Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Deepfake Image Detection Using Deep Learning Approach: A survey

Bushra Tariq Abdul Hafez [1]*, Farah Abbas Obaid[2]

[1]University of Kufa, College of Computer Science Mathematics/Department of Computer Science, Iraq.
bushrat.alsaalim@student.uokufa.edu.iq
[2]University of Kufa, College of Computer Science Mathematics/Department of Computer Science, Iraq.
faraha.altaee@uokufa.edu.iq
*Corresponding author E-mail: bushrat.alsaalim@student.uokufa.edu.iq

https://doi.org/10.46649/fjiece.v4.1.11a.25.3.2025

Abstract *The rapid advancements in artificial intelligence (AI) have brought forth deepfake technologies, leveraging sophisticated deep learning algorithms to generate highly realistic yet deceptive media. This poses a substantial threat to individuals' integrity, privacy, and security and can lead to widespread social and political instability. In response, there is an imperative necessity to create advanced computer models capable of efficiently identifying counterfeit content in real-time and notifying consumers of potential manipulations. This paper presents a comprehensive examination of recent studies on deepfake detection utilizing deep learning techniques. This paper aims to advance the forefront by systematically categorizing the diverse techniques employed for identifying counterfeit content. Furthermore, we outline the merits and drawbacks of each approach and propose many directions for future research to address the persistent challenges and shortcomings in deepfake content identification.*

Keywords: *Deepfake Detection; Convolutional Neural Networks (CNN); Image Manipulation; Generative Adversarial Networks (GANs); Transfer learning; AXI explanation.*

1. INTRODUCTION

The fast development of AI in the past few years has resulted in Deepfake technology, which uses deep learning algorithms to generate hyper-realistic, but ultimately—has been warned about for many images that seem indistinguishable from genuine ones. This phenomenon has elicited global anxiety, especially about the reliability and accuracy of digital content in sectors, e.g., media, cybersecurity, and legal systems [1]. Originally understood as tools for entertainment and creative manipulation, deepfakes have since evolved into a major threat, used to spread misinformation, commit identity theft and tilt public opinion. The emergence of deepfakes has developed alongside a rise in cybercrimes involving manipulated images, as shown by a staggering 67% growth in reported cases during the last three years [2]. Hence, the need for reliable detection systems has never been more important, especially to protect digital pieces of evidence from tampering and misuse[1]. Current detection methods largely rely on convolutional neural networks (CNNs) known for being efficient in image processing and capable of detecting subtle artefacts indicating tampering. Advancements in methods of creating deepfakes make it tremendously more challenging for traditional detection algorithms to deal with the complexities of modern-day manipulations. This has propelled the development of advanced deep-learning models capable of understanding images with greater accuracy and complexity [3]. In this paper, we focus on a comprehensive review of state-of-the-art techniques for detecting and classifying deepfake images. Focus is placed on evaluating the

effectiveness of such strategies together with an acknowledgement of their limitations, and a systematic framework is proposed in order to study alternative combinations according to different performance measures and evaluation criteria. This study explains the strengths and limitations of existing detection algorithms, emphasizing essential domains that necessitate additional investigation and innovation to address the increasing threats posed by deepfake technology [3].

Table 1: The top five datasets used in deep fake image

No.	Dataset Name	Description
1	FFHQ (Flickr-Faces-HQ)	70,000 high-quality real human faces; often used to generate fake faces with GAN models like StyleGAN.
2	100K-FACE	100,000 synthetic face images produced utilizing StyleGAN.
3	DFFD (Diverse Fake Face Dataset)	Combines images from other datasets like FFHQ and CelebA, consisting of both real and manipulated faces.
4	CASIA-WebFace	Over 500,000 real-face images from around 10,000 subjects.
5	CelebA (included in DFFD)	More than 200,000 real celebrity face images.
6	140k Real and Fake Faces	Consists of 140,000 images: 70,000 real and 70,000 generated using StyleGAN.

2. Literature Review

The Section examines existing literature involving deepfake technology.

Chia-Yen Lee et al. (2018) They proposed a deep forgery discriminator (DeepFD) to detect and identify computer generated image efficiently. To overcome this challenge, they employed contrastive loss to learn the unique features from synthesized images created by various Generative Adversarial Networks (GAN). By adding a classifier to improve detection performance, for the proposed DeepFD model, 94.7% of GAN images generated by various mainstream GANs were successfully identified. The test results demonstrate how well the model works in identifying real images and fake images [4] .

An investigation of a approach to identify deepfake facial images using deep ensemble neural network techniques along with transfer learning has been carried out by [104] Jannatul Mawa and Md. Humayun Kabir (2024). They used transfer learning and a weighted average ensemble technique to detect human fake faces by using three different pre-trained architectures (ResNet50, DenseNet201, InceptionV3) and "Real and Fake Face Detection" dataset. In the end, they obtained 64.71 accuracy[5].

Bogdan Ghita et al. [18] developed a deepfake detection approach using a Vision Transformer (VIT) model trained and tested on a composite dataset of real images and deepfakes collected from Kaggle of 40,000 images. This result indicates the VIT model got a high score, 89.9125%[6].

Suganthi ST and et al (2022) , They proposed a deep learning-based deepfake face image detection method via Fisherface + local binary pattern Histogram(FF-LBPH) The Fisherface algorithm initialises dimensionality reduction in the facial feature space using LBPH for facial recognition, and applies a Deep Belief Network (DBN) with Restricted Boltzmann Machine (RBM) in detecting the Images. The dataset

used in this study contains datasets like FFHQ, 100K-Faces, DFFD, and CASIA-WebFace. The accuracy rate on CASIA-WebFace dataset and DFFD dataset with the proposed FF-LBPH-DBN method was found to be 98.82% and 97.82% respectively[7].

Peng Zhou and ***et al*** (2018) , This method used a two-stream method to detect face tampering based on the training of GoogLeNet and patch-based triplet networks, where one stream of the GoogLeNet detects the tampering artifacts in a face classification stream, while the patch-based triplet network captures the local noise residuals and camera characteristics for the second stream. Moreover, they applied two different types of online face-swapping technologies to create a new dataset with 2010 modified images, each with update face: SwapMe and FaceSwap Dataset They used SwapMe + FaceSwap train So the outcome was SwapMe test and FaceSwap test (0.995, 0.999) respectively [8].

Iszuanie Syafidza Che Ilias and ***et al*** (2024) They took a look at how artificial intelligence could be used to detect deepfake pictures. They focused on three classes of convolutional neural network (CNN) algorithms, which are ResNet, VGG16, and VGG19. A dataset of 1,200 photos (both fake and real) was used to evaluate the accuracy of these CNN models. Those deepfake images were created using FaceApp, a popular image-altering app. In this regard, our findings show that VGG19 outshines both VGG16 and ResNet50, and accuracy rate of 98%[9].

Ananda Adhicitta Wangsadidjaja (2023) ,They created a tool to detect deepfakes using a Convolutional Neural Network, namely the ResNet-50 model that uses hoax data generating from ProGAN model. This approach achieved an accuracy of 85%, precision of 100% and recall of 65% in detecting these pictures. But original StyleGAN and BigGAN deepfakes were not as effective on the model [10].

