Journal of Chemical Health Risks

www.jchr.org

JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

Deep Learning–Based Fake Image Detection Using Transfer Learning: A Comprehensive Review

1Nisha Pathan, 2Unmukh dutta, 3Vani Agrawal

1,2MPCT, Gwalior,

3ITM University, Gwalior

(Received: 05 January 2026 Revised: 15 February 2026 Accepted: 05 March 2026)

KEYWORDS

Fake Image Detection, Deep Learning, Transfer Learning, GAN, CNN, Digital Forensics, Comparative Analysis

ABSTRACT:

The fast evolution of artificial intelligence has transformed the development of the digital content, allowing to create very realistic synthetic images and videos. Although these technologies have many legal uses, they have also given rise to the development of false visual content or deepfakes, which is highly dangerous to the integrity of information, social trust, computer security, and even investigations. Detecting these manipulated images has now become a crucial task and in some cases even standard algorithms based on handcrafted features has been found to be ineffective against advanced generative algorithms such as Generative Adversarial Networks (GANs) and diffusion-based algorithms. Transfer learning Deep learning-based methods, and in particular those that employ techniques of transfer learning, have become valuable solutions because they can extract discriminative features on small datasets at a lower cost. The given paper provides the in-depth discussion of the existing methods of fake image detection based on deep learning and transfer learning. It discusses popular ready-made convolutional neural network architectures, test datasets, metrics, and the current trends in research. Comparative studies point out the advantages and disadvantages of the current tools, and the paper also reveals key problems and sketches the way forward in creating robust, scalable and generalizable fake image detector systems that could help counter the emerging challenges in cyberspace.

1. Introduction

During the past several years, the creation of artificially created imagery has increased exponentially, mainly due to improvements in deep learning and generative models. Such methods as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models allowed computers to create photos that are frequently hard to distinguish between real photographs and those generated by the computer. Although there are many valid uses of these technologies in data augmentation, entertainment, and the creation of works of art, they can also be used to produce photorealistic fake images and deepfakes that harm information integrity, privacy, and digital trust. The abuse of such artificial images ranges between dishonesty and unfair election to identity theft, fraud, and damaged reputation, and there is urgency of finding a working detection tool [1].

Fig. 1 Deepfake and Real images[2]

Deepfakes and counterfeit images are created through deep learning models that are able to extract intricate patterns in data and recreate visual features with great fidelity. One of the most potent frameworks in this field has been GANs, a generator and a discriminator where the former is a model that generates images, and the latter that tries to differentiate between generated and real images. StyleGAN and ProGAN are the top contenders that have been able to generate high resolution images that are almost lifelike, and as a result, it has become more challenging to distinguish authentic and synthetic content using conventional forensic tools. Traditional image forensic techniques as well as machine learning algorithms generally utilize handdesigned features or

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

statistic incoherence in identifying image manipulation. Nevertheless, they perform poorly in the face of more advanced deep learning manipulations which leave only artifacts which can be detected. Consequently, the scientific community has shifted into deep learning-based detection systems which are able to learn discriminative features directly, yielding much better performance compared to traditional algorithms. In the current scenario, deep learning and transfer learning integration are the foundations of the majority of the more advanced fake image detecting system [3].

Transfer learning can be defined as the method of using a pretrained model that was initially trained on a large dataset (ImageNet) and updating it to a new and similar task. As it relates to the fake image detection, transfer learning allows an effective way of extracting features of images through the assistance of the pretrained convolutional neural networks (CNNs) backbone architectures. These trained models already possess rich hierarchical features (edges, textures, and object semantics) that can be fine-tuned to give the ability to differentiate fine differences between actual and fake images. Transfer learning can be especially beneficial in the situation where there is a small amount of labeled data, which is frequently the case in forensic datasets where it is difficult to gather as many annotated fake images as possible [4].

