IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Deep learning-based identification of deep fake images

Alap Mahar[1], Dr. Pushpneel Verma[2], Dr. Ajit Singh[3]

[1]Department of Computer Science and Engineering, Bhagwant University, India

[2]Department of Computer Science and Engineering, Bhagwant University, India

[3]Department of Computer Science and Engineering, VMSB University, India

Abstract

This study delves into the intricate realm of deep learning-based deepfake images, seeking to unravel the complexities and implications of this burgeoning technology. Deepfakes, driven by sophisticated neural network architectures such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have transformed the landscape of digital content creation, enabling the synthesis of remarkably realistic yet entirely fabricated images. This research aims to scrutinize the methodologies employed in the generation of deepfake images, exploring the nuances of GANs, VAEs, and other deep learning techniques that underpin their creation. According to the comparative study, the K-NN approach is not effective in detecting deepfake images, with a low precision and accuracy of 84% and 79%, respectively. On the other hand, the hybrid method (SVM+RF) demonstrates significant improvement, with an F1-score of 94% and an accuracy of 97%, exhibiting superiority over other methods. Furthermore, the study investigates the open challenges and research directions associated with detecting and mitigating the impact of deepfake images, addressing the ethical concerns arising from their potential misuse.

Keyword: Deep learning, deepfake, deepfake creation

1. Introduction

The progress of web technology has led to the widespread availability of information. While the Internet offers a wide range of information, the reliability of that material is influenced by several variables. An immense quantity of information is disseminated on a daily basis via internet and print media; nonetheless, it is challenging to ascertain the veracity of the material[1]. A comprehensive examination and analysis of the material is required, including the verification of facts via the evaluation of supporting sources, understanding the origin of the information, and establishing the trustworthiness of the writers. The dissemination of fabricated information is a deliberate effort to diminish or enhance the reputation of an organization, company, or individual, with the objective of gaining financial or political benefits[2]. Fabrication is the word used to describe this kind of contrived information, which misleads individuals. During the Indian election campaigns, several fabricated tales, news pieces, and manipulated photographs proliferated on social media[3]. Social media has become indispensable in our culture, intricately interwoven with our everyday lives, activities, and lifestyle, exerting a profound influence. Social media platforms have revolutionized the dissemination and consumption of information, spanning many forms of communication such as texting and blogging[4]. Deepfake technology enables individuals to expedite their fashion choices, hence yielding advantages for the fashion and e-commerce sectors. Moreover, this technology assists the entertainment industry by offering synthetic voices for artists who are unable to provide voiceovers within the designated timeframe[5]. Furthermore, deepfake technology enables filmmakers to faithfully reproduce several iconic scenes or include advanced visual effects into their films. The use of deepfake technology has the potential to enable those diagnosed with Alzheimer's disease to engage in communication with a digitally manipulated representation of their younger self[6]. This innovative approach may facilitate the preservation of their memories. GANs are now being investigation for their potential use in identifying abnormalities in X-ray images [7].

Deepfake techniques often need a substantial amount of picture, video, or audio data in order to produce authentic-looking photographs that may convincingly deceive observers[8]. In addition to all the attention, there are also notable disadvantages. Public personalities, such as celebrities, sports, and politicians, are particularly vulnerable to the negative effects of deepfakes due to the abundance of videos and photos of them accessible on the internet. While deep fake technologies may be used on occasion to mock individuals, their primary purpose is to generate explicit material[9]. Images including the visages of several celebrities and prominent figures have been digitally superimposed onto the physiques of pornographic models, and these explicit visuals are readily accessible on the Internet [10]. Deepfake technology enables the creation of satirical, pornographic, or political material using well-known individuals by utilizing their images and sounds without their permission. Thanks to the accessibility of many programmes, anybody may create false content that is indistinguishable from genuine material [11]. A significant number of adolescents are falling prey to cyber bullying. If the situation reaches its most extreme outcome, a significant number of individuals may resort to taking their own lives[12]. With a high degree of precision, the deep learning techniques were able to learn salient facial biometric patterns for the purpose of face identification [13]. The databases that include the human faces are used for the purpose of training models that are based on deep learning. The models that are based on deep learning perform better than the skills of people when it comes to facial recognition.

Figure 1. Example of deepfake [14]

1.1 Creation of deepfake

There is a term that may be used interchangeably with the term "deepfakes technology" to refer to any video or picture that has been altered via the use of deep neural networks. Standard computers are not capable of performing this manipulation, which necessitates the use of high-end desktop PCs equipped with powerful graphics cards or, even better, cloud computing capability. The amount of processing time necessary to train deep neural networks that are responsible for the creation of deepfakes is reduced using this method. These are the criteria that may be used to classify deepfakes in terms of face manipulation[15]:

1.1.1 Face swap

The type of face alteration that has become the most widespread in recent times is face swapping. The use of a certain kind of deep neural network is what makes this possible. This particular kind is an autoencoder, which is used in the process of feature extraction as well as picture compression [16], [17]. Reddit users have been using the autoencoder structure, which consists of an encoder and a decoder, as the first step in the process of creating deep fakes [18] [19]. It is the capacity to extract the original data from this medium representation that is the fundamental concept of autoencoder. This is accomplished by first representing the input data into a smaller and more compressed form. A visual representation of the process of making deepfakes may be seen in Figure 2. The training of two autoencoders is often required for this operation. Each autoencoder will operate on a collection of video clips of one individual from the two individuals whose identities will be reversed. The target video is then sent to the incorrect decoder in order to generate a deepfake face [20]. This occurs after the autoencoders have been trained. In general, deepfakes that are formed using autoencoder do not pay much attention to the differences between the identities of the source face and the target face. Instead, they make the swapped face seem to be the same as both the source face and the target face.