Wahidul Hasan Abir et al. Deep learning is the most utilized methodology for large-scale exploration of generative media such as deepfakes—(2023)—Advanced algorithms' proficiency in detecting deepfakes, and their evaluation methods via Local Interpretable Model-Agnostic Explanations (LIME) are extensively discussed in this article. This paper demonstrated style transfer from both sources: a dataset of 70,000 real images from the Flickr dataset (Nvidia authors) and 70,000 synthetic images created with StyleGAN at a resolution of 256 pixels. In an effort to achieve a high degree of accuracy, various Convolutional Neural Network (CNN) models (i.e., InceptionResNetV2, DenseNet201, InceptionV3, ResNet152V2) were employed. The LIME approach was applied to explain the regions of the image that affect the model's classification decisions. InceptionResNetV2 had the best accuracy of 99.87% of the models, followed by DenseNet201, InceptionV3, and ResNet152V2 with 99.81%, 99.68%, and 99.19%, respectively. The LIME technique further corroborated these findings, increasing the interpretability of models for explainable artificial intelligence (XAI)[11].

Majed M. Alwateer (2024) , In a novel pipeline involving three key components, A methodology for the production and categorization of explainable deepfake images The first part, called Instant ID, creates deepfake pictures from actual photos. The second module is Xception that classifies the image into real or deepfake. The third part is used for interpretability is a Local Interpretable Model-Agnostic Explanations (LIME). With the ImageNet dataset, the new model showed unparalleled results of 100% in both F1 score and accuracy. In comparison, the VGG16 and CNN models had an F1 score and accuracy of 94%, and the Multimodal Network had an F1 score and accuracy of 61%[12] .

It is important to summarize previous works, focusing on the most important keywords in them, as in Table 2 .

Table 2 Summarizing the key information from each study

No.	Reference	Year	Proposed Approach	Dataset	Accuracy (%)
1	Chia-Yen Lee et al. [4]	2018	Deep Forgery Discriminator (DeepFD)	Fake images produced by several GANs	94.7%
2	Jannatul Mawa and Md. Humayun Kabir [5][6]	2024	ResNet50, Dense201, and InceptionV3	Real and Fake Face Detection dataset	64.71%
3	Bogdan Ghita et al. [6]	2024	Vision Transformer (ViT)-based deepfake detection technique	Dataset from Kaggle (40,000 images)	89.91%
4	Suganthi ST et al. [7]	2022	Fisherface and Local Binary Pattern Histogram (FB-LBPH)	FFHQ, 100K-Face, DFFD, CASIA-WebFace	98.82% (CASIA-WebFace), 97.82% (DFFD)
5	Peng Zhou et al. [8]	2018	Dual-stream network for facial tampering detection using GoogLeNet	SwapMe and FaceSwap	99.5% (SwapMe), 99.9% (FaceSwap)
6	Iszuanie Syafidza Che Ilias et al. [9]	2024	VGG16, VGG19, and ResNet	Deepfake images generated using FaceApp	98%
7	Ananda Adhicitta Wangsadidjaja [10]	2023	Convolutional Neural Network (ResNet-50)	AI-generated photos (ProGAN)	Accuracy: 85% Precision: 100% Recall: 65%
8	Wahidul Hasan Abir et al. [11]	2023	Local Interpretable Model-Agnostic Explanations (LIME) for various CNN models	Flickr dataset	InceptionV3: 99.68%, ResNet152V2: 99.19%, DenseNet201: 99.81%, InceptionResNetV2: 99.87%

9	Majed M. Alwateer [12]	2024	Instant ID for creating deepfakes, Xception for classification, LIME for explainability	ImageNet dataset	Proposed Model: F1 Score 100%, Accuracy: 100% VGG16 and CNN: F1 Score 94%, Accuracy 94% Multimodal Network: F1 Score 61%, Accuracy 61%

Footnotes:
1. FFHQ: Flickr-Faces-HQ dataset, a collection of high-resolution facial images often used for training GANs.
2. DFFD: Diverse Fake Face Dataset, combining real and manipulated faces from multiple datasets.
3. GANs: Generative Adversarial Networks, a deep learning technique for generating synthetic images.
4. ProGAN: Progressive Growing of GANs, a model for generating high-quality synthetic images.
5. SwapMe and FaceSwap: Specialized datasets containing face-swapped images for tampering detection.
6. AUC: Area Under the Curve, a metric for evaluating the performance of classification models.

3. DEEPFAKE GENERATION

DNNs are made up of a collection of linked components known as neurons. Collectively, these units carry out computational tasks and aid in the resolution of challenging issues. The technologies commonly associated with the creation of deepfakes include the GAN architecture and the autoencoder-decoder model [13].

3.1. Autoencoders

Autoencoders are important for generating new data by learning how to compress and reconstruct data that passes through, and they are used widely in deep fake technology where visual features like faces are important. This system consists of two parts: an encoder that compresses the input images into a small latent representation, and a decoder that reconstructs the original image from this compressed representation. They detect essential features/expressions on the face from deepfakes and realistic alteration/exchange of the face in images and video[14].

3.1.1. How Autoencoders Work in Deepfake Generation:

1. **Facial Encoding**: Autoencoders are designed to capture and represent the distinctive facial structures and expressions of individuals by transforming the image into a compressed low-dimensional latent space.
2. **Reconstruction and Face Manipulation**: Following the encoding process, the decoder utilises the latent representation to reconstruct the image. The process of creating deepfakes can involve modifications to the face, including adjustments to expressions or the exchange of faces between individuals.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

3. **Training**: The effectiveness of autoencoders in generating realistic deepfakes is contingent upon the quantity and quality of the training data available. Extensive datasets enable autoencoders to enhance their generalization capabilities, yielding more credible outcomes [13]

3.2. *Generative Adversarial Networks (GANs)*

GAN is used in most of the deepfake technologies available today. In 2014, Ian Goodfellow made the initial proposal for the GAN architecture [15]. He presented a framework that makes use of two neural networks that compete with one another to produce new data. One of the neural networks creates the new data, while the other separates it from the initial training set. Contesting the two neural networks tends to enhance the neural network's discriminating capacity as well as the quality of fake data generated . GAN may produce a unique image on its own if a lot of images are provided to it [16]. Attaching a filter, however, is essential to assist determine whether or not these distinct outputs are appropriate. GANs use a discriminative network to do this, comparing the generated data with actual data. Both are taught to collaborate until the resultant output is mistakenly labeled as authentic, which happens around 50% of the time. This supports our conclusion that the generator model is producing credible examples with success. The schematic representation illustrating the GAN architecture's workflow is displayed in Fig. 1 . The generator and discriminator of the GAN architecture are trained using the min-max technique [15]. A fake output is epresented by the min(0), but a real output is represented by the max (1). The discriminator's objective is to approach the maximum value as closely as possible in order to produce a deepfake that seems realistic and can be used to swap faces in pictures and movies. For creating fresh data, GANs are better suited [17]. The primary benefit of GANs over autoencoders is their greater versatility, as evidenced by their ability to generate many data classes that resemble the MNIST dataset [18]. Conversely, autoencoders work better when compressing data to smaller dimensions or producing semantic