The recent methods take into consideration popular pretrained frameworks, including ResNet, Inception, Xception, DenseNet, and EfficientNet, among others. Recent literature has demonstrated that, transfer learning is much more accurate in its detection and generalization than training networks. As an example, it has been shown that EfficientNet models, finetuned using transfer learning, can perform classification at levels much higher than the state-of-the-art baselines on deepfake benchmark datasets. Likewise, comparative analysis of big datasets of people can show that transfer-learning models such as the DenseNet and Xception ones can achieve high precision and recall in the learning of fake versus real images. Although there is a significant improvement, the image detection of fake images is an ongoing issue. Detection systems need to be updated to handle images produced by hidden models and generative methods as the generative models improve in terms of their sophistication. Recent studies have suggested that conventional supervised classification models do not necessarily generalize to new generative algorithms, driving new detection strategies, which either insert generalized feature representations or applications of hybrid strategies. These methods tend to integrate transfer learning and domain adaptation, ensemble learning or multi-task learning to become robust to a large range of fake image sources [5].

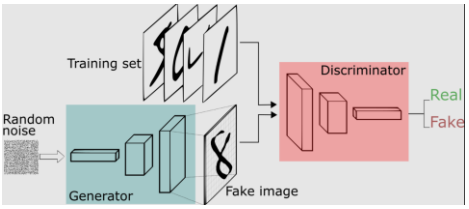

Fig. 2 Basic Structure of GANs Model[6]

The social effect of artificial images is not only limited to technology but also on ethics, legal matters and the state. In most jurisdictions, the ill intentional transmission of media which is manipulated can bear legal consequences especially when they encroach on the privacy rights and where they amount to defamation. As a result, effective detection systems are also needed by the automated systems as well as human moderators, journalists, and legal experts who will need to face the problem of manipulated media in the real world. In short, the analysis of fake image detection using deep learning and transfer learning is one of the most critical frontiers in the security of digital media. Using pre-trained neural networks and training them to use forensic datasets, the researchers have obtained significant advances in the accuracy of detection and generalization. Nevertheless, with the development of generative models and the increasing diversity of datasets, the current study of more robust and versatile detection models is critical. This is a review on important architecture, benchmark datasets, performance comparison, challenges and future research directions in this fast advancing area [7].

2. LITERATURE REVIEW

Ashani 2025 et.al To meet the digital world, the online image and video sharing has been expanding exponentially, and the technology of deepfakes has been developed, which is a product of the generative adversarial network (GAN) and the deep learning algorithm. This kind of technology allows producing extremely realistic manipulated videos and images that are extensively shared on social media, which are challenging to digital trust and information integrity. Understanding the duality of AI in creating and detecting deepfakes, Ashani et al. investigated how the convolutional neural networks (CNNs) can be used to detect a deep fake image. The experiment compared three CNNs VGG16, VGG19 and ResNet with a 1200 image set created with FaceApp. Findings indicated that VGG19 had the most accurate score of 98% which indicates greater capability of detecting manipulated images. This study demonstrates the usefulness of CNN-based AI tools in creating effective detection systems that

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

can reduce transmission of deepfake messages in the media [8].

Demir 2025 et.al Deepfake technology is based on artificial intelligence which creates extremely realistic images and videos that are closely similar to the original content, and has been brought to critical concern in terms of misinformation, identity theft, and reputational harm. Demir et al. examined the effectiveness of the transfer learning models used in the detection of deepfake, and it is essential that it is easy to identify the information with high reliability, and at the same time, it must happen quickly to ensure the safety of the information. The paper involved the use of popular transfer learning models, namely InceptionV3, EfficientNet, NASNet, ResNet, DenseNet, Xception and ConvNeXt on a large-scale publicly available dataset of 190,000 images consisting of both real and synthetic images. DenseNet showed the best test accuracy of 93 and was more accurate than other models. These results highlight the promise of the use of transfer learning in correctly identifying deepfakes, particularly in large and heterogeneous datasets, and offer a viable structure to the actual implementation of automated identification systems in practice to allow better checking of digital content and reducing the threats of deepfakes [9].