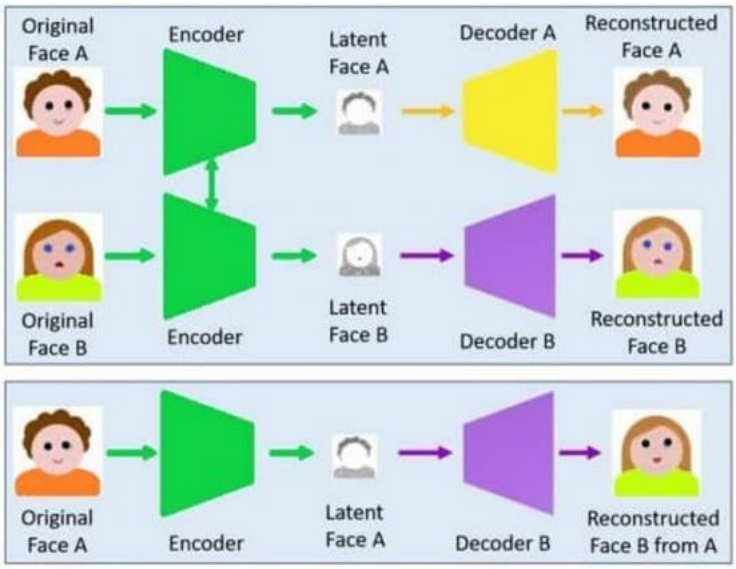

Figure 2. Deepfake creation process. Each autoencoder is trained on a set of images of one person

Face synthesis

For the purpose of generating non-existent genuine faces, generative adversarial networks, often known as GANs, are used in this field. Deepfakes came into being as a consequence of the development of GANs, which contributed to the production of results that were shockingly realistic [21]. The most common method is the employment of STYLEGAN, which is a subcategory of GANs. This method was used to generate the website that seemed to be "this person does not exist" [22]. Researchers enhance the capabilities of the style GAN architecture and propose a new version of the style GAN, which they call StyleGAN2 [23].

1.2 Deep learning techniques for deepfake

Deepfake technology, which involves the use of deep learning algorithms to create realistic-looking but fabricated content, has seen various approaches within the realm of deep learning[24]. These approaches leverage advanced neural network architectures and techniques to generate convincing fake images, videos, or audio recordings. Here are some key types of approaches in deep learning for deepfakes:

1.2.1 Generative Adversarial Networks (GANs): GANs are widely used in deepfake creation. GANs consist of two neural networks – a generator and a discriminator – engaged in a continuous adversarial dance. The generator fabricates synthetic data, such as images, while the discriminator scrutinizes both real and generated samples, striving to distinguish between the two. Through iterative training, the generator refines its abilities to produce increasingly realistic content, while the discriminator hones its discernment skills. This dynamic interplay results in the generation of remarkably convincing deepfakes that blur the line between reality and fabrication. GANs have demonstrated their prowess in various deepfake applications, from altering facial features to manipulating entire scenes in videos. The generator aims to produce realistic content, while the discriminatorlearns to distinguish between real and generated data. The iterative training process results in the generation of increasingly convincing deepfakes[25].

1.2.2 Variational Autoencoders (VAEs): VAEs are another class of generative models that learn to encode and decode data, enabling the generation of new content. In the context of deepfake generation, VAEs operate as generative models capable of encoding and decoding complex data distributions. Unlike traditional autoencoders, VAEs introduce a probabilistic framework that enables the generation of diverse and realistic outputs. In the deepfake domain, VAEs are employed to capture the inherent variability in facial expressions, poses, and other features of human subjects. By learning a latent space representation, VAEs facilitate the synthesis of new facial images that align with the learned distribution of real data. This method not only allows for the creation of more convincing and natural-looking deepfakes but also provides a means to explore the underlying structures of facial features.

1.2.3 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks: RNNs and LSTMs are used for sequential data generation, such as deepfake videos or audio. These networks capture temporal dependencies, allowing for the creation of more coherent and contextually relevant deepfake content[26]. In the realm of deepfake detection, RNNs and LSTMs prove beneficial in scrutinizing the temporal dynamics of facial expressions, lip movements, and overall consistency over time, allowing for more accurate differentiation between genuine and manipulated videos [27]. However, challenges persist, including

the need for extensive labeled training data to effectively train these networks and the computational complexity associated with processing high-dimensional video sequences. Despite these challenges, the application of RNNs and LSTMs underscores their potential to advance the state-of-the-art in both deepfake generation and detection, contributing to the ongoing dialogue on the ethical and technological implications of this rapidly evolving field.

2. Review of Literature

Ojha et al., (2023)[28]studied thatthe increasing prevalence of generative models necessitates the development of versatile counterfeit picture detection systems. In this study, we first demonstrate the inadequacy of the current approach, which involves training a deep neural network to classify genuine and false photos, in accurately identifying counterfeit images generated by more recent iterations of generative models when trained specifically to detect fake images produced by Generative Adversarial Networks (GANs). After conducting an examination, it has been shown that the classifier produced has an unbalanced tuning towards the identification of patterns indicative of picture falsification. When provided with access to the feature space of a large pretrained vision-language model, the basic approach of nearest neighbour classification demonstrates unexpectedly strong generalisation capability in identifying counterfeit images generated by various generative models. For instance, it outperforms the current state-of-the-art by an increase of 15.07 mean Average Precision (mAP) and 25.90% accuracy when evaluated on previously unseen diffusion and autoregressive models.