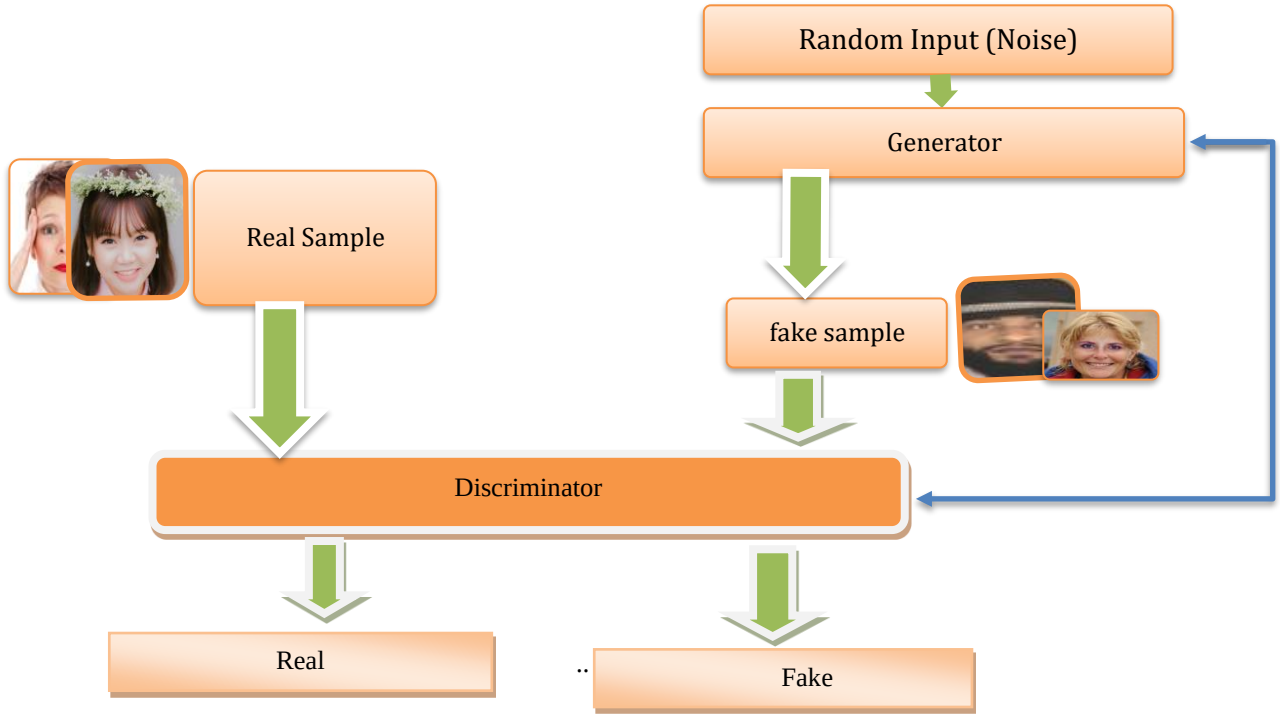

Figure 1. Basic GAN architecture

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

4. Types of manipulation in deep fake faces

As Figure 2 illustrates, There exist five primary categories of deepfake manipulation. A type of manipulation known as face synthesis involves creating images of non-existent human faces [19]. In attribute modification [20] , the only area that is changed is the one that is related to the characteristic. This technique enables various modifications, such as altering skin tone, adding or removing eyeglasses, and even more dramatic changes like adjusting age and gender. However, this study emphasizes manipulations that are predominantly found in video formats, as these tend to generate higher levels of engagement compared to image-based content. The dynamic characteristics of videos enhance the viewer's experience and interaction, making video manipulations particularly significant in the context of deepfake technology.

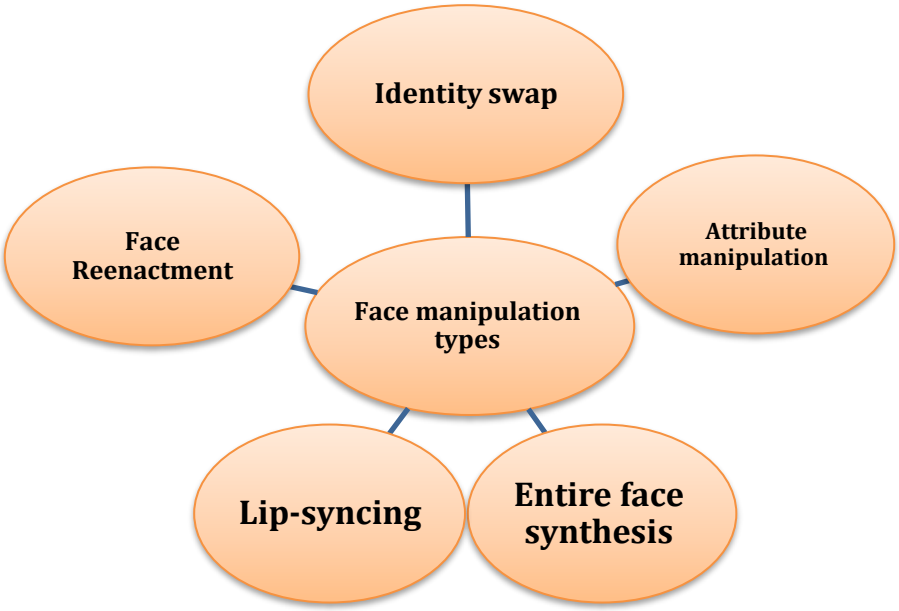

Figure 2. The five principal categories of types of deepfake manipulation.

Facial manipulation methods can be categorized into several distinct types based on the techniques used to alter or synthesize facial features. Below are the primary methods commonly employed:

4.1. Entire Face Synthesis:

This method involves generating an entirely new face using generative models, such as Generative Adversarial Networks (GANs). The generated face is not based on any real individual, and both the facial features and the identity are artificial [21].
Examples: ThisPersonDoesNotExist.com uses GANs to generate synthetic faces.

4.2. Identity Swap (Face Swapping):

In identity swapping, the facial identity of one person is replaced with another person's face while retaining the original facial expressions and pose. This technique is often used in deepfakes.
Tools and datasets: Popular datasets such as Celeb-DF provide real and manipulated images where face swapping has been applied [22].

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

4.3. Attribute Manipulation:

Attribute manipulation entails modifying specific facial attributes—such as age, gender, hairstyle, or facial expressions—while leaving the remainder of the face intact. This targeted approach allows for nuanced changes that can significantly alter a person's appearance without affecting the overall structure of the face. Such manipulations can create realistic alterations, making them particularly useful in applications such as entertainment, advertising, and virtual environments. This method can use apps like FaceApp or models that focus on facial features' transformation [23].
Example: Changing someone's age or adding/removing facial hair using FaceApp.

4.4. *Expression Swap:*

This method changes the facial expressions of a person without altering their identity. Techniques in this category manipulate the muscles and facial landmarks to change expressions like smiling, frowning, or surprise. Datasets: Face Forensics++ includes expression swaps, where the expression of a person is changed artificially while keeping the identity intact [24], as shown in figure .

