

Article

Explainable AI for DeepFake Detection

Nazneen Mansoor ^{1,*} and Alexander I. Iliev ^{1,2,*}¹ Berlin School of Technology, SRH Berlin University of Applied Sciences, D-10587 Berlin, Germany² Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria

* Correspondence: nazneenm1990@gmail.com (N.M.); ailiev@berkeley.edu (A.I.I.)

Abstract: The surge in technological advancements has resulted in concerns over its misuse in politics and entertainment, making reliable detection methods essential. This study introduces a deepfake detection technique that enhances interpretability using the network dissection algorithm. This research consists of two stages: (1) detection of forged images using advanced convolutional neural networks such as ResNet-50, Inception V3, and VGG-16, and (2) applying the network dissection algorithm to understand the models' internal decision-making processes. The CNNs' performance is evaluated through F1-scores ranging from 0.8 to 0.9, demonstrating their effectiveness. By analyzing the facial features learned by the models, this study provides explainable results for classifying images as real or fake. This interpretability is crucial in understanding how deepfake detection models operate. Although numerous detection models exist, they often lack transparency in their decision-making processes. This research fills that gap by offering insights into how these models distinguish real from manipulated images. The findings highlight the importance of interpretability in deep neural networks, providing a better understanding of their hierarchical structures and decision processes.

Keywords: explainable artificial intelligence; deep learning; deepfake detection; explainability; convolutional neural network; VGG-16; ResNet-50; inception V3; network dissection; face dictionary; model interpretability



Academic Editor: Jing Jin

Received: 24 October 2024

Revised: 25 November 2024

Accepted: 6 December 2024

Published: 13 January 2025

Citation: Mansoor, N.; Iliev, A.I. Explainable AI for DeepFake Detection. *Appl. Sci.* **2025**, *15*, 725. <https://doi.org/10.3390/app15020725>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Deep learning technology has advanced considerably in recent years, with notable developments in the area of computer vision. Its applications span diverse fields, including finance, the automotive sector, wellness [1–4], and healthcare, offering solutions to complex challenges. However, one downside of deep learning is its potential to create manipulated videos, images, and various multimedia forms, which create risks to privacy, national security, and democratic systems. With the widespread growth in modern technology concerning mobile phones and other editing tools, ease of access to social media platforms has made it convenient to create and distribute digital content [5–7]. These deceptive media are referred to as ‘deepfakes’. Immense amounts of research in the computer vision domain, machine learning, and the considerable amount of data availability make it harder to differentiate between altered and real media content. This has detonated social media platforms with forged videos, images, and fake content that appears realistic [5]. Lately, this has imposed a significant threat to society, business, democracy, and public discussion. Within the past few years, the journalism industry has been struggling to separate real content from fake news [5–7]. It harms the cybersecurity and national security of individuals and organizations.

As information and technology continue to evolve, the creation of deepfake content has become easier and it is an emerging concern in today's digital environment. There are different AI techniques and tools that help in creating these manipulated images. However, the creation of fake images is increasing to a higher level, which makes it difficult for observers to differentiate between fake and real images. Images generated by GANs (generative adversarial networks) have a wide range of applications, including text-to-image or image-to-image conversion, photo editing and blending, facial aging simulation, creating new human poses, generating 3D visuals, and restoring missing parts of photos through inpainting. However, it would be one-sided to dismiss the impact of deepfakes. GAN-generated images can convincingly fool the human eye. Additionally, several deepfake creation tools and software are publicly available, allowing users to produce highly realistic content. As deepfake generation algorithms advance rapidly, and the quality of these creations continues to improve, reliable detection tools have become increasingly important in today's society. A deep learning model is employed in this study to distinguish between authentic and fake instances. To construct the model, three distinct CNN architectures are utilized, allowing for a comparison of their effectiveness in detecting deepfakes.

While detection accuracy remains a vital focus, the interpretability of deep learning models is equally important to ensure their widespread acceptance in real-world applications. Users, decision-makers, and stakeholders must understand why a particular instance is classified as fake or real. Without this understanding, the trustworthiness and adoption of these systems may be limited, particularly in critical domains such as journalism, law enforcement, and content moderation. Additionally, the lack of transparency could hinder forensic investigations and legal proceedings where evidence-based reasoning is required. To address these challenges, this research incorporates explainable artificial intelligence (XAI) techniques to provide clear and interpretable explanations for the decisions made by the deepfake detection models. By doing so, this study aims to enhance trust in the system's outputs and improve its practicality in addressing the societal risks posed by deepfakes.

Explainable artificial intelligence stands out as an interesting field of study that has been emerging in recent times. Figure 1 shows the number of publications in the realm of explainable artificial intelligence for the last two decades and the plot shows how explainable artificial intelligence is blooming these days. This research deals with the reasoning behind the predictions made by the neural networks in classifying them as real or fake using explainable AI. 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 explainability and limited transparency of 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. Secondly, 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. Interpretability is crucial for fostering trust among end-users and stakeholders, a key factor in the widespread adoption of deepfake detection technologies. Given these

challenges, it has become clear that developing interpretation methods especially designed for neural networks is necessary.

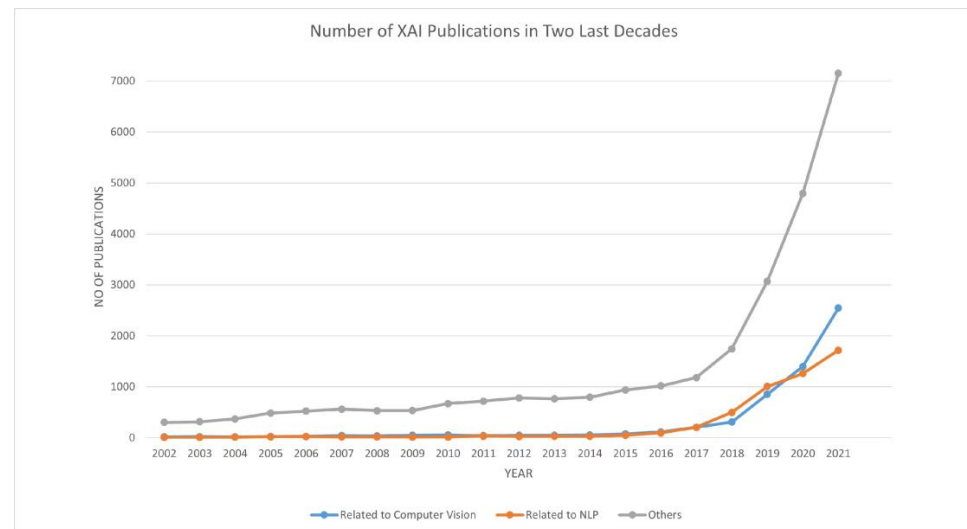


Figure 1. The Number of Publications Related to XAI [8].

With the help of explainable artificial intelligence, this research proposal aims to provide explainability for deepfake detection models, thereby improving their reliability. To explore the explainability potential of a deepfake detection model, a multi-step process is essential. This process involves creating a model that classifies images into real or fake, and then incorporating explainable AI techniques to enhance the transparency of the model's reasoning 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 acts as the foundation for training the model, enabling 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. Achieving the optimal balance between realism and diversity in the dataset is crucial for training a robust model that can accurately detect subtle deepfake modifications. The trustworthiness of AI systems depends on their explainability, especially when they are tasked with critical functions such as 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 [9–11]. This research contributes to addressing the pressing challenges in deepfake detection by leveraging a deep learning model combined with explainable artificial intelligence (XAI) techniques. By integrating XAI, this study not only aims to enhance detection accuracy but also focuses on the interpretability of the results, providing stakeholders with clear and evidence-based explanations for classification decisions. Unlike existing detection systems, which often lack transparency, our approach bridges this gap by offering insights into the internal decision-making process of the neural networks. The overview of this paper is as follows: Section 2 explores previous efforts in the fields of deepfake detection and explainability, highlighting the gaps this research seeks to address. Section 3 details the CNN architectures and XAI techniques employed. The Section 4 evaluates the performance of the proposed models and the Section 5 explains

their interpretability. Finally, Section 6 summarizes the findings and discusses the broader implications of this work.