Tiwari 2025 et.al The fast evolution of generative adversarial networks (GANs) and other image generation algorithms has contributed immensely to the realism of the deepfaked content, which poses privacy, security, and digital trust challenges. The review of state-of-the-art detection schemes by Tiwari et al. included both handcrafted forensic schemes and deep learning-based models, with limitations being noted in terms of bias in the datasets, narrow cross-domain generalization, and ability to resist adversarial attacks. The authors came up with a hybrid CNN-based model to solve these problems, which involves block-based ResNet-50 to extract robust features and VGG-16 to classify. Assessment was done using publicly available datasets such as Celeb-DF and DeepFake Detection Challenge, (DFDC) which focus on practical deployment issues. The paper ended with important gaps in the research and directions on future research, such as multimodal detection systems, continuous learning, and explainable AI methods to enhance the interpretability and scalability of the deepfake detection systems into more reliable and flexible solutions [10].

Yadav 2025 et.al The emergence of the deepfake technology has proposed hyperrealistic edited photographs and videos, which pose significant threats to misinformation, identity theft, and the distrust of digital media. In their work, Yadav et al. suggested a hybrid deepfake detection model that combines both spatial and frequency domain features in order to enhance the rate of classification. The technique involves the multiscale convolutional neural networks (CNNs), frequency analysis, attention-based transformer networks, and ensemble learning to identify subtle manipulations in manipulated images. It was tested on 140,000 large-scale dataset with the result of 93% training and 88% testing accuracy. Further improvements in robustness and generalizability of adversarial training and sophisticated feature extractors were made to a variety of manipulation strategies. The article shows that hybrid and multidomain methods are essential to identify deepfakes, which may be important in practice, when the reliability and verifiability of media is essential [11].

Sharma 2024 et.al Deepfake images, particularly in the social media, create a big problem in differentiating between the manipulated images and authentic images. To solve this problem Sharma et al. suggested using GAN-CNN model with a generative replay mechanism to alleviate catastrophic forgetting problem of CNN which generally decreases performance in the context of transfer or constant learning. The method produces and repeats the samples of the previous task in new task training that contributes to the strength of the models in detecting deepfakes. The model had a better DCGAN trained to be more stable and had 98.67 accuracy on training and 70.08 accuracy on test datasets with equal precision, recall, and F1-scores. This paper will show that generative replay is effective in retaining previously acquired knowledge and updating it to new tasks, and is a reliable solution to deepfake detection in dynamically changing and manipulated real-world contexts that have new techniques of manipulating content [12].

TABLE 1 LITERATURE SUMMARY

Author & Year	Methodology	Findings	Research Gap	Limitations
Magesh et al., 2025 [13]	CSWin Transformer with cross-shaped window attention; compared with CNNs (MTCNN, InceptionV3, Xception) on Deep Fake Face Detection dataset	Achieved 98.7% accuracy and 98.72% F1-score; transformer captured subtle global and local features better than CNNs	Limited exploration of hybrid models combining CNNs and Transformers for improved efficiency	Focused only on single dataset; computational efficiency in large-scale deployment not fully analyzed