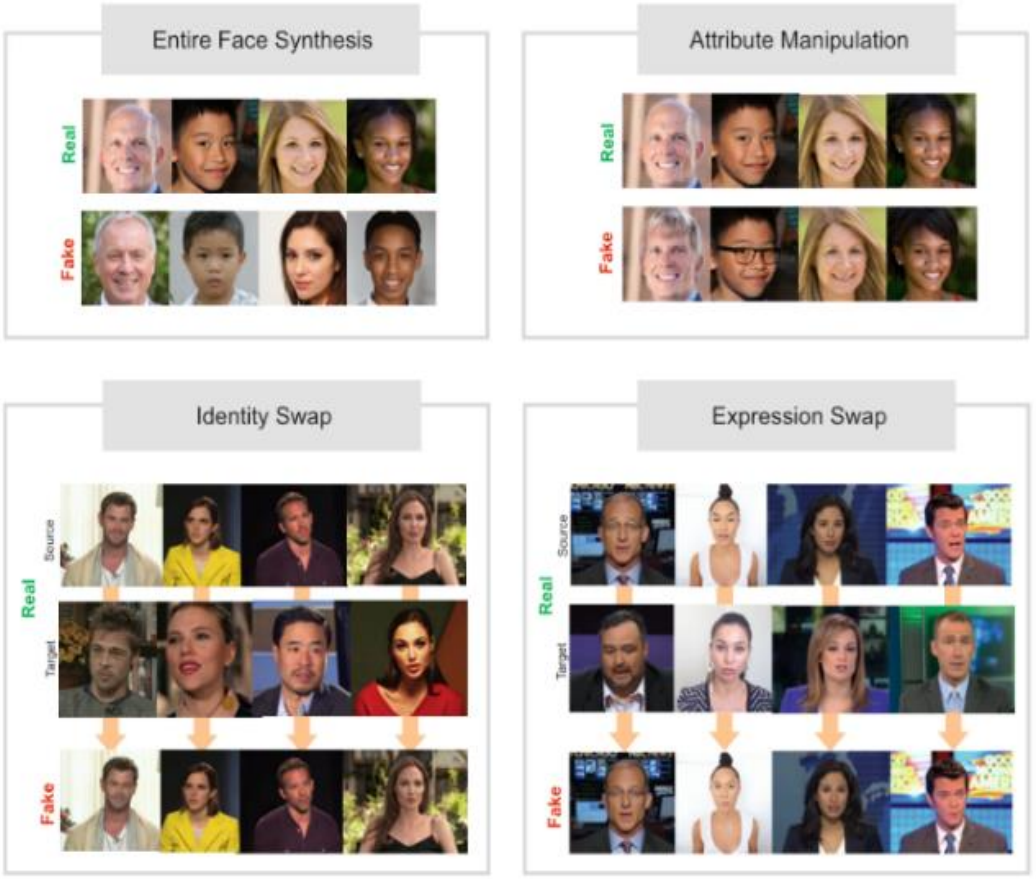

Figure 3. presents examples of real and manipulated facial images across four categories of facial manipulation [25] .

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

5. Deepfake Detection

In the field of deepfake detection, various approaches have been developed to identify manipulated content. Detection systems leverage certain inconsistencies and traces of manipulation left during the creation of deepfakes to classify content as either fake or authentic. As deepfake techniques expand to include images, videos, and audio content, it is essential to develop effective detection methods for these manipulations. Detection approaches are generally categorized into three main types: image/video-based detection, audio spoofing detection, and multimodal approaches [26] .

5.1. Deepfake Detection Images

Several techniques have been developed to leverage deep networks for the identification of images generated by Convolutional Neural Networks (CNNs). These neural network-based approaches enhance the detection of fake images, particularly those of faces, by employing a deep convolutional neural network that utilizes pre-processing techniques to evaluate the statistical properties of the images.Initially, the model extracts facial features using face recognition networks within a deep learning framework. A fine-tuning process is then applied to optimize these facial features for distinguishing between real and fake images. Preliminary results from validation datasets indicate promising outcomes with these methodologies [27] . However, a significant limitation in prior studies is the neglect of the forensics model's generalizability; they often train and test their models on the same type of dataset, which may hinder the robustness of the detection.To address this issue, a novel forensic convolutional neural network (CNN) is proposed, incorporating Gaussian Blur and Gaussian Noise as image pre-processing techniques to enhance the identification of fake human images. This model is designed to amplify high-frequency pixel noise while mitigating low-level pixel statistics, thereby ignoring low-level high-frequency artifacts typical in CNN-generated images. Consequently, the forensic classifier becomes more adept at differentiating between real and fake faces by learning more salient features associated with both image categories. Experimental results demonstrate the model's efficacy in detecting counterfeit images. Furthermore, to improve fake photo detection, a hybrid technique has been introduced alongside traditional deepfake detection models. For example, a two-stream network has been proposed to identify face tampering. In this architecture, GoogleNet employs a face categorization stream to train the model using both manipulated and real images [28]. Additionally, A patch triplet stream records low-level camera attributes and local noise residuals, which are then utilized to extract features through a dedicated feature extractor.

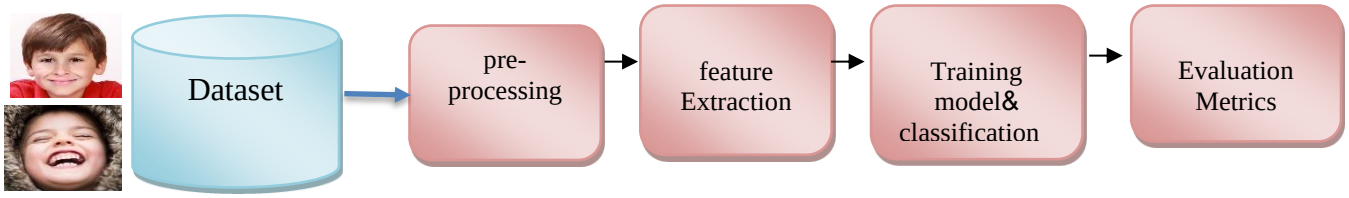

Figure 4. The steps of the deep fake detection Image

6. Using techniques of Transfer Learning Neural Network Techniques

This section looks at the neural network approaches that are used with transfer learning. Using a technique called transfer learning [4], we create models for the prediction process that are already learned. High prediction performance can be achieved by applying the features that have been learned through transfer learning. Using a pre-trained network, the transfer learning-based fine-tuning approach retrains a portion of the network using the new dataset. This section examines the operation of transfer learning methods used in deepfake detection. Neural network techniques are analysed architecturally and the configuration parameters are established.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

6.1. Xception Technique

Xception is a neural network based on transfer learning used mainly for image recognition tâches. Xception means "Extreme Inception". The Xception model is an extension of the Inception architecture, which is distinctly both an efficient and low-speed design on depthwise separable convolution layers. The Xception model architecture employs depth-wise separable convolution layers. The Xception model has the smallest weight serialization. When we talk about the architecture of the Xception model, it is compiled on 36 convolutional layers, which helps it to learn effective features[29].

6.2. NAS-Net Technique

Neural Search Architecture (NAS) Network is referred to as the NAS-Net. The NAS-Net [29] is a model from the convolutional neural network family that is based on transfer learning. The Google Brain analyzes the NAS-Net . The ImageNet database [30], which contains more than a million photos, is used to train the NAS-Net model. The calculation costs of the model are lower. The NAS-Net architecture's blocks are searched using the reinforcement learning search approach . In order to detect deepfakes, our research study substituted MLP blocks for the top layers of NAS-Net.