Hou et al., (2023)[29] examined that there has been significant advancement in the field of realistic face forging methods, often referred to as DeepFake. As a result, a growing number of detection approaches for DeepFake have been put out. This study aims to specifically reduce the statistical disparities in order to circumvent advanced DeepFake detection methods. In pursuit of this objective, we provide a statistical consistency attack (StatAttack) targeting DeepFake detectors, including two primary components. Initially, a set of statistical-sensitive natural degradations, including exposure, blur, and noise, are chosen and incorporated into the counterfeit photographs using an adversarial approach. Furthermore, it is observed that the statistical disparities between natural images and DeepFake images exhibit a positive correlation with the shifting of distributions between these two image types. Moreover, we propose an enhanced version of StatAttack called MStatAttack. In MStatAttack, we successively introduce multi-layer degradations and jointly optimise the combination weights using the loss function. The efficacy of our suggested attack strategy in both white-box and black-box scenarios is shown via extensive experimentation with four spatial-based detectors and two frequency-based detectors over four datasets.

Rafique et al., (2023)[30] analyzed that the proliferation of readily accessible material on social media, coupled with the use of sophisticated tools and cost-effective computer infrastructure, has facilitated the widespread production of deep fakes. Therefore, the development of a comprehensive system for discerning authentic and counterfeit information has become imperative in the contemporary era of social media. This study presents a novel approach for the automatic categorization of deep fake pictures via the use of advanced techniques in Deep Learning and Machine Learning. Conventional machine learning (ML) methods that rely on manual feature extraction are limited in their ability to catch intricate patterns that are not well-understood or readily represented using basic features. The proposed framework does an initial Error Level Analysis of the picture in order to ascertain if any modifications have been made to it. The provided picture is then used by Convolutional Neural Networks (CNNs) to perform deep feature extraction. The feature vectors obtained are next subjected to classification using Support Vector Machines and K-Nearest Neighbours, with hyper-parameter optimisation being carried out. The approach provided in this study demonstrated the best level of accuracy, reaching 89.5%, by using a combination of Residual Network and K-Nearest Neighbour techniques. The findings demonstrate the effectiveness and resilience of the suggested methodology, therefore indicating its potential use in identifying deep fake pictures and mitigating the associated risks of defamation and propaganda.

Dong et al., (2023)[31] examined the generalisation capacity of binary classifiers in the context of deepfake detection. The obstacle to their ability to generalise is attributed to the unforeseen acquired identity representation on visual stimuli. Referred to as the Implicit Identity Leakage, this phenomenon has been empirically validated via qualitative and quantitative analyses across many Deep Neural Networks (DNNs). Moreover, drawing upon this comprehension, we put forward a straightforward but efficacious approach referred to as the ID-unaware Deepfake Detection Model, with the aim of mitigating the impact of this occurrence. The experimental findings provide substantial evidence that our strategy surpasses the current

leading approach in terms of performance, as shown in evaluations conducted inside the dataset as well as across other datasets.

Choudhury et al., (2023) [32] studied the rapid dissemination of false information across many social media platforms has made it impossible to avoid incorporating it into our daily lives. As a result of the proliferation of fake news across social and news platforms, identifying such content has become a hot topic in the academic community. In this research, we evaluate the performance of the Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR) classifiers on four datasets including bogus news. In the three datasets tested (Liar, Fake Job Posting, and Fake News), the SVM classifier obtained the maximum accuracy at 61%, 97%, and 96%, respectively. In our unique GA-based false news detection algorithm, we again take into account SVM, Naive Bayes, Random Forest, and Logistic Regression as potential fitness functions. We found that our proposed technique yielded 61% accuracy using SVM and LR classifiers on the LIAR dataset, and 97% accuracy using SVM and RF on the fabricated job posting dataset.

Raza et al., (2022)[33] stated that deepfake technology is used in the realm of synthetic media with the purpose of creating fabricated visual and auditory information that is derived from pre-existing media of an individual. Deepfake technology is used to convincingly alter a person's appearance and vocal characteristics by substituting them with fabricated multimedia elements. The production of fabricated media information is considered immoral and poses a significant harm to the community. Based on a recent analysis by Sensity, a significant majority (over 96%) of deepfakes consist of explicit or offensive material. The main objective of our research investigation is to identify deepfake media via the implementation of a very effective framework. This study presents a unique strategy for predicting deepfakes, referred to as the Deepfake Predictor (DFP). The suggested method combines the VGG16 model with a convolutional neural network architecture. The transfer learning methods used for comparison are the Xception, NAS-Net, Mobile Net, and VGG16. The DFP technique, as presented, demonstrated a precision rate of 95% and an accuracy rate of 94% in the context of deepfake detection. The innovative DFP strategy shown in our study demonstrated superior performance compared to transfer learning approaches and other contemporary research in the field.

Birunda et al., (2022)[34] studied one of the foremost issues in the realm of online social networks pertains to the dissemination of counterfeit photos or manipulated colourized images, whereby individuals possess the ability to introduce, eliminate, or modify visual content. Currently, there is a lack of viable models or techniques capable of effectively discerning and classifying photographs as either authentic or counterfeit. This research aims to use the flood fill technique for the purpose of accentuating the counterfeit item inside the picture. Additionally, a solution based on Deep Learning is provided to ascertain the authenticity of the image. The Twitter dataset is gathered and used as input for deep learning models, which are then trained to discern the authenticity of images. The experimental assessments shown that the suggested framework exhibits a 96% accuracy rate in detecting counterfeit photographs disseminated over the Twitter platform.