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

Pontorno et al., 2025 [14]	DeepFeatureX-SN (Siamese network + CNN classifier) using contrastive learning to identify real, GAN, and DM images	97.29% detection accuracy; 67.40% average generalization to unseen architectures; robust to manipulations	Needs more real-world deployment studies; adaptation to emerging GAN/DM models	High computational cost; complexity in training tripartite architecture
Alfraihi et al., 2025 [15]	ECMVFD-FTLTDO: Fusion-based transfer learning with ResNet50, MobileNetV3, EfficientNetB7	Achieved 95.26% (GRIP) and 92.67% (VTD) accuracy; improved copy-move forgery detection in videos	Limited generalization to diverse video sources and other forgery types	Requires frame extraction and preprocessing; high dependence on hyperparameter tuning
Tanfoni et al., 2024 [16]	Transfer learning using modified DeepLabV3+ with ResNet50 or MobileNetV3 for face segmentation and synthetic image detection	Targeted segmentation improved detection of StyleGAN2 faces; higher classification rates than global analysis	Only tested on StyleGAN2 images; limited evaluation on diverse datasets	Binary classification only; may not generalize to non-facial images
Mad Sahar et al., 2024 [17]	InceptionResNetV2 for video deepfake detection; compared with EfficientNetB0 and ResNet50; Android app development	Highest accuracy: 97.72% (CelebDF), 93.20% (DFDC); InceptionResNetV2 outperformed other models	Need for larger-scale mobile deployment evaluation	Dataset limited to DFDC and CelebDF; performance may vary on unseen datasets
Singh et al., 2024 [18]	Fine-tuned VGG16 with facial embeddings for real-time video deepfake detection	Effective detection with robust real-time video processing; preserved credibility of digital media	Limited dataset diversity; may not generalize to new deepfake techniques	Focused on facial videos only; CNN limitations in detecting non-local artifacts
Hafez et al., 2024 [19]	Ensemble CNN: ResNet50, DenseNet121, InceptionV3 combined via stacking + Logistic Regression; applied Error Level Analysis	Achieved 98.7% accuracy on 140k images; ensemble improved robustness and feature extraction	Need for evaluation on real-world mixed datasets	High computational cost; requires large labeled dataset for training
Zhang et al., 2024 [20]	X-Transfer: GAN detection using interleaved parallel gradient transmission; combined AUC and cross-entropy loss	Outperformed standard transfer learning; 99.04% best metric; generalized to non-face datasets	Limited exploration for video content; adaptation to evolving GANs	Complex architecture; performance may drop on highly imbalanced datasets

3. Objectives

FAKE IMAGE GENERATION TECHNIQUES: Due to the rapid development of deep learning, the generation of highly realistic fake images using the powerful generative models became possible. The models are trained to acquire the statistical distribution of natural images and produce the synthetic content that is highly similar to natural visual patterns. The three most notable types of fake image generating methods are the Generative Adversarial Networks (GANs), autoencoders-based models, and diffusion models [21].

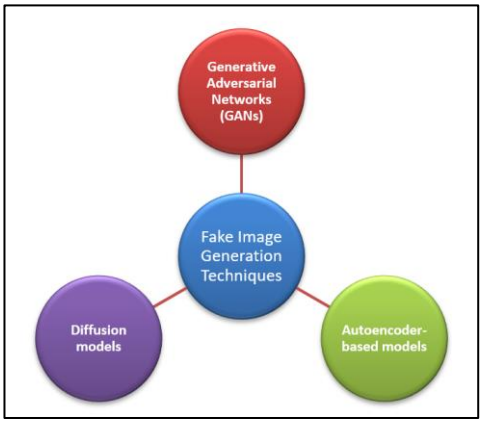

Fig. 3 Fake Image Generation Techniques

Generative Adversarial Networks (GANs) are the most commonly used frameworks in generating high-quality fake images. A GAN is composed of two opposing neural networks the generator and a discriminator. The generator is one that produces synthetic images and the discriminator is one that tries to distinguish between the real and synthetic images. In continuous adversarial training, the generator is gradually enhanced with capability to generate

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

photorealistic images. The superior GAN models like StyleGAN, ProGAN and CycleGAN have shown a stunning performance in high-resolution human faces, object images, and translation across domains. These models have the ability to maintain a high level of detail including lighting, texture, and facial expression that makes the created images look real and harder to notice with each release [22].

Autoencoder-based models are a different type of generative systems that are applied mostly in face swapping and reconstruction. Autoencoders reduce input images in latent representations and reassemble them to the original images. These models are able to swap identities of different people by training individual encoders and decoders using separate individuals, without altering facial expressions and head movements. This method has found extensive application in the early deepfake applications and has continued to be popular because of its relatively simple architecture and popularity in face manipulation tasks [23].

Diffusion models are a more recent approach to image synthesis methods that have recently become very popular. The models produce images by updating randomly added noise in a series of learned steps until it is reduced to weak noise. The diffusion models can generate highly realistic images with a high level of texture quality, sharpness and variety than the conventional GAN-based systems. The fact that they produce high fidelity images has also made the issue of fake image detection even more problematic. Taking all of these generative models, they learn more complicated visual distributions and generate synthetic images that are close to real data, which makes automated detection a more and more challenging research problem [24].