6.3. Mobile Net Technique

In our study, Mobile Net , we have created a transfer learning model for deepfake detection[31]. Google made the Mobile Net model open-source. Fast processing applications related to computer vision heavily depend on **MobileNet**.. The model architecture is simpler and requires less processing. Depthwise separable convolutions are used in the architecture's construction [32]. In depthwise separable convolutions, the two operations—depthwise and pointwise—are carried out. When compared to standard convolutions, this results in fewer parameters. In our research project, MLP blocks were used in place of the top layers of Mobile Net.

6.4. VGG16 Technique

The VGG16 is one of the most commonly used pre-trained neural networks for image recognition tasks[33]. The VGG16 model was proposed by K. Simonyan and A. The VGG16 architecture is based on CNN. The VGG16 Model architecture was introduced in the year 2014 in the ILSVR competition. Based on our dataset, we implemented the VGG16 model to detect deepfake videos. In our research work, we use a multi-layer perceptron (MLP) [34] block replaces the top layers of VGG16.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

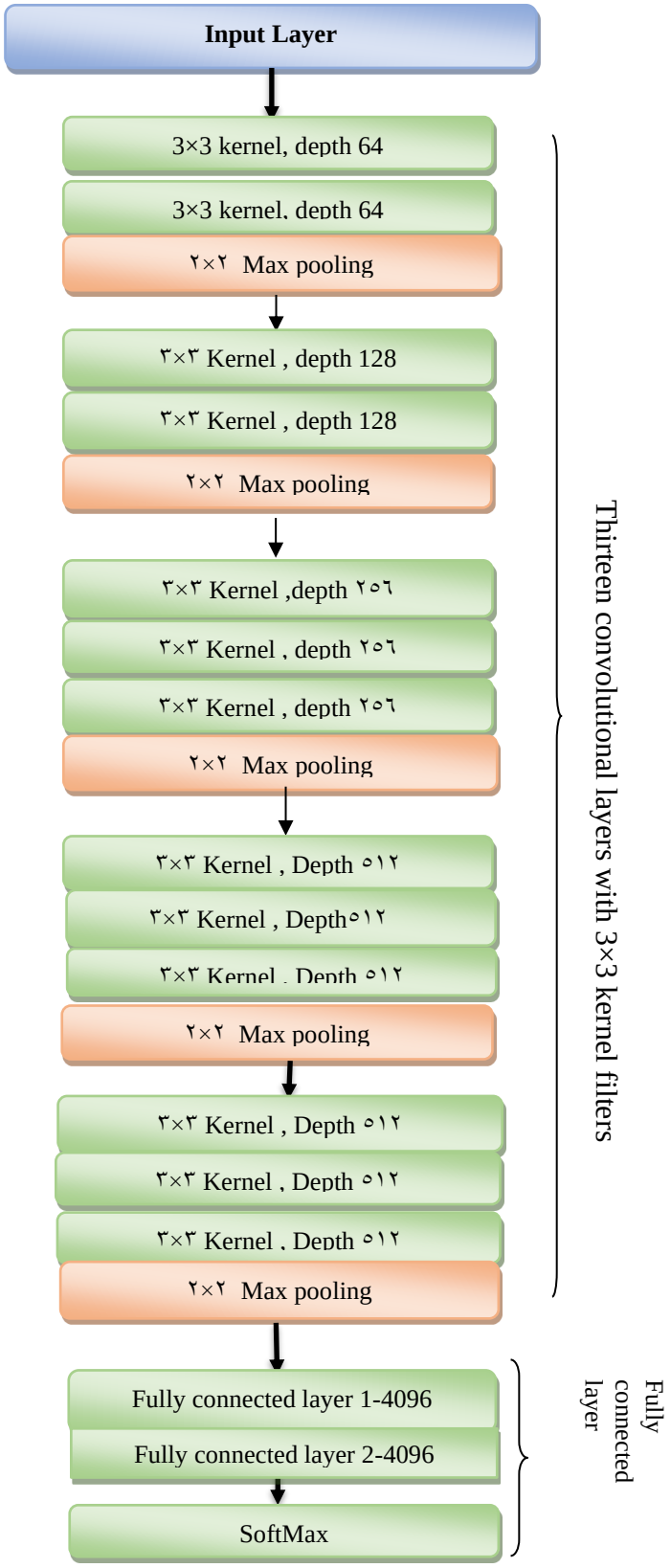

Figure 5: VGG-16 Architecture of a VGG16 mode

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

6.5. VGG19

Then VGG19 which is the convolutional neural network model is better than the first one which is only the one convolutional layer because the VGG19 have many convolutional layers and nonlinear activation layers. Because of the layer structure, rectified linear unit (ReLU) which means the largest value in the picture region will be the pooled value of that area, can be used as activation function and for down sample through max- pooling. The role of the down sampling layer is to reduce the number of parameters, retain the main properties of the sample, and improve the network's ability to resist picture distortion.