2. Related Works

Within the broader fields of artificial intelligence (AI) and computer vision, explainable artificial intelligence (XAI) is gaining growing significance. It is characterized by the development of various techniques and approaches aimed at interpreting and elucidating the predictions generated by complex deep learning algorithms [9]. This rapidly growing field focuses on enhancing the transparency and interpretability of AI systems for human users, fostering trust, and facilitating the effective management of these complex “black box” models. The relevance and growth of XAI can be attributed to the widespread adoption of deep learning solutions across diverse industries. In tasks such as natural language processing, autonomous decision-making, and image recognition, deep learning models have proven highly effective. 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 such as healthcare, finance, and autonomous vehicles, where accountability and interpretability are paramount. XAI aims to close this gap by creating machine learning techniques that allow humans to better understand, trust, and manage AI systems effectively. These techniques serve as a means to uncover the decision-making processes within deep learning models. XAI offers clear explanations for model predictions, enabling users to make informed decisions based on AI recommendations while also helping developers improve the model’s robustness and fairness. 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 approaches that empower human users to understand, trust, and control the next wave of artificial intelligence partners [10,12,13]. Given the high complexity of deep learning models, it becomes challenging to understand and assess the internal decision process underlying predictions. In the case of images, explainability describes which parts of an image caused the model to predict specific classifications. Image explanations are essential for two different groups of people, 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 2 describes the scenario with and without explainable artificial intelligence. Unlike the traditional deep learning approaches which help in future predictions, explainable AI models provide the reasoning behind the predictions. For explainability models, two elements draw attention along with the new machine learning algorithm, as shown in Figure 2 [12]. One of them is the explanatory model and the other is the explanation interface.

Neural network predictions involve millions of mathematical calculations determined by the structure of the neural network. It is difficult to understand the predictions made by neural networks as we need to consider the many 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 features and concepts from the hidden layers and special tools are required to uncover them [14]. Secondly, other model-agnostic approaches such as local models or partial dependence plots are mainly useful for tabular data, and text and image data require different techniques. The commonly used methods to explain neural network models include LIME, SHAP, gradient-based methods, saliency maps, Grad-CAM, Occlusion, DeepLift, and others [13,15].

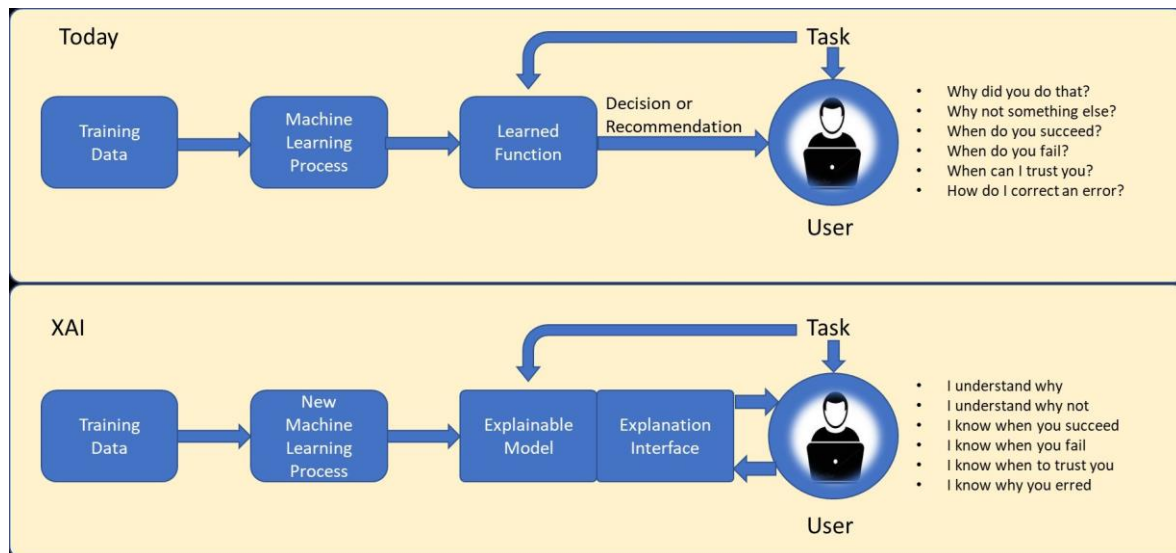


Figure 2. Explainable Artificial intelligence (XAI) Approach [12].

The existing approaches or techniques used by researchers for deep learning algorithms developed for deepfake detection can be explained in this section. As discussed in [6,7], Nazneen has explored different datasets and model architectures, contributing to advancements in distinguishing between real and manipulated media, which are crucial for forensic applications and combating misinformation. Vitor [16] 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 [17] used XAI methods such as GradCAM and SHAP for explaining the predictions made by an EfficientNet model trained on a 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 [18] proposed a quantitative framework to estimate the informativeness and visual quality of explanations of deepfake classifiers. The paper highlights the need for intuitive explanations to validate deepfake detections, ensuring both user trust and compliance with legal frameworks such as GDPR. It evaluates existing explainable AI (XAI) methods, such as saliency maps, Grad-CAM, SmoothGrad, SHAP, and LIME, for their effectiveness in generating interpretable explanations for visual classifiers. They benchmark state-of-the-art video recognition models on popular datasets and explore how heatmap-based explanations can be communicated effectively to users, offering insights for improving the interpretability of deepfake detection systems. Shichao [19] proposed an FST-matching model (fake source target matching) for the interpretation of deepfake detection to learn the features of images that are used to categorize them as fake or real. Sara [20] introduced an evaluation method for GANs through the XAI technique known as a logic learning machine, which has been applied to enhance data augmentation techniques. Loc [21] proposed an effective and interpretable method known as a dynamic prototype network for explaining deepfake temporal artifacts. This paper introduces DPNet, an XAI (explainable AI) approach for deepfake detection that emphasizes interpretability and temporal dynamics. Unlike traditional methods, which primarily focus on frame-by-frame analysis and offer limited explanations, DPNet enhances transparency by learning

the dynamic prototypes of temporal inconsistencies in videos. These prototypes serve as case-based evidence, allowing the model to explain why a particular prediction was made in a way that reflects its underlying computational process. Additionally, DPNNet provides visual dynamic explanations by projecting learned prototypes onto representative video patches, making the detection process accessible and understandable for humans. Wahidul [22] used the LIME XAI technique to interpret the CNN model used for identifying real and fake images, which also detailed the part of the image that caused it to make the specific classification. This paper focuses on deepfake detection using deep learning (DL) models, specifically convolutional neural networks (CNNs), and enhances their interpretability through explainable AI (XAI) methods, particularly local interpretable model-agnostic explanations (LIME). It examines the efficacy of different CNN models, including InceptionResNetV2, DenseNet201, InceptionV3, and ResNet152V2, in classifying real and deepfake images from a dataset containing 70k real images and 70k fake faces generated by StyleGAN. This approach combines high performance with XAI to provide a reliable and interpretable solution for deepfake detection.

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 [9,10]. Network dissection is defined as an interpretability technique for convolutional neural networks (CNNs) that 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. In [9], the researchers suggested a general framework 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 such as ResNet, AlexNet, VGG, and GoogLeNet, all of which are trained on objects and scenic concepts.

This review highlights that the integration of XAI methods across various domains has demonstrated both quantitative and qualitative improvements, fostering greater trust in AI-driven decisions and predictions. As trust plays a crucial role in assessing the reliability of AI models, explainable AI serves as a pivotal approach in enhancing the transparency and trustworthiness of these systems. As an extension to our published proceedings [6,7], this research aims 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.

3. Experiments

In recent years, considerable progress has been made within the domain of computer vision, especially in the areas of image manipulation and detection. With the proliferation of manipulated or fake images and videos, detecting such alterations has become a critical challenge. To test the explainability of deepfake detection models, the proposed work has been categorized into two stages—detection and explainability. 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. During the detection phase, the classification process is carried out using three CNN models: ResNet-50, Inception V3, and VGG-16. A comparison is then made based on the accuracy of each model to determine how effectively they can distinguish between authentic and manipulated 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 images as real or fake. The implementation stages of deepfake detection and explainable artificial intelligence are portrayed in Figures 3 and 4.