Dong et al., (2022)[35] analyzed the process by which deepfake detection algorithms acquire knowledge of artefact characteristics in photos only via binary label supervision. In order to achieve this objective, three theories pertaining to image matching are put forward. Deepfake detection methods discern the authenticity of pictures by evaluating visual ideas that are unrelated to the original source or intended target, instead focusing on identifying visual artefacts. In addition to overseeing binary labels, deepfake detection algorithms acquire knowledge about visually significant artefacts via the FST-Matching process, which involves comparing fake, source, and target pictures within the training dataset. The visual ideas of artefacts learnt implicitly using FST-Matching in the unprocessed training set are susceptible to degradation when subjected to video compression techniques. The aforementioned ideas are validated across several deep neural networks in experimental settings. Moreover, building upon this comprehension, we provide the FST-Matching Deepfake Detection Model as a means to enhance the efficacy of forgery detection on compressed video content. The experimental findings demonstrate that our approach has exceptional performance, particularly when used to films with high levels of compression, such as c40.

Chang et al., (2020)[36] analyzed that DeepFake technology has the capability to produce manipulated photos and videos of superior quality that closely resemble authentic data. The quick pace of its growth elicits both terror and introspection among individuals. This research introduces an enhanced VGG network, referred to as NA-VGG, for the purpose of detecting DeepFake face images. The proposed network leverages image noise and image augmentation techniques to increase its performance. To begin with, the use of the SRM filter layer is employed as a means to identify tampering artefacts that may not be readily discernible in the RGB channels. Subsequently, the image noise features are enhanced by augmentation of the image noise map, hence diminishing the prominence of facial characteristics. Ultimately, the enhanced noise pictures are fed into the

neural network for the purpose of training and evaluating the authenticity of the image. The use of the Celeb-DF dataset in experimental analysis has shown that NA-VGG has exhibited significant advancements in comparison to other contemporary false image detection methods.

Kesarwani et al., (2020) [37] examined the consumption of news obtained via social media is steadily growing as a result of the fact that it is simple to access, that it is inexpensive, and that it is more appealing, as well as the fact that it is capable of spreading "fake news." The pervasiveness of false news has the potential to have negative repercussions, both for individuals and for society as a whole. On social media, some users intentionally publish false material to get attention or to further their own financial or political interests. The ability to distinguish between authentic and false news is a skill that has to be improved upon. The distinctive characteristic of identifying bogus news on social media renders existing detection algorithms inefficient or inappropriate. After then, it is necessary to take into consideration any secondary information. The social actions of the user as documented on various social media platforms are an example of secondary information. With the assistance of the K-Nearest Neighbor classifier, we will now provide a straightforward method for identifying bogus news on social media platforms. This method was developed as part of this study endeavor. When evaluated on the dataset consisting of Facebook news postings, we were able to get a classification accuracy of about 79% using this model.

Jain et al., (2019) [38] studied that most people who use smartphones would rather get their news from social media than the web. The news websites are the ones disseminating the information and serving as the authority. The challenge is verifying the veracity of information shared on platforms like WhatsApp chats, Facebook pages, Twitter, and other microblogs and social networking sites. Spreading false information as news is bad for society. In emerging nations like India, putting an end to rumors and shifting attention to accurate, certified news pieces is an urgent need. In this study, we provide both the model and the methods for this task. The author made an effort, aided by machine learning and NLP, to compile the news and then use a Support Vector Machine to establish whether or not the news was genuine. The suggested model's findings are compared to those of previously used models. The suggested model performs well, with results correctness defined to a level of 93.6%.

Bahad et al., (2019) [39] analyzed that the media is crucial to informing people about what's going on in the world. Due to the Internet's fast growth, news may now travel rapidly via many online mediums such as social networks and websites. Unverified or fraudulent news is disseminated via social networks and reaches thousands of people without any scrutiny. Oftentimes, fake news is created to serve the economic and political goals of misinforming and attracting readers. Fake news is a serious problem in modern culture. The study of how to automatically evaluate the veracity of news items is an active area of study. Language models built using deep learning techniques are more popular. Common deep learning models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), can identify intricate patterns in textual data. The recurrent neural network architecture known as Long Short-Term Memory (LSTM) may be used to examine sequences of varying lengths. In this study, we provide a Bi-directional Long Short-Term Memory (LSTM)-recurrent neural network-based model for detecting bogus news. The model's efficacy is evaluated using two publicly accessible datasets of unstructured news items. The findings demonstrate that the Bi-directional LSTM model is more accurate at detecting false news than the competing approaches of convolutional neural networks (CNNs), regular neural networks (RNNs), and unidirectional LSTMs.

2.1 Comparison of reviewed technique

There is a wide range of authors who studied on Deep learning-based identification of deepfake images and give their findings as shown below.

Table 1. Comparison of reviewed technique

Authors [Ref.]	Technique	Outcome
Ojha et al., (2023)[28]	GAN	The proposed method outperforms the current state-of-the-art by an increase of 15.07 mean Average Precision (mAP) and 25.90% accuracy when evaluated on previously unseen diffusion and autoregressive models.
Hou et al., (2023)[29]	DeepFake technique	The efficacy of our suggested attack strategy in both white-box and black-box scenarios is shown via extensive experimentation with four spatial-based detectors and two frequency-based detectors over four datasets.
Rafique et al., (2023)[30]	K-NN and Residual network	The approach provided in this study demonstrated the best level of accuracy, reaching 89.5%, by using a combination of Residual Network and K-Nearest Neighbour techniques.
Dong et al., (2023)[31]	DNN	The experimental findings provide substantial evidence that our strategy surpasses the current leading approach in terms of performance, as shown in evaluations conducted inside the dataset as well as across other datasets.
Choudhury et al., (2023) [32]	SVM and LR	The findings indicate that our proposed technique yielded 61% accuracy using SVM and LR classifiers on the LIAR dataset, and 97% accuracy using SVM and RF on the fabricated job posting dataset.
Raza et al., (2022)[33]	DFP	The DFP technique demonstrated a precision rate of 95% and an accuracy rate of 94% in the context of deepfake detection.
Birunda et al., (2022)[34]	Deep learning	The experimental assessments shown that the suggested framework exhibits a 96% accuracy rate in detecting counterfeit photographs disseminated over the Twitter platform.
Dong et al., (2022)[35]	FST-Matching DeepFake detection	The experimental findings demonstrate that our approach has exceptional performance, particularly when used to films with high levels of compression, such as c40.
Chang et al., (2020)[36]	NA-VGG	The use of the Celeb-DF dataset in experimental analysis has shown that NA-VGG has exhibited significant advancements in comparison to other contemporary false image detection methods.