4. Methods

Transfer Learning in Fake Image Detection: The concept of transfer learning is an essential part of the contemporary fake image detector systems because deep learning models are able to use visual knowledge acquired in large-scale databases like ImageNet. Alternatively, instead of using raw data to train a model, pre-trained convolutional neural networks are trained on forensic databases to detect subtle information and anomalies in fake images. The method is much more efficient in detection, particularly where the data on labeled fake images is scarce [25].

A. Reduced Training Time

Transfer learning greatly saves time in total training that it takes to obtain fake image detecting models. Pre-trained networks do not need to be retrained since they already possess all the necessary low-level and mid-level visual features such as edges, color patterns and textures. The last layers only require fine-tuning to be used in classification tasks to achieve faster model convergence, reduce computational costs, and detect systems are deployed faster [26].

B. Improved Performance on Small Datasets

In digital forensics, the volume of labelled fake images that can be gathered is a challenge because of privacy issues and the dynamism of generative methods. Transfers learning enables models to provide high accuracy with little data, through the use of generalized representations, which have been trained on large amounts of data. This set of pre-learned features assists in capturing the minor artifacts in the image and in any irregularity of the texture that is usually added in the synthetic image generation procedures [27].

C. Better Generalization Capability

The initial training of pre-trained deep learning models is exposed to a large set of image distributions. This allows them to come up with strong and generalizable feature representations that can generalize well to novel and unknown fake image generation methods. Consequently, this means that transfer learning-based detection systems do not suffer performance instability when faced with new manipulation strategies or even new outputs of generative models [28].

D. Use of VGGNet and ResNet Architecture

One of the first and most universal convolutional neural networks that are employed in fake image detection is VGGNet. It is successful in fine-grained texture patterns that are found in manipulated images because of its uniform structure and small convolutional filters. Nevertheless, VGGNet has many parameters that present a high memory usage and computational expense, which could restrict its application in real-time detection systems. ResNet adds residual connections, which can address the problem of vanishing gradients and allow training very deep networks. The architecture enables models to learn discriminative features that are important in detecting small inconsistencies in fake images. Transfer learning models based on ResNet have shown high efficiency in the detection of deepfake images in

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

various data sets and are still in use in forensic image classification [29].

E. Use of Inception and MobileNet Architecture

Inception architectures use multi-scale convolutional filters in the same network layer to enable the model to gain fine and coarse spatial features. This multi-scale element-removal step enhances the capacity to identify the areas of manipulation, lighting anomaly and texture inconsistencies that occur in the fake photos. Transfer learning models based on inception have demonstrated improved classification on complicated forensic data. MobileNet is a lightweight and computationally efficient design and is capable of running in mobile and edge devices. It has depthwise separable convolutions that make this model very compact in terms of model size and inference time, with decent detection performance. This renders the MobileNet-based fake image detector systems as the most suitable in real-time applications as a social media moderation system and mobile authentication [30].

F. Use of EfficientNet and DenseNet

EfficientNet uses compound scaling to evenly scale the network depth, width, and resolution in order to obtain high accuracy with fewer parameters. DenseNet ties all the layers to all the other layers, enhancing the flow of features and improving the flow of gradients. A combination of these architectures offers stable learning, less overfitting, and better detection performance, and it is therefore an effective use of these architectures in fake image detection models based on transfer learning [31].

I. BENCHMARK DATASETS FOR FAKE IMAGE DETECTION

The design and testing of deep learning-based fake image detectors requires the access to various and quality data. Benchmark datasets offer normalized sources of both genuine and manipulated images, which allows researchers to train, test, and compare detection models across. Such datasets are required to determine the strength of detection systems on various types of images, manipulation modes, and resolutions. They also assist in knowing the way models can be generalized to new or unknown fake image generation mechanisms [32].