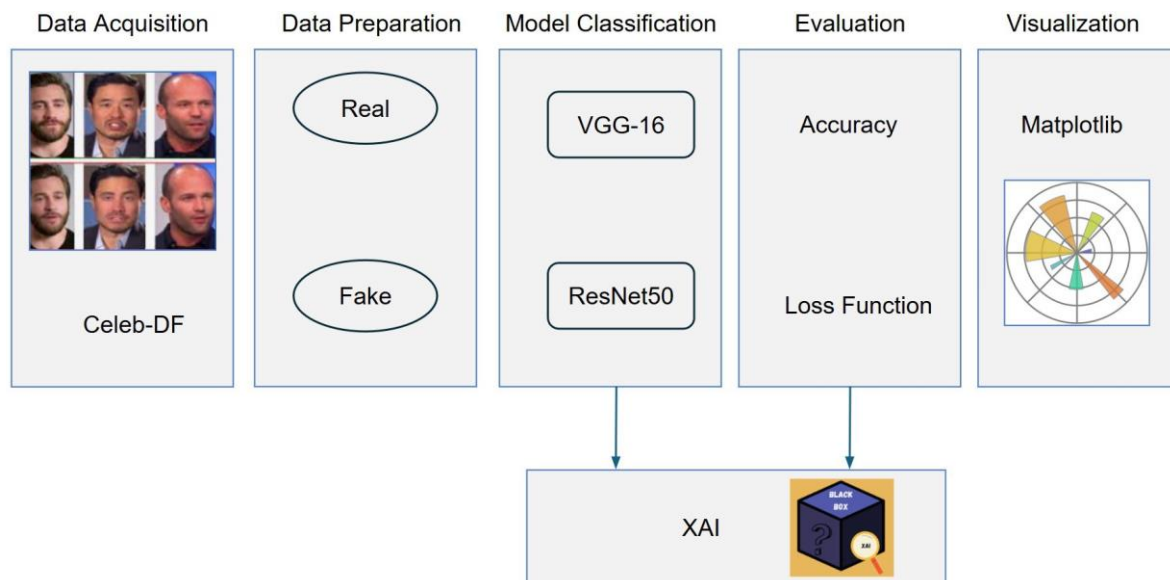


Figure 3. Implementation Stages for DF Detection.

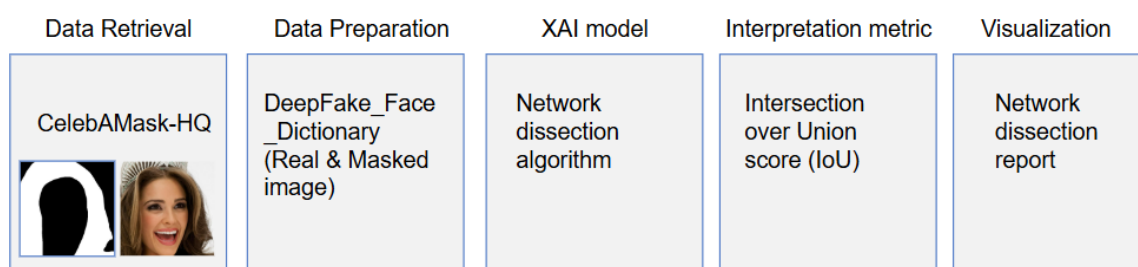


Figure 4. Implementation Stages for XAI.

The dataset used to train the three convolutional neural network models is crucial in advancing deep learning systems for image classification. In this instance, the dataset contains a mix of real and fake images, taken from the Celeb-DF dataset [23]. This dataset is crucial for advancing deep learning and computer vision research, particularly in addressing the challenges posed by deepfake videos, which have the potential to spread misinformation and manipulate digital media. The Celeb-DF dataset provides researchers with a substantial collection of both deepfake and genuine videos, making it an invaluable resource in this field. The dataset comprises a considerable number of images, with 19,457 frames obtained from the videos. To effectively train, validate, and test the CNN models, the dataset is divided in a balanced manner. Specifically, a percentage of the images, equivalent to 15,565 frames, are used for training. This comprehensive training set enables the models to identify key features and patterns essential for differentiating between deepfake and genuine content. Additionally, a percentage of the data are designated for the purpose of validation, a key process in monitoring model performance and adjusting hyperparameters to avoid overfitting, ensuring the model's capability to be applied to unseen data. Lastly, the final 10 percent of the dataset is set aside for testing purposes, providing a final evaluation of the CNN models' effective-

ness in classifying real and fake images accurately. Figure 5 presents a visual sample of both fake and real images belonging to the Celeb-DF dataset that were utilized for training purposes.

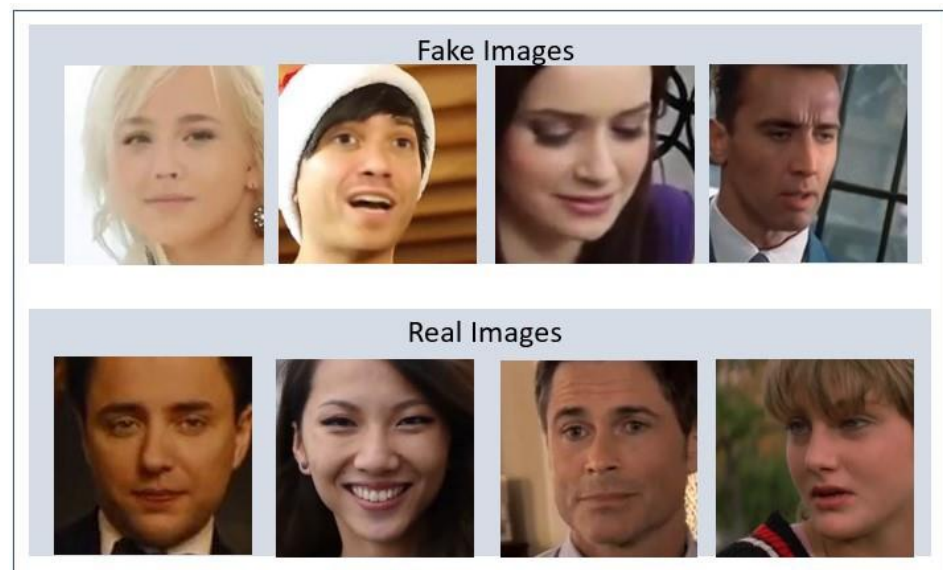


Figure 5. Real and Fake Images—Training Dataset [23].

To test the explainability potential of the models, the network dissection algorithm is used to dissect various layers of the CNN models. This aims to assess the feasibility of implementing the deepfake detection models. Unit–concept mapping for network dissection is built upon our deepfake face dictionary, featuring numerous facial concepts. In the proposed work, the network dissection algorithm uses the deepfake face dictionary that incorporates concepts of different features of faces. 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 6 describes an overview of the probabilistic approach to determine what units of the CNN models serve as facial concept detectors, which helps in categorizing images as real or fake. The proposed method involves performing a forward pass through all images in the deepfake face dictionary to collect activation maps from the dissected layer, followed by running the network dissection algorithm to calculate IoU scores for each concept at the unit level. Intersection over union (IoU) is a metric used to measure the overlap between two regions, typically a predicted area and a ground truth area.

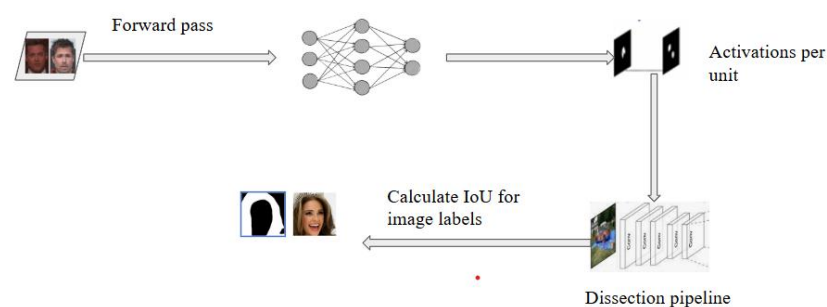


Figure 6. XAI Approach for Dissecting CNN Models.

