfake Video Detection Using Data Augmentation Techniques. (n.d.). Improving Deepfake Video Detection Using Data Augmentation Techniques | Research Square. https://doi.org/10.21203/rs.3.rs-1844392/v1
13. Deepfake Technology - Organization for Social Media Safety. (n.d.). Organization for Social Media Safety. <https://www.socialmediasafety.org/advocacy/deepfake-technology/>
14. Westerlund, M. (2019, November 1). The Emergence of Deepfake Technology: A Review | TIM Review. The Emergence of Deepfake Technology: A Review | TIM Review. <https://timreview.ca/article/1282>
15. What are deepfakes? How fake AI-powered audio and video warps our perception of reality. (n.d.). Business Insider. <https://www.businessinsider.com/guides/tech/what-is-deepfake>
16. Team, G. L. (2021, May 14). All You Need to Know About Deepfake AI. Great Learning Blog: Free Resources What Matters to Shape Your Career! <https://www.mygreatlearning.com/blog/all-you-need-to-know-about-deepfake-ai/>
17. J. Bao, D. Chen, F. Wen, H. Li and G. Hua, "CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2764-2773, doi: 10.1109/ICCV.2017.299.
18. Agrawal, R. (2021, October 20). An End-to-End Introduction to Generative Adversarial Networks(GANs). Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/10/an-end-to-end-introduction-to-generative-adversarial-networksgans/>
19. Box, B. B. (2020, February 3). How Deepfake Technology Impact the People in Our Society? Medium. <https://becominghuman.ai/how-deepfake-technology-impact-the-people-in-our-society-e071df4ffc5c>
20. Nagahisarchoghaei, M.; Nur, N.; Cummins, L.; Nur, N.; Karimi, M.M.; Nandanwar, S.; Bhattacharyya, S.; Rahimi, S. An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives. Electronics 2023, 12, 1092. <https://doi.org/10.3390/electronics12051092>
21. P. (2019, November 19). Inside the Generative Adversarial Networks (GAN) architecture. Medium. <https://medium.com/@Packt_Pub/inside-the-generative-adversarial-networks-gan-architecture-2435afbd6b3b>
22. Agrawal, R. (2021, October 20). An End-to-End Introduction to Generative Adversarial Networks(GANs). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/an-end-to-end-introduction-to-generative-adversarial-networksgans/
23. Kakogeorgiou, Ioannis & Karantzalos, Konstantinos. (2021). Evaluating Explainable Artificial Intelligence Methods for Multi-label Deep Learning Classification Tasks in Remote Sensing.
24. Bento, V., Kohler, M., Diaz, P. et al. Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches. Discov Artif Intell 1, 9 (2021). <https://doi.org/10.1007/s44163-021-00008-y>
25. Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17). JMLR.org, 3145–3153.
26. A. Kapishnikov, T. Bolukbasi, F. Viegas and M. Terry, "XRAI: Better Attributions Through Regions," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 4947-4956, doi: 10.1109/ICCV.2019.00505.
27. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, C (Jun 2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
28. S. Xu, S. Venugopalan and M. Sundararajan, "Attribution in Scale and Space," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9677-9686, doi: 10.1109/CVPR42600.2020.00970.
29. Pino, Samuele & Carman, Mark & Bestagini, Paolo. (2021). What's wrong with this video? Comparing Explainers for Deepfake Detection.
30. Baldassarre, Federico, Quentin Debard, Gonzalo Fiz Pontiveros and Tri Kurniawan Wijaya. “Quantitative Metrics for Evaluating Explanations of Video DeepFake Detectors.” British Machine Vision Conference (2022).
31. Shao, R., Wu, T., Liu, Z. (2022). Detecting and Recovering Sequential DeepFake Manipulation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13673. Springer, Cham. <https://doi.org/10.1007/978-3-031-19778-9_41>
32. Dong, S., Wang, J., Liang, J., Fan, H., Ji, R. (2022). Explaining Deepfake Detection by Analysing Image Matching. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13674. Springer, Cham. <https://doi.org/10.1007/978-3-031-19781-9_2>
33. S. Narteni, V. Orani, E. Ferrari, D. Verda, E. Cambiaso and M. Mongelli, "A New XAI-based Evaluation of Generative Adversarial Networks for IMU Data Augmentation," 2022 IEEE International Conference on E-health Networking, Application & Services (HealthCom), Genoa, Italy, 2022, pp. 167-172, doi: 10.1109/HealthCom54947.2022.9982780.
34. R. R. Kontham, A. K. Kondoju, M. M. Fouda and Z. M. Fadlullah, "An End-To-End Explainable AI System for Analyzing Breast Cancer Prediction Models," 2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), BALI, Indonesia, 2022, pp. 402-407, doi: 10.1109/IoTaIS56727.2022.9975896.
35. V. Pitroda, M. M. Fouda and Z. M. Fadlullah, "An Explainable AI Model for Interpretable Lung Disease Classification," 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), Bandung, Indonesia, 2021, pp. 98-103, doi: 10.1109/IoTaIS53735.2021.9628573.
36. S. Bhattacharjee, Y. -B. Hwang, K. Ikromjanov, R. I. Sumon, H. -C. Kim and H. -K. Choi, "An Explainable Computer Vision in Histopathology: Techniques for Interpreting Black Box Model," 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Korea, Republic of, 2022, pp. 392-398, doi: 10.1109/ICAIIC54071.2022.9722656.