TABLE 2 BENCHMARK DATASETS

Dataset Name	Description	Size
CelebA-HQ	High-quality real human face images	30,000
FFHQ	Flickr-Faces-HQ dataset	70,000
FaceForensics++	Real and manipulated face images/videos	1M+
DeepFake Detection Challenge (DFDC)	Real and deepfake images/videos	470,000+
GANFake Dataset	GAN-generated fake face images	200,000

CelebA-HQ database consists of high-resolution images of humans faces that can be used in image synthesis and detection, which forms the basis on which subtle facials can be assessed. The FFHQ data presents a wide range of faces in Flickr, having differences in age, ethnicity, and pose, and also improves the generalization of the model under demographic variation [33]. One of the most commonly used datasets is FaceForensics++ that contains more than one million genuine and manipulated photos and videos, so it is very beneficial to train deep learning models. DeepFake Detection Challenge (DFDC) dataset concerns real-life deepfake videos gathered in various sources to allow researchers to compare models with other modern techniques of deepfake generation. Finally, GANFake consists of images created with the help of GAN models specifically, and it aids the models in learning and understanding artificial patterns peculiar to generative networks [34].

As a whole, these datasets represent a wide range of fake image situations, such as varying resolutions, generation methods and real-world variability. They are an important resource in designing, evaluating and comparing transfer learning-based fake images detectors, both in terms of accuracy and generalization. Using these benchmark datasets, scientists are able to create effective detection frameworks that have the ability to detect a variety of synthetic content with reduced false positives and enhanced reliability in practice [35].

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

II. TRANSFER LEARNING MODELS

Transfer learning takes advantage of the already trained convolutional neural net (CNN) to enhance the performance of detecting fake images especially where there is limited labeled data. It has used different CNN models, and the models themselves have their strong and weak points. Table provides a summary of the properties of popular pre-trained networks, including depth of the architecture, benefits, and shortcomings [36].

TABLE 3 COMPARISON OF PRE-TRAINED CNN MODELS

Model	Architecture Depth	Advantages	Limitations
VGG16	16 layers	Simple and stable	High memory usage
ResNet50	50 layers	Solves vanishing gradient problem	Computationally heavy
InceptionV3	48 layers	Efficient multi-scale feature extraction	Complex architecture
MobileNetV2	53 layers	Lightweight and fast	Slightly lower accuracy
EfficientNet-B4	Compound scaling	High accuracy and efficiency	Needs fine-tuning

VGG16 is also an easier and more stable framework, which can be used in the extraction of features in fake image detection, but this model demands extensive memory and computational power. ResNet50 solves the disappearance of the gradient issue by applying residual connections and thus, allows extremely deep networks with high accuracy at the expense of higher computational complexity. InceptionV3 represents multi-scale spatial information well, enhancing the detection capabilities of subtle artifacts, however, its architecture is complicated to execute. MobileNetV2 is also small and can be deployed in real-time applications, which is why it is applicable to mobile devices, but the accuracy is reduced a bit [37]. EfficientNet-B4 balances both network depth and width and resolution by scaling the compound, and it is as efficient and accurate as possible, but it must be fine-tuned to work optimally. Such comparative analysis shows that the choice of an appropriate CNN model is determined by the trade-off between accuracy, computational cost and deployment requirements. Models such as EfficientNet-B4 and ResNet50 are more appropriated when one wants to detect with high accuracy whereas MobileNetV2 is more pertinent in resource constrained settings [38].

III. CHALLENGES AND LIMITATIONS

Regardless of the greater progress made in the field of deep learning-based fake image detection, there are a number of key problems that continue to exist. The first problem is the active development of generative models, in particular, GANs and diffusion networks. Recent architecture, including StyleGAN3 and latent diffusion models, is capable of generating ultra-realistic images with extremely fine textures, facial expressions and lighting. Older datasets used to train models tend to miss these advanced fakes and as such, constant training and retraining of models is required to ensure that the model remains reliable [39].

The other problem is the bias and issues of generalization of data sets. Numerous benchmark datasets are large, but in terms of diversity, they are limited in terms of demographics, image resolutions, lighting, and the types of manipulation. Models that are trained using these datasets can be overfitted to a set of patterns and therefore poorly predict previously unseen data in a real-world context. This restricts the feasibility of existing methods and highlights the importance of more representative data. Deep learning models are also relatively expensive to compute. Big CNN models, hybrid models or ensemble models demand vast memory and processing power to train and to infer. This limits scalability, real time deployment and running on edge devices like smart phones or surveillance cameras [40].