3.1. Dataset Description

Our deepfake face dictionary includes 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 including masked face images [24]. 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. Among them, the labels that were used in the proposed work are nose, skin, ears, eyes, lip, hair, mouth, hat, neck, earrings, eyeglasses, and cloth. Figure 7 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 8 shows the histogram of the number of labeled occurrences for each facial concept present in the deepfake face dictionary.

In this section, we describe the process of extracting activation maps for all input images in the ‘Deepfake Face Dictionary’ dataset, which contains labeled facial concepts. The extraction of these activation maps is carried out for each layer of the selected CNN models, with the choice of model determined by the specific architecture being evaluated. For each input image (denoted as N), the activation maps are generated by running the model through the script that stores the activations in a NumPy memory map. These activations represent the response of each neuron in the layer when processing the input images. The size of the activation maps decreases in deeper layers of the network due to lower resolution at those levels.

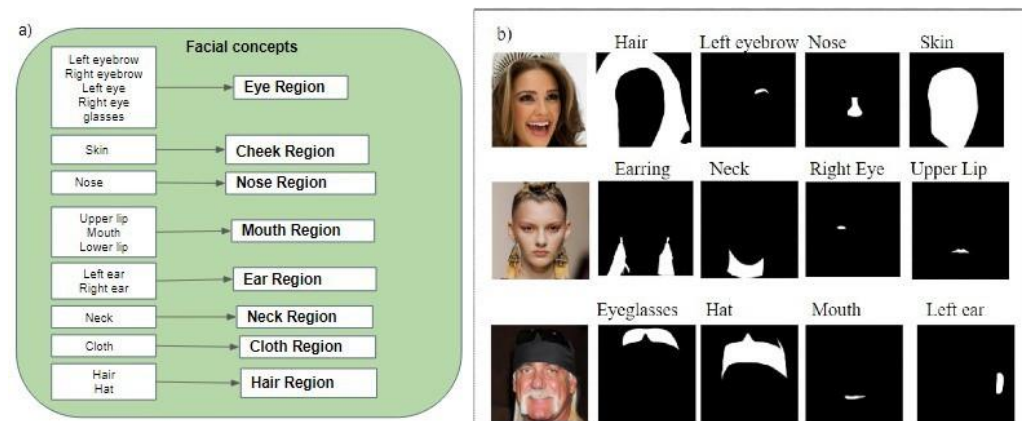


Figure 7. Deepfake Face Dictionary. (a) Facial Concepts and corresponding regions from the dictionary. (b) Samples of images and respective segmentation masks for facial concepts.

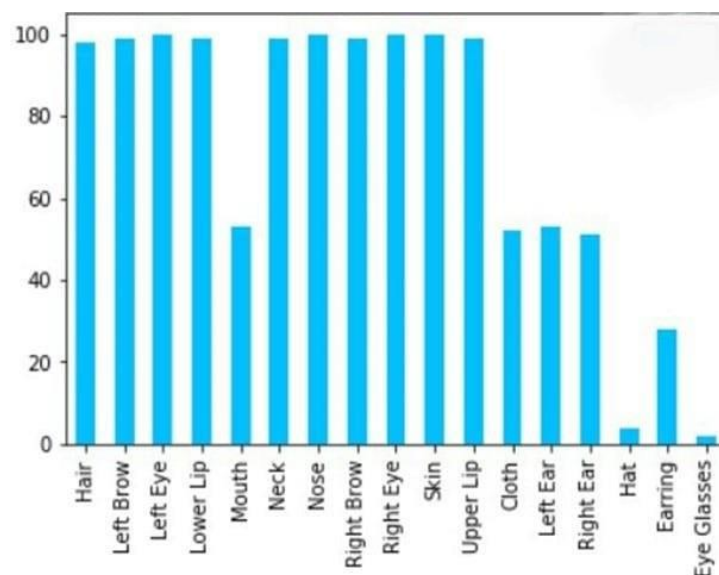


Figure 8. Number of Images per Facial Concept in our Deepfake Face Dictionary.

3.2. Implementation

The network dissection algorithm is used to dissect the CNN models trained to categorize the images as genuine or fake. Figure 9 provides a general overview of network dissection for deepfake detection CNN models.

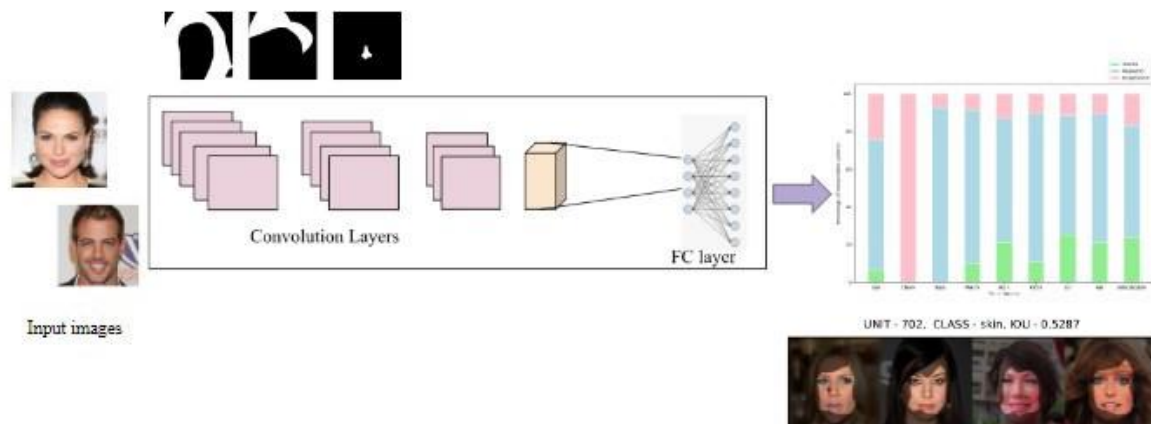


Figure 9. Overview of Network Dissection for Deepfake Detection Models.

Next, we evaluate threshold values for each unit within the activation maps. This is achieved by analyzing spatial features and setting a quantile threshold for any spatial region with values above this threshold. The experiments are conducted using different threshold values (0.4, 0.04, and 0.005) to segment the activation maps. These segments correspond to the regions in the image that are most salient for distinguishing between real and fake content. The threshold values determine how many units are interpretable. For each unit, the intersection over union (IoU) score is calculated to quantify how well the unit corresponds to a specific facial concept. For each unit–concept pair, the IoU score is computed. A neuron is considered interpretable if its IoU score for a concept exceeds a defined threshold (0.04 in this case). The IoU score provides a measure of the overlap between the unit’s activation map and the facial concept it represents.

Table 1 provides a summary of the models that were dissected, outlines the layer that was dissected during the experiments, and depicts the results for each model. These interpretability experiments focused on the deeper layers of the models, as previous studies have shown that high-level concepts are primarily located in the layers 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 the intersection over union (IoU) score, which is given in Equation (1).

Table 1. Dissected Deepfake Detection models.

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

Equation (1): IoU Score Calculation

$$\text{Intersection Over Union} = (\text{Area of Intersection}) / (\text{Area of Union}) \quad (1)$$

The concepts with the highest IoU score for each unit are noted in the output net dissect report, and the unit is considered interpretable for top concepts with values greater than 0.04. Although we can obtain the dominant concepts for each unit based on IoU scores, sometimes

there is more than one concept with a high IoU score and it lies in a similar area of the face. In these situations, it is better to find a hierarchy of concepts that lie in the same facial region as that of the concept with the high IoU score. This is achieved by identifying the facial region and generating probabilities for each concept that lies within that facial region.

4. Findings and Evaluation

This research work employs three widely used CNN architectures for deepfake detection: Resnet50, Inception V3, and VGG-16. To assess how well these models perform, metrics such as accuracy, recall, precision, and F1-score are applied. The test data results are presented in Table 2.

Table 2. Performance Comparison of CNN models.

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

Figures 10–12 depict the learning curves for CNN models trained with various hyperparameters, illustrating the accuracy and loss trends over the course of the epochs. The plots reflect the results across 15 epochs. The Adam optimizer was consistently applied across all three models. ResNet-50 and Inception V3 were trained with a learning rate of 0.00001 and a batch size of 32, whereas VGG-16 was trained using a batch size of 16 and a learning rate of 0.0001. These hyperparameters were selected based on several experimental trials conducted on the models.