37. J. G, P. S and S. D, "An Explainable Deep Learning Model for Vision-based Human Fall Detection System," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1223-1229, doi: 10.1109/ICICICT54557.2022.9917979.
38. K. A. Hasenstab, J. Huynh, S. Masoudi, G. M. Cunha, M. Pazzani and A. Hsiao, "Feature Interpretation Using Generative Adversarial Networks (FIGAN): A Framework for Visualizing a CNN’s Learned Features," in IEEE Access, vol. 11, pp. 5144-5160, 2023, doi: 10.1109/ACCESS.2023.3236575.
39. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W. (2022). Explainable AI Methods - A Brief Overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science(), vol 13200. Springer, Cham. <https://doi.org/10.1007/978-3-031-04083-2_2>
40. S. -H. Kang, H. -G. Jung and S. -W. Lee, "Interpreting Undesirable Pixels for Image Classification on Black-Box Models," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 4250-4254, doi: 10.1109/ICCVW.2019.00523.
41. Montavon, Grégoire & Samek, Wojciech & Müller, Klaus-Robert. (2018). Methods for Interpreting and Understanding Deep Neural Networks. Digital Signal Processing. 73. 1-15. 10.1016/j.dsp.2017.10.011.
42. Bau, David & Zhu, Jun-Yan & Strobelt, Hendrik & Lapedriza, Àgata & Zhou, Bolei & Torralba, Antonio. (2020). Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences. 117. 201907375. 10.1073/pnas.1907375117.
43. L. Trinh, M. Tsang, S. Rambhatla and Y. Liu, "Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 1972-1982, doi: 10.1109/WACV48630.2021.00202.
44. K. Siddiqui and T. E. Doyle, "Trust Metrics for Medical Deep Learning Using Explainable-AI Ensemble for Time Series Classification," 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Halifax, NS, Canada, 2022, pp. 370-377, doi: 10.1109/CCECE49351.2022.9918458.
45. Khan, Mohammad & Alam, Kazi Nabiul. (2022). Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods. Intelligent Automation and Soft Computing. 35. 10.32604/iasc.2023.029653.
46. Daglarli, Evren. (2020). Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. 10.5772/intechopen.92172.
47. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 2021, 23, 18. <https://doi.org/10.3390/e23010018>
48. Bartelt, Christiann & Marton, Sascha & Stuckenschmidt, Heiner. (2020). xRAI: Explainable Representations through AI.
49. D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3319-3327, doi: 10.1109/CVPR.2017.354.
50. B. Zhou, D. Bau, A. Oliva and A. Torralba, "Interpreting Deep Visual Representations via Network Dissection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9, pp. 2131-2145, 1 Sept. 2019, doi: 10.1109/TPAMI.2018.2858759.
51. Molnar, C. (2023, July 17). Interpretable Machine Learning. Interpretable Machine Learning. <https://christophm.github.io/interpretable-ml-book/>
52. Teotia, D., Lapedriza, A. & Ostadabbas, S. Interpreting Face Inference Models Using Hierarchical Network Dissection. Int J Comput Vis 130, 1277–1292 (2022). <https://doi.org/10.1007/s11263-022-01603-x>
53. A. Rawal, J. McCoy, D. Rawat, B. Sadler and R. Amant, "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives" in IEEE Transactions on Artificial Intelligence, vol. 3, no. 06, pp. 852-866, 2022.doi: 10.1109/TAI.2021.3133846
54. KDD Process in Data Mining - Javatpoint. (n.d.). www.javatpoint.com. <https://www.javatpoint.com/kdd-process-in-data-mining>
55. Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.
56. Team, G. L. (2021, October 1). Introduction to VGG16 | What is VGG16? Great Learning Blog: Free Resources What Matters to Shape Your Career! <https://www.mygreatlearning.com/blog/introduction-to-vgg16/>
57. Understanding VGG16: Concepts, Architecture, and Performance. (n.d.). Datagen. <https://datagen.tech/guides/computer-vision/vgg16/>
58. Ali, L.; Alnajjar, F.; Jassmi, H.A.; Gocho, M.; Khan, W.; Serhani, M.A. Performance Evaluation of Deep CNN-Based Crack Detection and Localization Techniques for Concrete Structures. Sensors 2021, 21, 1688. <https://doi.org/10.3390/s21051688>
59. Y. (2023, July 12). GitHub - yuezunli/celeb-deepfakeforensics: Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. GitHub. <https://github.com/yuezunli/celeb-deepfakeforensics>
60. CelebAMask-HQ Dataset. (n.d.). CelebAMask-HQ Dataset. <https://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html>
61. Papers with Code - Inception-v3 Explained. (n.d.). Inception-v3 Explained | Papers With Code. <https://paperswithcode.com/method/inception-v3>
62. Precision, Recall, Sensitivity and Specificity. (2021, November 4). OpenGenus IQ: Computing Expertise & Legacy. <https://iq.opengenus.org/precision-recall-sensitivity-specificity/>
63. How to interpret F1 score (simply explained). (2022, August 19). Stephen Allwright. <https://stephenallwright.com/interpret-f1-score/>

# AFFIDAVIT

Under the penalty of perjury, I, **Nazneen** **Mansoor**, hereby declare under oath, that this master thesis, titled

**Explainable AI for DeepFake Detection** **Network Dissection Approach**

has been independently authored by me, that I have used no other sources or aids other than those quoted, and that I have denoted all direct and indirect citations thereof within the text. This manuscript has never before been published or submitted to this or another academic institution for examination.

|  |  |
| --- | --- |
| Berlin, 10.08.2023 |  |
| Location, Date | Signature of Author |