There is another concern among adversarial attacks. Even with the slightest perturbation of images, detection models may be fooled into misclassifying the image. These vulnerabilities allow security risks especially in social media moderation or legal forensic investigations. Lastly, there are no studies of actual deployment. The majority of the studies have been done on controlled datasets under perfect conditions and few studies have been done concerning performance in dynamic, uncontrolled environments. Images of mixed quality, alternate camera origins and invisible manipulation methods raise serious practical issues that have to be resolved before their widespread use [41].

Journal of Chemical Health Risks

www.jchr.org

JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

5. Results

Future Research Directions: In order to address these limitations, a number of research directions are coming up. Multi-modal detection frameworks have potential, in that they combine visual with other modalities, e.g., audio, video or metadata. This is able to display inconsistencies that might be missed in image only methods, which make it more robust in real life situations [42].

The hybrid CNN-Transformer models, who integrate the local feature extraction power of CNNs and the global context availability of Transformers, were created. These architectures are better at detecting fine, high-resolution artifacts, especially in complex images when manipulation can be large. It is important to develop lightweight mobile and edge models. Model pruning, quantization and efficient architecture such as MobileNet or EfficientNet are some of the techniques that can be used to achieve near real-time detection with a sufficient level of accuracy, making it easy to apply the techniques at scale [43].

Another major area is continuous learning strategies, which enable models to be updated on new generating methods in a gradual way without losing the previously learned patterns. This method will make sure that the detection systems keep up with the synthesis of images generated by artificial means [44]. Finally, there is the need to have Explainable AI (XAI) when it comes to forensic transparency. The benefits of XAI methods include identifying areas of manipulation and having an interpretable decision explanation, which can justify human checking, legal responsibility, and trust of the system among the population regarding automated systems. Combining these issues and these research directions will ensure that fake image detection systems become more resilient, flexible, and able to work in the real world [45].

6. Discussion

The blistering development of deep learning has altered the situation in the sphere of digital media, allowing generating extremely realistic fake images and deepfakes with the help of GANs, autoencoders, and diffusion models. These attacks have a serious risk on privacy, security, and information integrity, as well as, there is an urgent necessity of powerful detecting systems. Transfer learning has become an effective methodology in this field, where the use of discriminative features can be obtained efficiently using a pre-trained convolutional neural network even when using small labeled datasets. The tuning of models, including ResNet, EfficientNet, DenseNet, and MobileNet, has enabled researchers to obtain high-performance detection on benchmark datasets at low computational cost and training time. Comparative studies have shown that the choice of a model should find a compromise between accuracy, scalability and deployment limits, and that hybrid and ensemble models can also add extra robustness. Although these advances exist, there are still issues such as the fast development of generative models, dataset bias, high computational costs, adversarial attacks, and the lack of studies on the real implementation. The directions to follow in the future are in multi-modal detection frameworks, CNNTransformer, lightweight CNN-based architectures designed to support mobile devices, continuity-based learning strategies, and explainable AI to build transparency and trust. In general, the concept of fake image detection with the help of deep learning and transfer learning is an important aspect of the digital forensics field that provides effective mechanisms to prevent the synthetic media, yet the necessity to continue research and ensure flexibility in the response to more specific image creation methods remains.

Refrences

[1] B. B. G. Khoo, C. H. Lim, and R. C.-W. Phan, “Transferable Class-Modelling for Decentralized Source Attribution of GAN-Generated Images,” no. Figure 1, 2022, [Online]. Available: http://arxiv.org/abs/2203.09777

[2] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer, “Adversarial threats to deepfake detection: A practical perspective,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 923–932, 2021, doi: 10.1109/CVPRW53098.2021.00103.

[3] S. A. Fezza, M. Y. Ouis, B. Kaddar, W. Hamidouche, and A. Hadid, “Evaluation of Pre-Trained CNN Models for Geographic Fake Image Detection,” *2022 IEEE 24th Int. Work. Multimed. Signal Process. MMSP 2022*, 2022, doi: 10.1109/MMSP55362.2022.9949282.