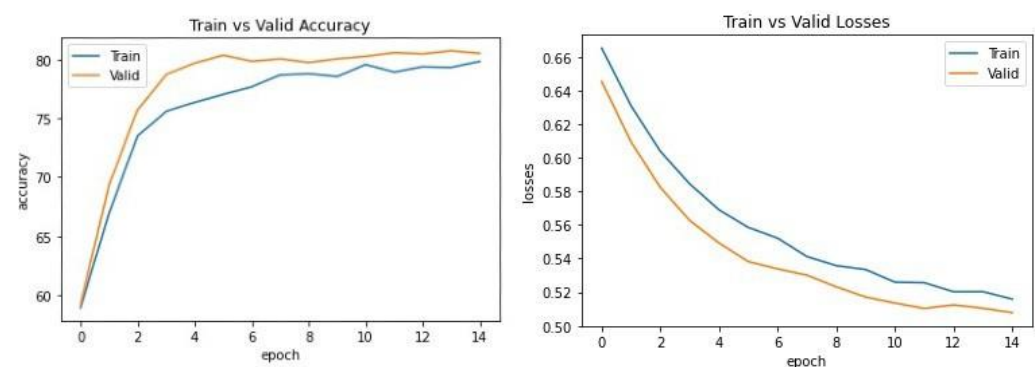


Figure 10. Learning Curves of ResNet-50.

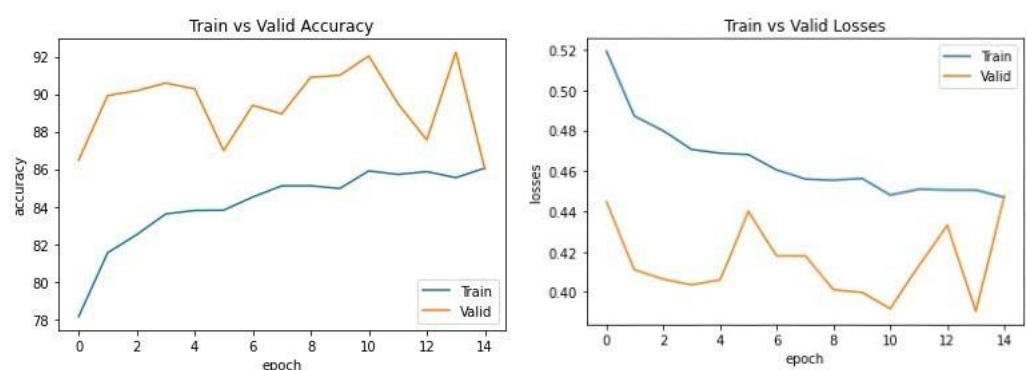


Figure 11. Learning Curves of VGG-16.

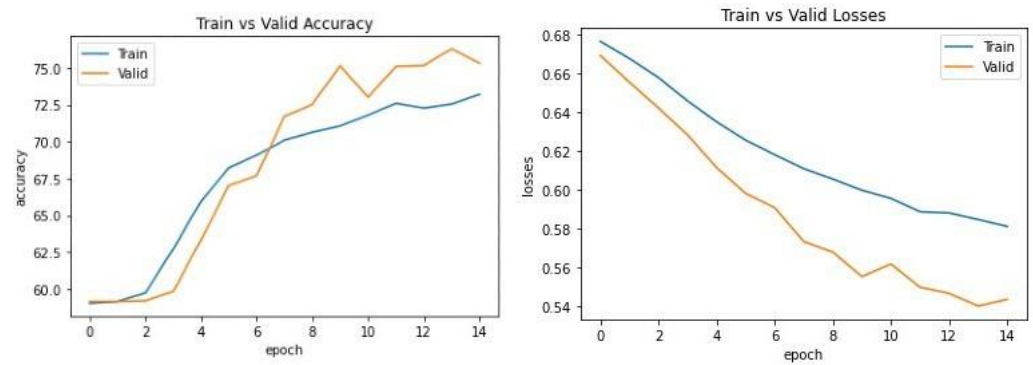


Figure 12. Learning Curves of Inception V3.

This section of the proposed work aims to explain the underlying decision process of the deepfake detection models. After performing multiple experiments with a 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. To test the explainability of the models, different threshold values are set to compare the features learned by different architectures. This provides an overview of how models learn different facial concepts based on segmentation. Figures 13–15 depict the number of interpretable units detected by each model with regard to the various facial regions for different threshold values. Figure 16 represent a sample of a few of the facial concepts, along with their IoU scores as detected by the models. In Tables 3–5, information about the amount of interpretable and uninterpretable units, along with the top IoU scores and their corresponding facial concepts, is listed. Additionally, the top intersection over union (IoU) scores and their associated facial concepts are included, reflecting the models' effectiveness in identifying critical facial features. Together, these results illustrate the comparative performance of the CNN architectures in capturing distinct and relevant features for distinguishing between real and fake images, underscoring their interpretability and relevance in deepfake detection tasks.

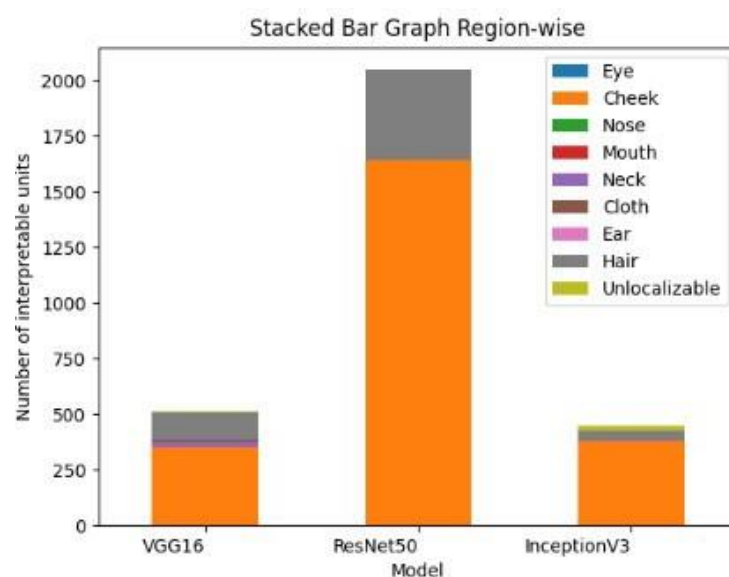


Figure 13. Distribution of Interpretable Units per Facial Region with a Threshold value of 0.4.

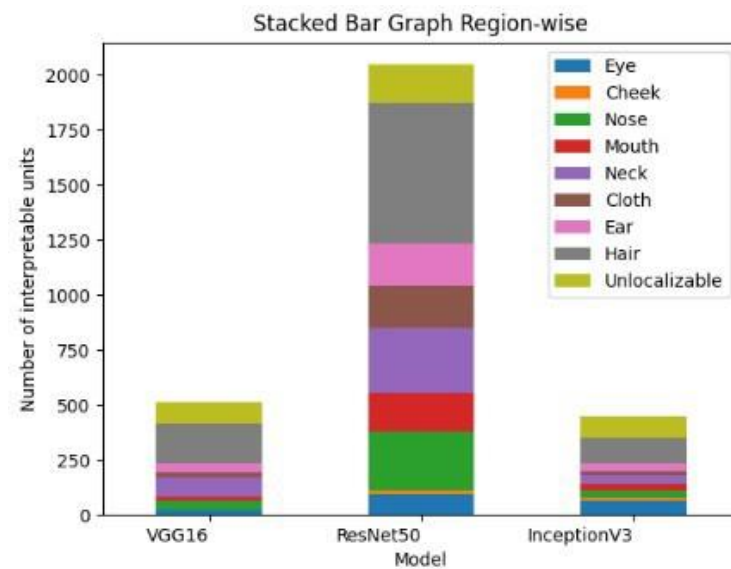


Figure 14. Distribution of Interpretable Units per Facial Region with a Threshold value of 0.04.