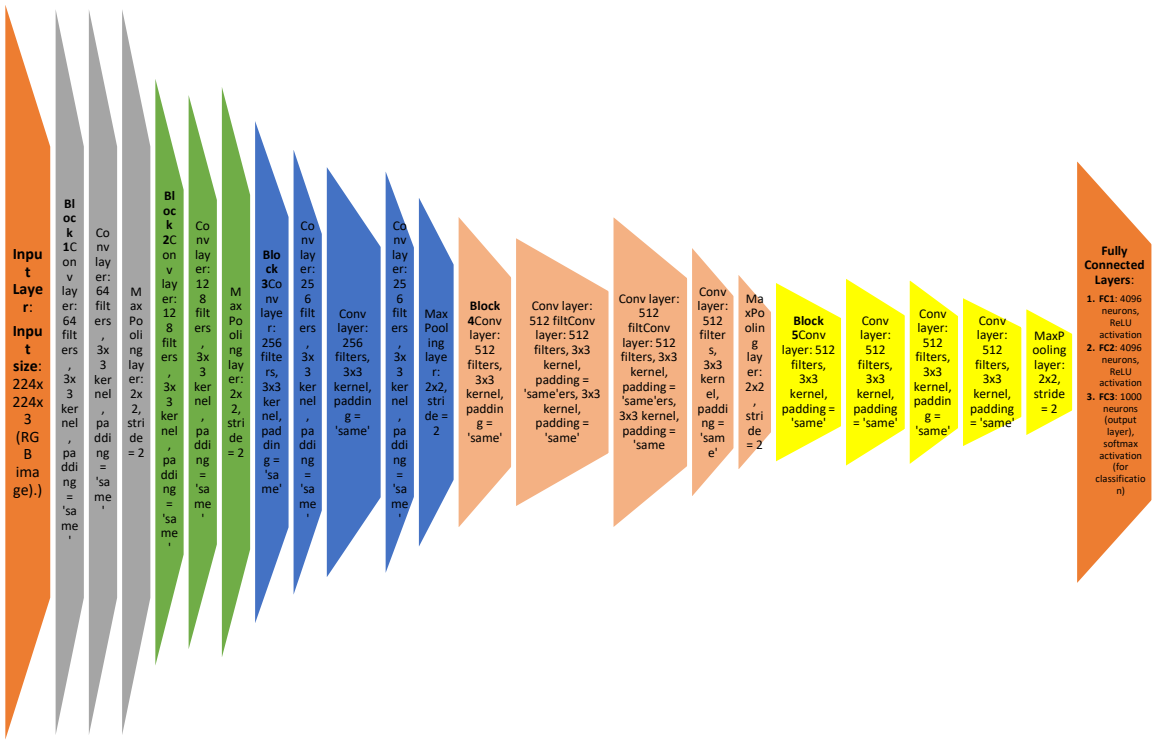

Figure 6.Architecture of VGG19

6.6. VGGFace

VGGFace is a model for image recognition that leverages standard datasets for face identification from Oxford's Visual Geometry Group researchers to provide the most sophisticated results [35]. With this method, we can use only a moderate amount of annotation power to create a big training data set. To construct the model, we followed Tai Do Nhu and Kim's [36] VGGFace design. Convolutional and max-pooling layers were present in each of the five layer blocks in this model. The first and second blocks each contained two 3 × 3 convolution layers, which were followed by a pooling layer. A max-pooling layer comes after three 3 × 3 convolution layers, each made up of the third, fourth, and fifth blocks.
In all convolutional layers, the ReLU activation function was used. We have had to modify VGGFace to meet our demands because it employs pretrained weights. We added dense layers to improve the facial traits that were obtained from the five-layer blocks. Lastly, there was also a dense layer included in the output layer that had sigmoid activation.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

7. Evaluation metrics of deepfake detection models

Five distinct Metrics have been employed to assess the models' performance: area under the ROC curve, F1-score, accuracy, precision, and recall.

7.1. Accuracy

Accuracy is defined as the frequency of correct estimations. This formula is used to calculate accuracy.

accuracy= (number of correct predictions/ total number of predictions made) (1)

7.2. Precision

Precision, which is another name for positive predictive value, is the degree to which the model correctly predicts positive values out of all the positive values it is capable of predicting , "Precision" refers to the following :

precision=True Positive / (True Positive + False Positive) (2)

7.3. Recall

The model's Recall can be utilized to assess the efficacy of identifying true positives. A high recall indicates that the model has effectively identified true positives. Conversely, a low recall value causes the model to have a lot of false negatives. The following is what is meant to be remembered:

recall=True Positive /(True Positive + False Negative) (3)

7.4. F1-Score

It is the precision and recall harmonic mean. Comparing the F1-score to the accuracy measure of the incorrectly classified cases yields a more accurate estimate.

F1 _score=2(Precision/Precision + Recall) (4)

Recall and precision must be balanced in the F1-score. As previously observed, True Negatives play a significant role in accuracy. If there is an unequal class distribution (many Actual Negatives), and we need to balance precision and recall, the F1-score would be a preferable metric [37].

7.5. Receiver Operating Characteristic Curve (ROC) and Area under the ROC Curve (AUC)

The AUC-ROC curve is used to evaluate the algorithm's performance for classification tasks. The probability curve is called ROC, and the degree or level of separability is indicated by AUC. It demonstrates the model's ability to distinguish between different classes. Generally speaking, the AUC shows how accurately the model predicts classes 0 and 1. For instance, the more accurately the model distinguishes between patients who are ill and those who are not, the higher the AUC. First, let's define a few terms. At different classification levels, the connection between True Positive Rate and False Positive Rate is depicted by the receiver operating characteristic (ROC) curve. More items are labeled as positive when the categorization criterion is lowered, which raises the number of True Positives and False Positives [36] . A model is considered good if its AUC is close to 1, which suggests a high level of separability. AUC values near zero indicate that a model is insufficient since they have the lowest degree of separability. Indeed, it appears that the outcome is reciprocal. It includes mixing together 1s and 0s and 0s with 0. Moreover, an AUC of 0.5 implies that the model has no ability to distinguish between classes at all.

8. Explainable AI (XAI)

Although artificial intelligence (AI) systems are being used in many advanced applications today, the decisions from many AI models are difficult to interpret and trust due to their opaque natures. It can often be important to know WHY these models modelled this way. Therefore, there is a demand for Explainable AI (XAI) methods to maintain faith in AI techniques. The main goal of eXplainable AI (XAI) is to create models that people are able to comprehend, especially in sensetive sectors like military, banking, healthcare, etc. Domain experts in these fields expect more than solutions to their problems; they expect solutions they can trust and understand the reasoning behind. This intuitiveness is useful not just for experts looking to analyze output, but also for developers, as bad output can lead to more exploration about how

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

the system functions. AI approaches assist (i) the assessment of current knowledge, (ii) the enhancement of knowledge, and (iii) the development of new hypothesis or theory[38]. Moreover, the overarching goals of XAI methods are to enhance justification, improve control, foster system refinement, and facilitate discovery [39] .The benefits of XAI can be summarized as follows, offering greater transparency into black-box systems [40]:

- Do so in a way that empowers people to reduce the harms of automated decision-making.
- To help people make better decisions.
- To identify and help mitigate security vulnerabilities.
- – Human alignment of algorithms is a primary goal
- Helping brands set higher standards in the development of AI using products thus building trust among businesses and consumers.