Kesarwani et al., (2020) [37]	K-NN	When evaluated on the dataset consisting of Facebook news postings, we were able to get a classification accuracy of about 79% using this model.
Jain et al., (2019) [38]	SVM	The suggested model performs well, with results correctness defined to a level of 93.6%.
Bahad et al., (2019) [39]	CNN and RNN	The findings demonstrate that the Bi-directional LSTM model is more accurate at detecting false news than the competing approaches of convolutional neural networks (CNNs), regular neural networks (RNNs), and unidirectional LSTMs.

3. Open Issues and Research direction

In spite of the fact that a significant amount of work has been put into developing deepfake creation and detection, there are still a number of problems that have not been satisfactorily resolved. There will be a discussion of some of them in the following.

3.1 Generalization Capability

It is evident in the literature that the majority of deepfake detection frameworks experience a significant loss in performance when tested with deepfakes, manipulations, or datasets that were not included in their training. Therefore, the task of identifying unfamiliar and innovative deepfakes or methods used to create deepfakes remains a significant obstacle. The capacity of deepfake detectors to generalise is crucial for ensuring accurate and reliable identification of manipulated media, hence fostering public confidence in the authenticity of online information. Several first generalisation strategies have been suggested, but their effectiveness in addressing newly developing deepfakes remains an unresolved matter.

3.2 Explain ability of Deepfake Detectors

The deepfake detection framework's interpretability and reliability are insufficient. The majority of deepfake or face alteration detection approaches in existing research often lack an explanation for the final detection conclusion. The black box nature of deep learning methods is mostly responsible for this. Existing deepfake or face manipulation detection systems provide just a classification, confidence level, or probability score indicating the likelihood of fakeness, without offering a detailed explanation of the findings. Having such a description would be valuable in understanding the rationale behind the detector's specific conclusion. Furthermore, the act of deepfake or face alteration, such as the application of digital cosmetics, may be carried out with either harmless or harmful motives. However, current deepfake or face manipulation detection algorithms are unable to differentiate the intention behind the manipulation. To enhance the interpretability and reliability of the deepfake detection framework, using advanced approaches such fuzzy inference systems [40], layer-wise relevance propagation [41], and the Neural Additive Model [42] may be advantageous.

3.3 Next-Generation Deepfake and Face Manipulation Generators

Enhanced deepfake and face manipulation generating techniques will facilitate the development of more sophisticated and comprehensive deepfake detection algorithms. Some of the current datasets and generation methods have several limitations. Firstly, they lack ultra-high-resolution samples, as the existing methods typically generate samples with a resolution of 1014 × 1024, which is insufficient for the next generation of deepfakes. Secondly, there are limited options for manipulating face attributes. The types of face attribute manipulations are dependent on the training set, which restricts the range of manipulation characteristics and attributes. It also means that novel attributes cannot be generated. Thirdly, there is a problem with video continuity. The deepfake and face manipulation techniques, particularly identity swap, do not consider the smooth continuation of video frames or physiological signals. Lastly, the current databases do not include obvious fake samples, such as a human face with three eyes.

3.4 Vulnerability to Adversarial Attacks

Recent research has shown that deep learning-based techniques for detecting deepfake and face manipulation are susceptible to adversarial samples [43]. While contemporary detectors can well handle many forms of deterioration, such as compression and noise, their accuracy significantly diminishes when subjected to adversarial assaults. Hence, it is essential for future methods to include the capability to address both deepfakes and hostile cases. In order to achieve this objective, the use of diverse multistream and filtering algorithms might prove to be efficacious.

3.5 Mobile Deepfake Detector

Due to the large number of parameters and processing expense, neural network-based deepfake detection algorithms, known for their impressive accuracy, are often not suitable for mobile platforms or apps. Compact and efficient deep learning-based detection systems, suitable for deployment on mobile and wearable devices, can significantly aid in combating deepfakes and fake news.

3.6 Lack of Large-Scale ML-Generated Databases

The majority of research on the identification of AI-generated face samples included the creation of a unique database using different Generative Adversarial Networks (GANs). As a result, several published research exhibit varying levels of performance when it comes to GANs samples, due to the fluctuating and mostly unknown quality of the samples created by GANs. Multiple publicly available databases of artificially generated faces using Generative Adversarial Networks (GANs) should be created to facilitate progress in this challenging area of study.

4. Comparative analysis

In this section, several authors provide their results following the accuracy performance metrics, which are described in table 2. According to Table 2, Rafique and his fellow students were able to greatly boost the accuracy using the K-NN method and Residual network for detect the deepfake images, which resulted in 89.5%. By using a DFP method, Raza and his colleagues obtained 94% accuracy, while Jain and his colleagues attained 93.4% accuracy using the SVM, which is minimum as compared to DFP. By using hybrid method (SVM+RF), Choudhury and his colleagues achieved a superior accuracy of 97% which is greater as compared to SVM, DFP method, and all other methods.

Table 2. Comparative analysis

Author	Year	Technique	Accuracy
Rafique et al., [30]	2023	K-NN + Residual Network	89.5%
Choudhury et al., [32]	2023	SVM+RF	97%
Raza et al., [33]	2022	DFP	94%
Jain et al., [38]	2019	SVM	93.6%
Kesarwani et al., [37]	2020	K-NN	79%

The highly achieved accuracy is revealed in Fig. 3., as can be seen in the following graph. The SVM and Random Forest (SVM+RF) has attained maximum accuracy which is 97% for detect the deepfake images as compared to other methods as shown in the graph.