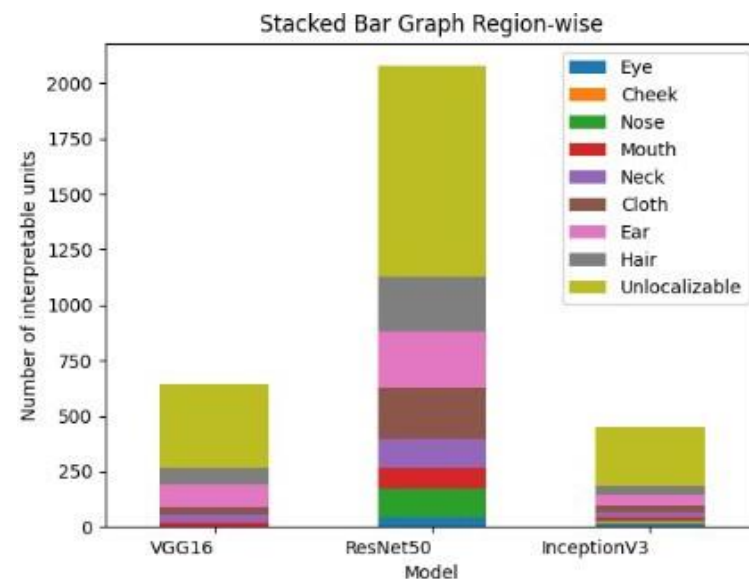


Figure 15. Distribution of Interpretable Units per Facial Region with a Threshold value of 0.005.

Table 3. Details of Interpretations by CNN Models for a threshold value of 0.4.

Models	ResNet-50	VGG-16	InceptionV3
Interpretable units	509/512	2048/2048	423/448
Uninterpretable units	3/512	0	25/448
Top IoU score	0.5918	0.5287	0.7601
Facial concept	skin	skin	skin

Table 4. Details of Interpretations by CNN Models for a threshold value of 0.04.

Models	ResNet-50	VGG-16	InceptionV3
Interpretable units	417/512	1875/2048	350/448
Uninterpretable units	95/512	173/2048	98/448
Top IoU score	0.4235	0.3524	0.4286
Facial concept	hat	hat	cloth

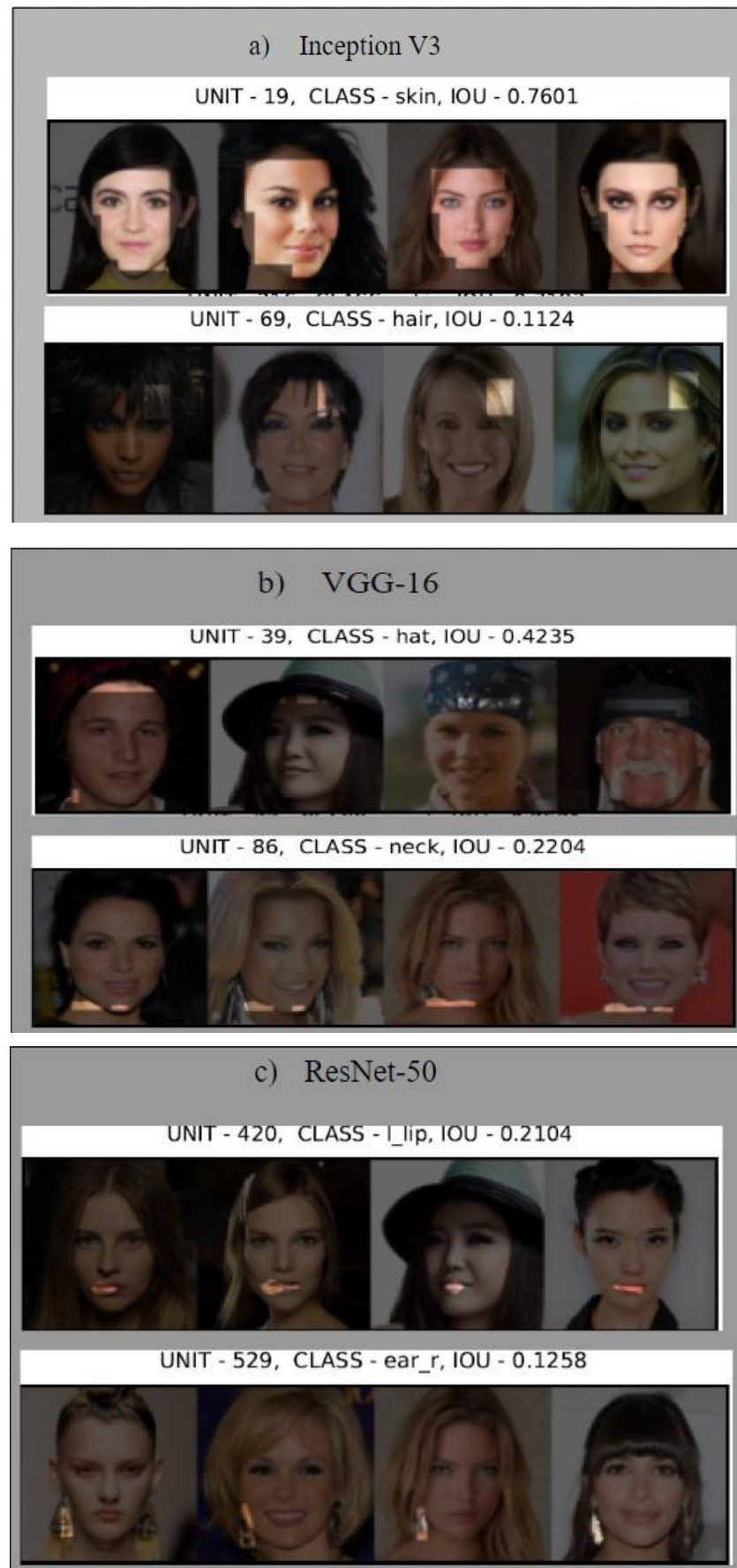


Figure 16. Two examples for each model (a) Inception V3 (b) VGG-16 (c) ResNet-50.

Table 5. Details of Interpretations by CNN Models for a threshold value of 0.005.

Models	ResNet-50	VGG-16	InceptionV3
Interpretable units	133/512	1128/2048	182/448
Uninterpretable units	379/512	950/2048	266/448
Top IoU score	0.3187	0.3510	0.4229
Facial concept	cloth	earring	earring

As stated before, Figure 17 shows samples of two facial concepts that describe the formulation of a probabilistic hierarchy.

**Figure 17.** Probabilistic Hierarchy Results.

Figures 18–20 illustrate the interpretability reports of the dissected models and represent the number of units learned by each model for each facial concept for a threshold value set at 0.04. This representation provides information about the top facial concepts detected by different CNN architectures. All the results are obtained by dissecting the final output layers of the models. Figures 21–23 show the count of interpretable units across convolution layers for the various trained models. The detailed analysis and interpretation of the network dissection results, including the implications of the observed patterns and insights into the internal representations learned by the CNN models, will be further explored and elaborated on in Section 5.

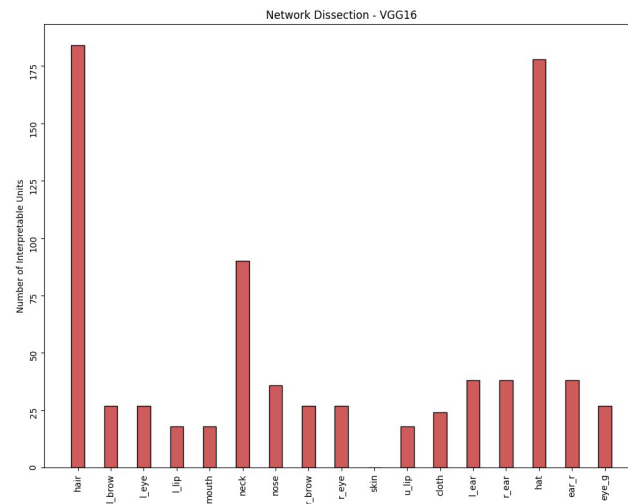


Figure 18. Dissection Histograms of VGG-16.