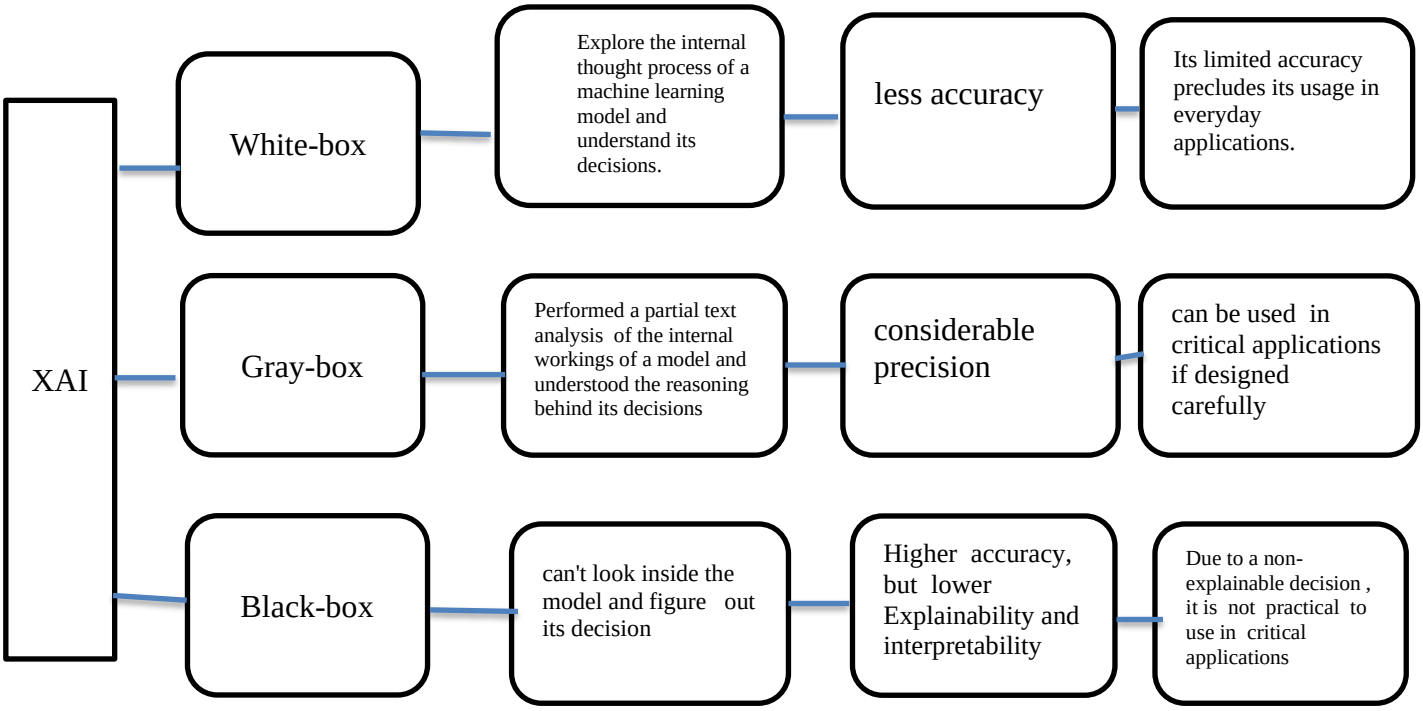

Fig.7. White-box, gray-box, and black-box models depicted.

8.1. Black-Box Model

In explainable AI (XAI), we call a model a black-box model when it is a machine learning model that is treated as a sensory approach, where its internal processes and reasons for its predictions are obscured from users and not understandable. Although these models generate predictions from input data, their decision-making processes are not transparent, preventing the user from learning how the model works, finding out if the model is biased or produces mistakes, or holding any responsibility regarding its conclusions. In the context of XAI, "black box" models are often contrasted with "white box" or "transparent" models, where the internal mechanisms and the reasoning behind the predictions are fully visible and interpretable. Transparent models allow individuals to more easily understand and trust the conclusions made by the system. However, despite their effectiveness, highly predictive models such as deep neural networks (DNNs) are often lacking with respect to interpretability, representing challenges that

needs be overcome, particularly in domains that demand rational and accountable decision making. the following explanatory techniques:

8.1.1.Grad-CAM++ [41] makes use of the gradients produced during back-propagation to generate visual explanations. To produce the visual explanation, a weighted combination of the positive partial derivatives of the last convolutional layer w.r.t a class score is computed, which is a straightforward extension of the original Grad-CAM.This approach permits improved exploration of the object and can avoid the case where one single image could have multiple instances of the same object [42].

8..1.2. *RISE (Randomized Input Sampling for Explanation*) is a visually explainable method which follows perturbation-based approach. The random masking of patches of an input image is central to evaluating the effect the patches have on a machine learning model output.This approach first generates a stream of binary masks that hide certain parts of the image, giving rise to a sequence of altered images. The model inspected each of the altered images and generated predictions, which it then used to assign weights to the corresponding binary masks. Multiple weighted masks combine to create the final visual appearance [43].

8.1.3. SHAP (SHapley Additive exPlanations) is an approach based on attribution principles and relies on game theory concepts, particularly on the Shapley values. It constructs an additive model for a local feature attribution that distributes an effect on to each input feature and then aggregates that effect, called SHAP values, to approximate the model output locally. This method treats each pixel of the input image as a player in a coalition game, where their inclusion or exclusion impacts the final prediction. The outcome of the coalition matches what the model predicts and Shapley values are used to fairly distribute this outcome between the pixels, using the prediction of the test sets through modified images to analyze each pixel's contribution .

8.1.4. LIME (Local Interpretable Model-agnostic Explanations) [44] Such methods use perturbation approaches to attain visual explanations through randomly occluding regions of an input image to evaluate their impact on a model prediction. LIME's core idea is to create a local approximation of a model's behavior using a simpler one that is more interpretable around a specific instance. The first step is segmentation of the input image, then random masking is applied on the segments to perturb them. These modified images are entered into the model producing the predictions. Finally, induce simpler model to find visual explanation on binary masks by applying linear model (linear regressor) to binary masks for every perturbations, and find weights/coefficients from the simple model.

8.1.5. SOBOL [45] It is based on attribution and uses Sobol' indices (developed by Ilya M. Sobol) which help in identifying how much input variables contribute to the output variability of a model. They use a Quasi-Monte Carlo sequence to generate a set of real-valued masks. Next, the masks are applied to an input image using several perturbation methods (e.g., blurring), resulting in warped copies of the image. The model analyzes the modified photos for prediction scores. SOBOL works by examining the relation between the generated masks and their corresponding prediction scores, computing the entire order of Sobol' indices.

9. Limitations and Future Directions

9.1.Limitations:

Generalization Problem: Most deepfake detection models have issues generalizing to unobserved or more recent forms of GAN-generated media. GAN technology itself is evolving quickly, and models trained only on older datasets may not detect deepfakes created by newer methods.

Data Dependency The effectiveness of deepfake detection models is highly dependent on the diversity and quality of the training dataset. A lack of diverse datasets may limit the model's ability to accurately detect different kinds of deepfakes .

Computational Complexity: Most deep learning-based detection methods, especially those involving complex CNN architectures, are computationally expensive, making real-time or low-resource application difficult .

Dataset Limitations: Several methods have only been tested on specific datasets like CelebA or LFW, limiting their generalizability to other datasets with varying attributes or more diverse conditions .

Over-reliance on Training Data: Models may overfit on specific features from training datasets, reducing their performance on real-world, previously unseen data.

9.2.Future Directions:

Improving Generalization: To improve generalization, The development of models that are able to detect deepfakes from a wider range of GAN architectures and data conditions .

Adaptive Models: Researchers could develop adaptive learning models that can evolve alongside advancements in deepfake creation techniques. These models could continually update to recognize newer manipulations.

Diverse and Larger Datasets: Increasing the variety and volume of datasets used for training could improve model robustness. This includes incorporating different environments, lighting conditions, and demographic variations to simulate real-world scenarios.

Real-time and Low-resource Applications: Future efforts could prioritize optimizing models to work in real-time and in resource-constrained environments, making them more applicable for practical use in cybersecurity and law enforcement.

Multi-modal Approaches: There is potential to combine audio, video, and spatial analysis for more comprehensive deepfake detection. Multimodal detection methods could exploit the inconsistencies between visual and auditory data.

10. Conclusion

This review examined various sophisticated methods for detecting images generated by convolutional neural networks and for identifying deepfake content. Present methodologies predominantly utilize deep convolutional neural networks (CNNs) alongside pre-processing techniques to examine the statistical characteristics of images. Although these methods have demonstrated impressive accuracy, numerous earlier studies have neglected the essential concern of generalizability, given that models are frequently trained and evaluated on identical datasets. To tackle this issue, innovative models like forensic CNNs have been introduced, employing methods such as Gaussian Blur and Gaussian Noise to improve the model's capacity to differentiate between authentic and counterfeit images. By concentrating on high-frequency pixel noise and disregarding unnecessary low-level artefacts, these models are more adept at identifying deepfakes. Furthermore, networks employing a dual-stream methodology, which combines facial classification with low-level feature extraction, have demonstrated improved efficacy in identifying modified images. Transfer learning has become a potent method, enabling the application of pre-trained networks such as Xception, MobileNet, NAS-Net, and VGG for deepfake detection. Enhancing these

models with updated datasets can achieve great accuracy. Hybrid deep learning approaches, especially those employing paired learning, have markedly improved detection efficacy by addressing the limitations of traditional deepfake detection models.In summary, these varied methodologies have demonstrated considerable success in identifying deepfake images, attaining impressive accuracy levels. Nonetheless, the persistent challenge lies in enhancing the applicability of models across varied datasets and addressing the growing complexity of deepfake generation methods. Innovative pre-processing techniques and transfer learning persist in presenting encouraging pathways for the creation of more resilient and dependable models in the battle against image falsification.