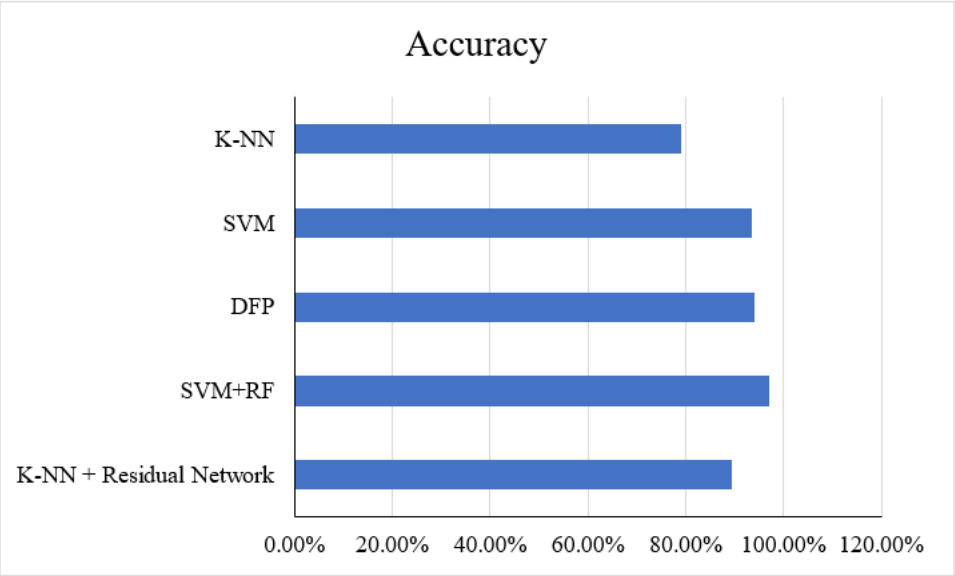

Figure 3. Comparative analysis

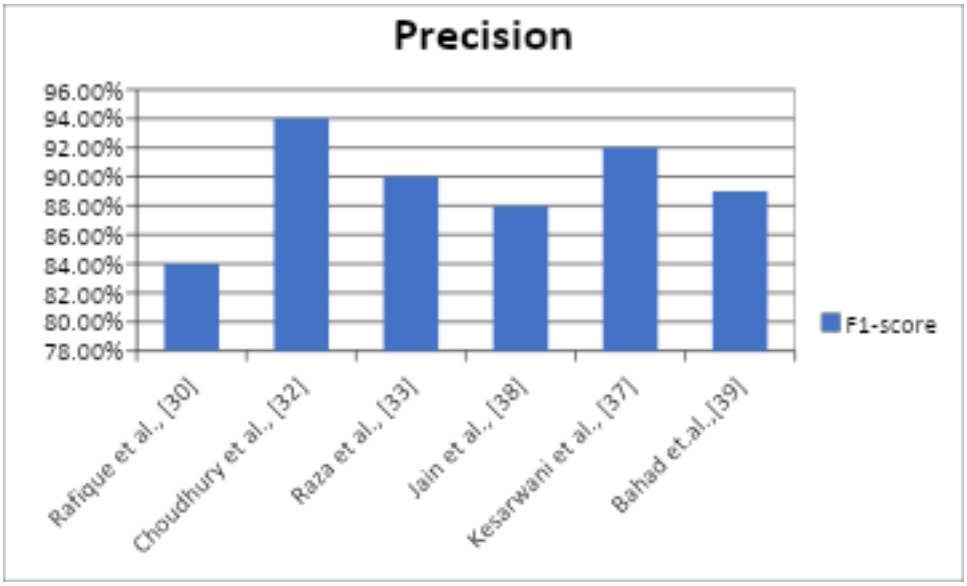

Figure 4. Precision analysis

the Precision analysis shown in figure 4 of various studies, showing that Raza et al., [33] achieved the highest precision at around 94%, while Rafique et al., [30] had the lowest, slightly above 80%. Other notable performances include Bahad et al., [39] and Jain et al., [38], with F1-scores close to 92% and 89%, respectively. The remaining studies by Choudhury et al., [32] and Kesarwani et al., [37] also demonstrated commendable precision, with F1-scores near 82% and 86%. Overall, the graph highlights the variability in precision across different research works.