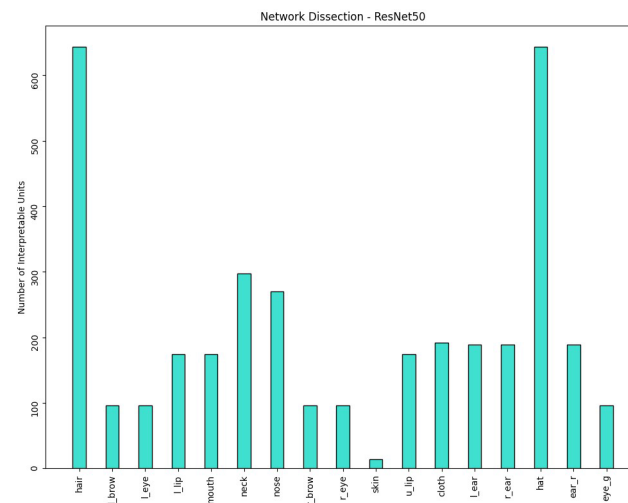


Figure 19. Dissection Histograms of ResNet-50.

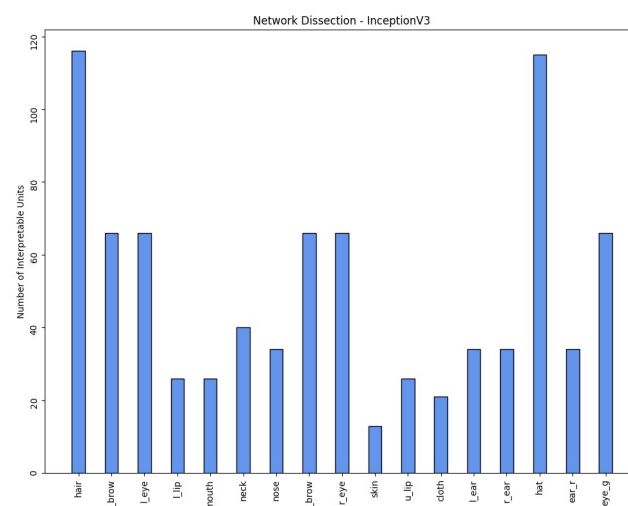


Figure 20. Dissection Histograms of Inception V3.

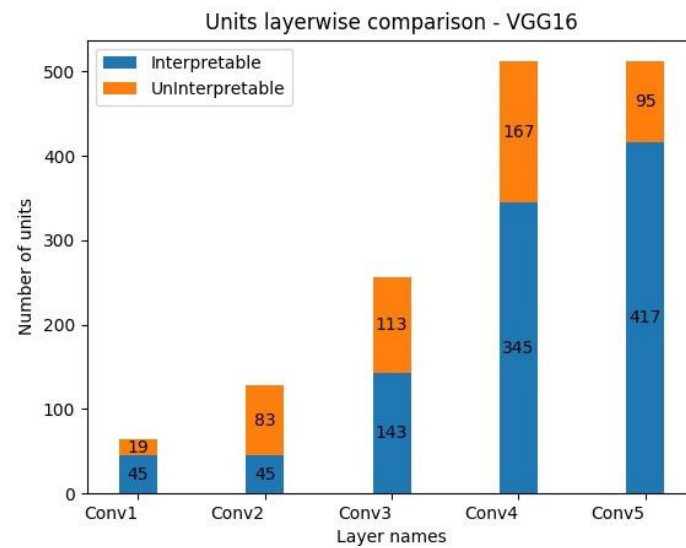


Figure 21. Number of Units Within Multiple Layers for VGG-16.

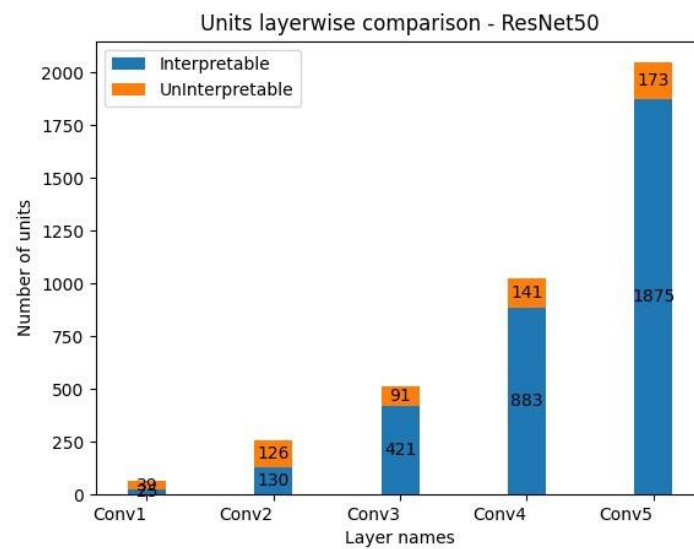


Figure 22. Number of Units Within Multiple Layers for ResNet-50.

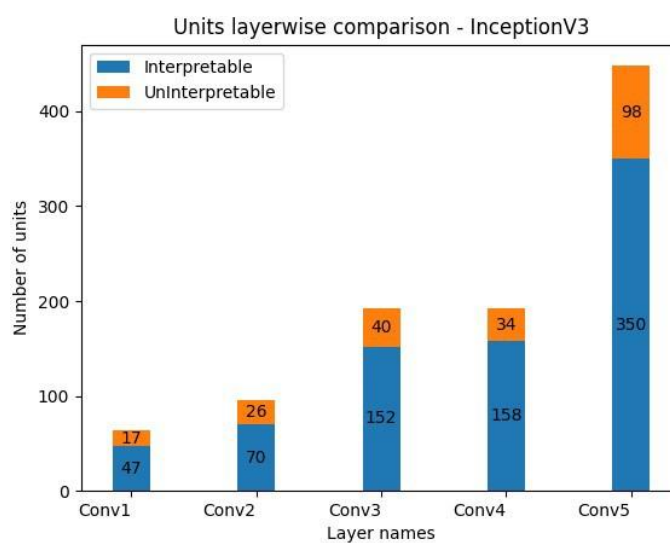


Figure 23. Number of Units Within Multiple Layers for Inception V3.

5. Discussion

This work conducted various studies to understand the internal parts of the developed CNN models by interpreting the facial features learned by these models. It is known that neural networks can learn features through their hidden layers, and they are considered as black box models. In Section 4 (Findings and Evaluation), the outcomes of the explainable AI analysis are presented, which reveal the internal neuronal representation of the three CNN models used for deepfake detection. The data in Section 4 form the basis for the detailed insights discussed here. Specifically, the network dissection results, as summarized in Tables 3–5, provide evidence that the units of the CNN models have learned various facial concepts. These results show a balance between interpretable and uninterpretable units, depending on the threshold value used. However, there are fewer uninterpretable units for a threshold of 0.4, and the types of different facial concepts learned by the models are comparatively fewer. All three models mostly detected the cheek region as the major concept, followed by the hair region. With the decreasing threshold value, e.g., 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 best threshold value as it identifies a variety of facial concepts for all CNN architectures used in the proposed study. It is observed from Figures 13–15 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 the threshold value of 0.4 for all three models, whereas for threshold values of 0.04 and 0.005, the 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 Tables 3–5, 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 a threshold value of 0.04, as given in Figures 20–22, 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 should be avoided due to the model's inability to learn higher-level concepts. This statement can be proven by analyzing the results from Figure 23. From the figure, it is observed that with the initial layers, the proportion of the number of interpretable units is smaller compared to the final layers of the models. As previously stated, applying the IoU approach based on the network dissection algorithm to the images from our deepfake face dictionary makes it difficult to conclude that a concept with an IoU score similar to the top concept is interpretable without conducting a thorough analysis of the activations associated with that concept in relation to other concepts present in the same facial regions. Consequently, we discover that a unit can correspond to several concepts by constructing a hierarchy among the concepts from that specific facial region. This characteristic is explained in Figure 19 and clearly shows how challenging it is for a neuron to differentiate between concepts that appear similar but possess different features.