REFERENCES

[1] B. U. Mahmud and A. J. a. p. a. Sharmin, "Deep insights of deepfake technology: A review," 2021.
[2] G. Pei *et al.*, "Deepfake Generation and Detection: A Benchmark and Survey," vol. abs/2403.17881, 2024.
[3] A. S. Priya, T. J. I. J. o. S. Manisha, and R. Archive, "CNN and RNN using Deepfake detection," 2024.
[4] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. J. A. S. Lee, "Deep Fake Image Detection Based on Pairwise Learning," 2020.
[5] J. Mawa and M. H. J. I. J. o. C. A. Kabir, "Study on Deepfake Face Detection Using Transfer Learning Approach," 2024.
[6] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, J. J. I. I. B. S. C. o. C. Marchang, and Networking, "Deepfake Image Detection Using Vision Transformer Models," pp. 332-335, 2024.
[7] S. St *et al.*, "Deep learning model for deep fake face recognition and detection," vol. 8, 2022.
[8] P. Zhou, X. Han, V. I. Morariu, L. S. J. I. C. o. C. V. Davis, and P. R. Workshops, "Two-Stream Neural Networks for Tampered Face Detection," pp. 1831-1839, 2017.
[9] Z. N. Ashani *et al.*, "Comparative Analysis of Deepfake Image Detection Method Using VGG16, VGG19 and ResNet50," 2024.
[10] A. J. b.-T. Adhicitta, "Application Of Deep Learning For Image Deepfake Detector Using Convolutional Neural Network Algorithm," vol. 6, no. 2, pp. 198-207, 2023.
[11] W. H. Abir *et al.*, "Detecting deepfake images using deep learning techniques and explainable AI methods," vol. 35, no. 2, pp. 2151-2169, 2023.
[12] M. M. J. J. o. C. Alwateer and Communications, "Explainable Deep Fake Framework for Images Creation and Classification," vol. 12, no. 5, pp. 86-101, 2024.
[13] P. S. J. I. J. f. R. i. A. S. Rekha G and E. Technology, "Deepfake: Creation and Detection using Deep Learning," 2023.
[14] P. Sharma, M. Kumar, H. Sharma, S. M. J. M. T. Biju, and Applications, "Generative adversarial networks (GANs): Introduction, Taxonomy, Variants, Limitations, and Applications," 2024.
[15] G. Cohen and R. Giryes, "Generative adversarial networks," in *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*: Springer, 2023, pp. 375-400.
[16] J. Kossen and M. J. W. M. l. Müller, "Generative gegnerische Netzwerke," 2019.
[17] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-y. J. I. C. J. o. A. S. Wang, "Generative adversarial networks: introduction and outlook," vol. 4, pp. 588-598, 2017.
[18] M. Mirza and S. J. A. Osindero, "Conditional Generative Adversarial Nets," vol. abs/1411.1784, 2014.
[19] Y. Shi, X. Liu, Y. Wei, Z. Wu, W. J. I. C. C. o. C. V. Zuo, and P. Recognition, "Retrieval-based Spatially Adaptive Normalization for Semantic Image Synthesis," pp. 11214-11223, 2022.

[20] M. Liu *et al.*, "STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing," pp. 3668-3677, 2019.
[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. J. I. C. o. C. V. Nießner, and P. Recognition, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," pp. 2387-2395, 2016.
[22] L. Li, J. Bao, H. Yang, D. Chen, and F. J. A. Wen, "FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping," vol. abs/1912.13457, 2019.
[23] M. Ferrara, A. Franco, and D. J. I. I. J. C. o. B. Maltoni, "The magic passport," pp. 1-7, 2014.
[24] K. Wang *et al.*, "MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation," in *European Conference on Computer Vision*, 2020.
[25] Y. Li, X. Yang, P. Sun, H. Qi, S. J. I. C. C. o. C. V. Lyu, and P. Recognition, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," pp. 3204-3213, 2019.
[26] K. N. Ramadhani, R. J. r. I. C. o. I. Munir, and C. Technology, "A Comparative Study of Deepfake Video Detection Method," pp. 394-399, 2020.
[27] R. A. Sattar and D. M. N. M. Raouf, "Deepfake Image and video detection using deep learning method."
[28] B. Liu, B. Liu, M. Ding, T. Zhu, and X. Yu, "TI2Net: temporal identity inconsistency network for deepfake detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4691-4700.
[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. J. I. C. o. C. V. Wojna, and P. Recognition, "Rethinking the Inception Architecture for Computer Vision," pp. 2818-2826, 2015.
[30] T. T. Nguyen *et al.*, "Deep learning for deepfakes creation and detection: A survey," vol. 223, p. 103525, 2019.
[31] D. Afchar, V. Nozick, J. Yamagishi, I. J. I. I. W. o. I. F. Echizen, and Security, "MesoNet: a Compact Facial Video Forgery Detection Network," pp. 1-7, 2018.
[32] Y. Li, M.-C. Chang, and S. J. A. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," vol. abs/1806.02877, 2018.
[33] U. A. Ciftci, I. Demir, and L. J. I. I. J. C. o. B. Yin, "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals," pp. 1-10, 2020.
[34] U. A. Ciftci, I. J. I. t. o. p. a. Demir, and m. intelligence, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," vol. PP, 2019.
[35] N.-T. Do, I.-S. Na, and S.-H. J. I. Kim, "Forensics face detection from GANs using convolutional neural network," vol. 2018, pp. 376-379, 2018.
[36] G. Developers. (2024, November 1, 2024). *Classification: ROC and AUC (Not applicable ed.)* [Online Resource].
[37] S. M. Borstelmann and S. J. v. Jha, "Confusion in the Matrix: Going Beyond the Roc Curve," 2019.
[38] "Human Protein Sequence Classification using Machine Learning and Statistical Classification Techniques %J International Journal of Recent Technology and Engineering," 2019.
[39] A. Adadi and M. J. I. A. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," vol. 6, pp. 52138-52160, 2018.
[40] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. J. A. C. S. Giannotti, "A Survey of Methods for Explaining Black Box Models," vol. 51, pp. 1 - 42, 2018.
[41] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. J. I. W. C. o. A. o. C. V. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," pp. 839-847, 2017.
[42] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. J. I. J. o. C. V. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," vol. 128, pp. 336 - 359, 2016.

[43] V. Petsiuk, A. Das, and K. J. A. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," vol. abs/1806.07421, 2018.

[44] M. T. Ribeiro, S. Singh, C. J. P. o. t. n. A. S. I. C. o. K. D. Guestrin, and D. Mining, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," 2016.

[45] T. Fel, R. Cadène, M. Chalvidal, M. Cord, D. Vigouroux, and T. J. A. Serre, "Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis," vol. abs/2111.04138, 2021.