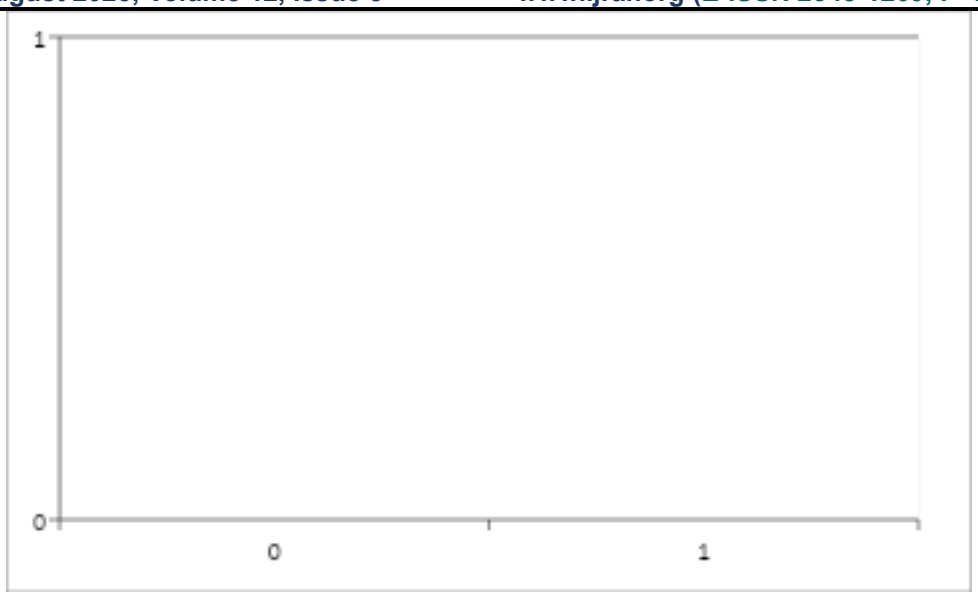

Figure 5. F1-Score analysis

the F1-scores for different methods, demonstrating how well they work in a comparison. With an F1-score of 84%, the k-NN+Residual Network method produced the lowest results. On the other hand, the SVM+RF method performed better, achieving the maximum F1-score of 94%. Additionally successful were the k-NN and DFP approaches, with F1-scores of 92% and 90%, respectively. Meanwhile, with F1-scores of 88% and 89%, SVM and CNN approaches demonstrated a moderate level of precision. Overall, the graph shows a notable amount of variation in performance amongst the various approaches, with SVM+RF emerging as the most successful strategy.

5. Discussion

Studying deep learning-based deepfake images has become a critical focus within the field of artificial intelligence and computer vision, owing to the increasing sophistication and prevalence of deepfake technology. Deep learning techniques, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), play a pivotal role in the generation of highly convincing yet entirely synthetic visual content. This research area involves a multifaceted discussion that encompasses both the creation and detection of deepfake images. Researchers delve into the intricacies of GANs, exploring their ability to capture and mimic intricate details such as facial expressions, gestures, and contextual nuances. The study of deepfake generation extends to novel approaches, including the fusion of different architectures and the integration of additional modalities like audio to create more convincing and sophisticated forgeries. The challenge lies in adapting detection methods to keep pace with the rapid evolution of deepfake generation techniques, necessitating a continuous cycle of innovation and refinement. he potential misuse of deepfake technology for malicious purposes, such as misinformation, identity theft, or privacy invasion, adds a layer of urgency to understanding and mitigating the risks associated with deepfakes. The discourse includes discussions on responsible use, legal implications, and the development of countermeasures to protect individuals and society at large.

6. Conclusion

The study of deep learning-based deepfake images represents a critical exploration into the complex and evolving intersection of artificial intelligence and digital media manipulation. The rapid advancements in generative models, particularly those employing techniques like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have ushered in an era where the creation of highly convincing yet entirely synthetic content is possible. The implications of deepfake technology extend across various domains, from entertainment and creative arts to more serious concerns such as misinformation, identity theft, and privacy breaches. According to the comparative analysis, the K-NN method is not efficient for detecting deepfake images, showing low precision and accuracy at 79% and 84%, respectively. In contrast, the hybrid method (SVM+RF) demonstrated superior performance with an F1-score of 94% and an impressive accuracy of 97%. This highlights the significant potential of advanced hybrid models in outperforming traditional methods. The study of deep learning-based deepfake images not only unravels the intricacies of algorithmic creativity but also challenges us to develop robust countermeasures, fostering a technological landscape where the benefits of AI are harnessed responsibly for the greater good.

References

[1]. Frank, Joel, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. "Leveraging frequency analysis for deep fake image recognition." In *International conference on machine learning*, pp. 3247-3258. PMLR, 2020.

[2]. Suganthi, S. T., Mohamed Uvaze Ahamed Ayoobkhan, Nebojsa Bacanin, K. Venkatachalam, Hubálovský Štěpán, and Trojovský Pavel. "Deep learning model for deep fake face recognition and detection." *PeerJ Computer Science* 8 (2022): e881.

[3]. Chang, Xu, Jian Wu, Tongfeng Yang, and Guorui Feng. "Deepfake face image detection based on improved VGG convolutional neural network." In *2020 39th chinese control conference (CCC)*, pp. 7252-7256. IEEE, 2020.

[4]. Jeon, Hyeonseong, Youngoh Bang, and Simon S. Woo. "Fdftnet: Facing off fake images using fake detection fine-tuning network." In *IFIP international conference on ICT systems security and privacy protection*, pp. 416-430. Cham: Springer International Publishing, 2020.

[5]. Negi, Shweta, Mydhili Jayachandran, and Shikha Upadhyay. "Deep fake: an understanding of fake images and videos." *International Journal of Scientific Research in Computer Science Engineering and Information Technology* 7, no. 3 (2021): 183-189.

[6]. Rahman, Ashifur, Md Mazharul Islam, Mohasina Jannat Moon, Tahera Tasnim, Nipo Siddique, Md Shahiduzzaman, and Samsuddin Ahmed. "A qualitative survey on deep learning based deep fake video creation and detection method." *Aust. J. Eng. Innov. Technol* 4, no. 1 (2022): 13-26.

[7]. Westerlund, Mika. "The emergence of deepfake technology: A review." *Technology innovation management review* 9, no. 11 (2019).

[8]. Badale, Anuj, Lionel Castelino, Chaitanya Darekar, and Joanne Gomes. "Deep fake detection using neural networks." In *15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. 2018.

[9]. Ananthi, M., P. Rajkumar, R. Sabitha, and S. Karthik. "A secure model on Advanced Fake Image-Feature Network (AFIFN) based on deep learning for image forgery detection." *Pattern Recognition Letters* 152 (2021): 260-266.

[10]. Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. "Deepfakes and beyond: A survey of face manipulation and fake detection." *Information Fusion* 64 (2020): 131-148.

[11]. Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. "Deep learning for deepfakes creation and detection: A survey." *Computer Vision and Image Understanding* 223 (2022): 103525.