6. Conclusions

The rise of technology in the multimedia sector has brought about various malicious activities, including the exploitation of fabricated images and videos in areas such as entertainment and politics. This advancement has blurred the line between manipulated and authentic visuals, prompting extensive research in this field. The term 'deepfake' is now widely recognized for its ability to create highly realistic imagery. To deal with the latest manipulation methods in visual media, an explainable deepfake detection approach is critical. This research focuses on developing a deepfake detection method that delivers an

explainable model. The proposed research study involves the detection and interpretation of deepfake detection models using the network dissection algorithm. This is implemented in two stages. (1) To detect forged images, three cutting-edge CNNs (VGG-16, Inception V3, and ResNet-50) are used, and a comparative analysis of these models is performed. (2) A network dissection algorithm is employed 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 compulsory 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. The experiments demonstrated that the CNN models employed in this study performed well, with F1-scores ranging from 0.8 to 0.9. The explainability part of the proposed study is based on a network dissection algorithm, which involves generating interpretability reports for the three deepfake detection models that have been trained on a well-known deepfake dataset. This also offers a comprehensive discussion on how model interpretability can be leveraged to gain a deeper understanding of the model mechanisms and the representation of information or features learned by these models.

Although there are several deepfake detection tools and models available to detect manipulated videos or images, existing research lacks explanations or reasoning behind the classification of images when using the network dissection algorithm. This study aims to shed light on how these models distinguish between real and altered images by analyzing their internal neural representations. Additionally, while this study focuses on CNN architectures, it is important to note that the network dissection algorithm is specifically a strong interpretability tool for CNNs, as it is inherently designed to map internal CNN components to concepts comprehensible to humans. However, it is not directly applicable to architectures such as vision transformers (ViTs) or pure transformer models, which operate on fundamentally different principles. Nevertheless, the general principle of associating model components with concepts comprehensible to humans could inspire the development of similar techniques adapted for non-CNN architectures. As a future direction, adapting and extending the core ideas of the network dissection algorithm to these advanced architectures could provide deeper insights into their decision-making processes and enhance interpretability in tasks such as deepfake detection. This would bridge the gap in interpretability techniques across different model families, ensuring that the rapid advancements in model architectures are matched by equally sophisticated explainability methods.

In conclusion, this study underscores the importance of interpretability in deep neural networks, providing insights into their hierarchical structures. By addressing the critical challenge of explainability in deepfake detection, this work not only improves the reliability of these systems but also lays the foundation for extending interpretability methods to more advanced architectures in future research.

While the proposed method effectively enhances the interpretability of CNN-based deepfake detection models, certain limitations remain. The interpretability of the network dissection results is constrained by the limited range of concepts in the deepfake face dictionary, which could be expanded in future work to include additional facial features and more diverse images. Further, the efficiency of models could be enhanced by fine-tuning neuron activations, and incorporating other XAI techniques could provide complementary insights. Additionally, adapting the method to more advanced architectures, such as vision transformers and hybrid models, would extend its applicability, paving the way for more robust explainable AI frameworks.

Author Contributions: Conceptualization, N.M. and A.I.I.; methodology, N.M.; software, N.M.; validation, N.M. and A.I.I.; formal analysis, N.M. and A.I.I.; investigation, N.M.; resources, N.M. and A.I.I.; data curation, N.M.; writing—original draft preparation, N.M.; writing—review and editing, N.M. and A.I.I.; visualization, N.M.; supervision, A.I.I.; project administration, A.I.I.; funding acquisition, A.I.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This dataset is released under the Terms to Use Celeb-DF, which is provided “as it is” and we are not responsible for any subsequence from using this dataset. All original videos of the Celeb-DF dataset are obtained from the Internet which are not property of the authors or the authors’ affiliated institutions. Neither the authors or the authors’ affiliated institution are responsible for the content nor the meaning of these videos. If you feel uncomfortable about your identity shown in this dataset, please contact us and we will remove corresponding information from the dataset. Available online: <https://github.com/yuezunli/celeb-deepfakeforensics> (accessed on 5 December 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ignatova, D. Affirming wellness culture through innovative methodology related to Blaze-pod trainer system. *Bulg. Educ. J.* **2023**, *31*, 212–225. [\[CrossRef\]](#)
2. Ignatova, D. Specificity of the motor potential for achieving Scholar Wellness. *Trakia J. Sci.* **2021**, *19*, 867–873.
3. Ignatova, D. The effects of swimming on preschool children with spinal abnormalities. In Proceedings of the 17th International Balkan Society for Pedagogy and Education/BASOPED/Conference, Varna, Bulgaria, 24–26 September 2015; pp. 207–212, ISBN 978-954-326-370-7.
4. Ignatova, D. Motor activity based on learning—Contemporary trends in School Wellness. *Smart Innov. Recreat. Wellness Ind. Niche Tour. Sci. J.* **2023**, *5*, 22–26.
5. Rong, I. Detection and Segmentation of Deepfake Face Images Generated by GANs Using Segmented Based CNN. Master’s Thesis, Dublin Business School, Dublin, Ireland, 25 August 2020.
6. Mansoor, N.; Iliev, A. Artificial Intelligence in Forensic Science. In *Advances in Information and Communication. FICC 2023. Lecture Notes in Networks and Systems*; Arai, K., Ed.; Springer: New York City, NY, USA, 2023; Volume 652.
7. Mansoor, N.; Iliev, A. Artificial Intelligence in Forensic Science. In *Intelligent Computing. SAI 2024. Lecture Notes in Networks and Systems*; Arai, K., Ed.; Springer: New York City, NY, USA, 2023; Volume 1018.
8. 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. [\[CrossRef\]](#)
9. Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HA, USA, 21 July 2017; pp. 3319–3327.
10. Bau, D.; Zhou, B.; Oliva, A.; Torralba, A. Interpreting Deep Visual Representations via Network Dissection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2131–2145.
11. Teotia, D.; Lapedriza, A.; Ostadabbas, S. Interpreting Face Inference Models Using Hierarchical Network Dissection. *Int. J. Comput. Vis.* **2022**, *130*, 1277–1292. [\[CrossRef\]](#)
12. Daglarli, E. Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In *Advances and Applications in Deep Learning*; Aceves-Fernandez, M.A., Ed.; InTechOpen: London, UK, 2020.
13. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [\[CrossRef\]](#)
14. Molnar, C. *Interpretable Machine Learning*, 2nd ed.; Leanpub: Victoria, BC, Canada, 2023.
15. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI Methods—A Brief Overview. In *xxAI—Beyond Explainable AI. Lecture Notes in Computer Science*; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W., Eds.; Springer: New York City, NY, USA, 2020.

16. Bento, V.; Kohler, M.; Diaz, P.; Mendoza, L. Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches. *J. Discov. Artif. Intell.* **2021**, *1*, 9. [[CrossRef](#)]
17. Pino, S.; Carman, M.; Bestagini, P. What's wrong with this video? Comparing Explainers for Deepfake Detection. *arXiv* **2021**, arXiv:2105.05902.
18. Baldassarre, F.; Debar, Q.; Pontiveros, G.F.; Wijaya, T.K. Quantitative Metrics for Evaluating Explanations of Video DeepFake Detectors. In Proceedings of the British Machine Vision Conference, London, UK, 21–24 November 2022.
19. Dong, S.; Wang, J.; Liang, J.; Fan, H. Explaining Deepfake Detection by Analysing Image Matching. In *xComputer Vision—ECCV 2022. Lecture Notes in Computer Science*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: New York City, NY, USA, 2022.
20. Narteni, S.; Orani, V.; Ferrari, E.; Verda, D.; Cambiaso, E.; Mongelli, M. A New XAI-based Evaluation of Generative Adversarial Networks for IMU Data Augmentation. In Proceedings of the IEEE International Conference on E-Health Networking, Application Services, Genoa, Italy, 17 October 2022; pp. 167–172.
21. Trinh, L.; Tsang, M.; Rambhatla, S.; Liu, Y. Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3 January 2021; pp. 1972–1982.
22. Abir, W.H.; Khanam, F.R.; Alam, K.N.; Hadjouni, M.; Elmannai, H.; Bourouis, S.; Dey, R.; Khan, M.M. Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods. *Intell. Automat. Soft Comput.* **2023**, *35*, 2151–2169. [[CrossRef](#)]
23. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online, 14–19 June 2020.
24. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.