[12]. Dong, Shichao, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. "Explaining deepfake detection by analysing image matching." In *European Conference on Computer Vision*, pp. 18-35. Cham: Springer Nature Switzerland, 2022.

[13]. Hsu, Chih-Chung, Chia-Yen Lee, and Yi-Xiu Zhuang. "Learning to detect fake face images in the wild." In *2018 international symposium on computer, consumer and control (IS3C)*, pp. 388-391. IEEE, 2018.

[14]. Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. "Do (microtargeted) deepfakes have real effects on political attitudes?." *The International Journal of Press/Politics* 26, no. 1 (2021): 69-91.

[15]. Korshunova, Iryna, Wenzhe Shi, Joni Dambre, and Lucas Theis. "Fast face-swap using convolutional neural networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 3677-3685. 2017.

[16]. Zhou, Lei, Chunlei Cai, Yue Gao, Sanbao Su, and Junmin Wu. "Variational autoencoder for low bit-rate image compression." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2617-2620. 2018.

[17]. Cheng, Zhengxue, Heming Sun, Masaru Takeuchi, and Jiro Katto. "Deep convolutional autoencoder-based lossy image compression." In *2018 Picture Coding Symposium (PCS)*, pp. 253-257. IEEE, 2018.

[18]. Vyas, Harshal. "Deep fake creation by deep learning." *Extraction* 7, no. 07 (2020).

[19]. Katarya, Rahul, and Anushka Lal. "A study on combating emerging threat of deepfake weaponization." In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 485-490. IEEE, 2020.

[20]. Pu, Jiameng, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. "Deepfake videos in the wild: Analysis and detection." In *Proceedings of the Web Conference 2021*, pp. 981-992. 2021.

[21]. Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. "Fighting deepfake by exposing the convolutional traces on images." *IEEE Access* 8 (2020): 165085-165098.

[22]. Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110-8119. 2020.

[23]. Zhu, Peihao, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. "Improved stylegan embedding: Where are the good latents?." *arXiv preprint arXiv:2012.09036* (2020).

[24]. Zhang, Tao. "Deepfake generation and detection, a survey." *Multimedia Tools and Applications* 81, no. 5 (2022): 6259-6276.

[25]. Kumar, Manoj, and Hitesh Kumar Sharma. "A GAN-based model of deepfake detection in social media." *Procedia Computer Science* 218 (2023): 2153-2162.

[26]. Al-Dhabi, Yunes, and Shuang Zhang. "Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn)." In *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, pp. 236-241. IEEE, 2021.

[27]. PODDAR, RAUNAK. "DEEPFAKE VIDEO DETECTION: A MULTI-MODEL APPROACH USING CNN, RNN & LSTM." PhD diss., 2023.

[28]. Ojha, Utkarsh, Yuheng Li, and Yong Jae Lee. "Towards universal fake image detectors that generalize across generative models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480-24489. 2023

[29]. Hou, Yang, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Jianjun Zhao. "Evading DeepFake Detectors via Adversarial Statistical Consistency." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12271-12280. 2023

[30]. Rafique, Rimsha, Rahma Gantassi, Rashid Amin, Jaroslav Frnda, Aida Mustapha, and Asma Hassan Alshehri. "Deep fake detection and classification using error-level analysis and deep learning." *Scientific Reports* 13, no. 1 (2023): 7422

[31]. Dong, Shichao, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. "Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3994-4004. 2023

[32]. Choudhury, Deepjyoti, and Tapodhir Acharjee. "A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers." *Multimedia Tools and Applications* 82, no. 6 (2023): 9029-9045

[33]. Raza, Ali, Kashif Munir, and Mubarak Almutairi. "A novel deep learning approach for deepfake image detection." *Applied Sciences* 12, no. 19 (2022): 9820

[34]. Birunda, S. Selva, P. Nagaraj, S. Krishna Narayanan, K. Muthamil Sudar, V. Muneeswaran, and R. Ramana. "Fake Image Detection in Twitter using Flood Fill Algorithm and Deep Neural Networks." In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 285-290. IEEE, 2022

[35]. Dong, Shichao, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. "Explaining deepfake detection by analysing image matching." In *European Conference on Computer Vision*, pp. 18-35. Cham: Springer Nature Switzerland, 2022

[36]. Chang, Xu, Jian Wu, Tongfeng Yang, and Guorui Feng. "Deepfake face image detection based on improved VGG convolutional neural network." In *2020 39th chinese control conference (CCC)*, pp. 7252-7256. IEEE, 2020

[37]. Kesarwani, Ankit, Sudakar Singh Chauhan, and Anil Ramachandran Nair. "Fake news detection on social media using k-nearest neighbor classifier." In *2020 international conference on advances in computing and communication engineering (ICACCE)*, pp. 1-4. IEEE, 2020

[38]. Jain, Anjali, Avinash Shakya, Harsh Khatter, and Amit Kumar Gupta. "A smart system for fake news detection using machine learning." In *2019 International conference on issues and challenges in intelligent computing techniques (ICICT)*, vol. 1, pp. 1-4. IEEE, 2019

[39]. Bahad, Pritika, Preeti Saxena, and Raj Kamal. "Fake news detection using bi-directional LSTM-recurrent neural network." *Procedia Computer Science* 165 (2019): 74-82

[40]. Mishra, Sunny, Amit K. Shukla, and Pranab K. Muhuri. "Explainable Fuzzy AI Challenge 2022: Winner's Approach to a Computationally Efficient and Explainable Solution." *Axioms* 11, no. 10 (2022): 489.

[41]. Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE access* 6 (2018): 52138-52160.

[42]. Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." *arXiv preprint arXiv:2006.11371* (2020).

[43]. Hussain, Shehzeen, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples." In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3348-3357. 2021.