[4] Y. Jeong, D. Kim, Y. Ro, and J. Choi, “FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations,” *Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022*, vol. 36, pp. 888–896, 2022, doi: 10.1609/aaai.v36i1.19990.

[5] A. H. Khalifa, N. A. Zaher, A. S. Abdallah, and M. W. Fakhr, “Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition,” *IEEE Access*, vol. 10, pp. 22678–22686, 2022, doi: 10.1109/ACCESS.2022.3152029.

[6] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, “DeepFake Detection for Human Face Images and Videos: A Survey,” *IEEE Access*, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[7] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, “Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples,” *Proc. - 2021*

Journal of Chemical Health Risks
www.jchr.org
JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

IEEE Winter Conf. Appl. Comput. Vision, WACV 2021, pp. 3347–3356, 2021, doi: 10.1109/WACV48630.2021.00339.

[8] Z. N. Ashani, I. S. C. Ilias, K. Y. Ng, M. R. K. Ariffin, A. D. Jarno, and N. Z. Zamri, "Comparative Analysis of Deepfake Image Detection Method Using VGG16, VGG19 and ResNet50," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 47, no. 1, pp. 16–28, 2025, doi: 10.37934/araset.47.1.1628.

[9] L. E. Demir and Y. Canbay, "Deepfake Image Detection with Transfer Learning Models," *Bitlis Eren Üniversitesi Fen Bilim. Derg.*, vol. 14, no. 1, pp. 546–560, 2025, doi: 10.17798/bitlisfen.1610300.

[10] M. R. Tiwari and S. S. Patil, "Deepfake Image Detection: A Review of Existing Methods and a Hybrid CNN-Based Proposed Framework," *EPJ Web Conf.*, vol. 341, p. 01043, 2025, doi: 10.1051/epjconf/202534101043.

[11] U. Yadav, P. Dasarwar, S. Bondre, and S. Kalamkar, "A Hybrid Approach for Robust Deep Fake Image Detection using Spatial and Frequency Domain Features," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 3, pp. 22786–22791, 2025, doi: 10.48084/etasr.10458.

[12] P. Sharma, M. Kumar, and H. K. Sharma, "GAN-CNN Ensemble: A Robust Deepfake Detection Model of Social Media Images Using Minimized Catastrophic Forgetting and Generative Replay Technique," *Procedia Comput. Sci.*, vol. 235, no. 2023, pp. 948–960, 2024, doi: 10.1016/j.procs.2024.04.090.

[13] A. P. Magesh, S. S. M. Ramakrishnan, R. Arumuga Arun, N. Priyanka, and M. N. Kartheek, "Building an efficient Deep Fake detection system using the recognition capabilities of convolutional neural networks and transformers," *Discov. Comput.*, vol. 28, no. 1, 2025, doi: 10.1007/s10791-025-09586-2.

[14] O. Pontorno, L. Guarnera, and S. Battiato, "DeepFeatureX-SN: Generalization of deepfake detection via contrastive learning," *Multimed. Tools Appl.*, pp. 47721–47740, 2025, doi: 10.1007/s11042-025-21060-1.

[15] H. Alfraihi *et al.*, "A multi-model feature fusion based transfer learning with heuristic search for copy-move video forgery detection," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-88592-2.

[16] M. Tanfoni, E. G. Ceroni, N. Pancino, M. Bianchini, and M. Maggini, "Facial Segmentation in Deepfake Classification: a Transfer Learning Approach," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 4160–4168, 2024, doi: 10.1016/j.procs.2024.09.255.

[17] N. Mad Sahar *et al.*, "Advances in DeepFake Detection: Leveraging InceptionResNetV2 for Reliable Video Authentication," *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 1, no. 1, pp. 90–105, 2024, doi: 10.37934/araset.62.1.90105.

[18] A. V. Singh, D. Moghe, and K. Meenakshi, "Deepfake detection using fine-tuned VGG16 model: A transfer learning approach," *Challenges Information, Commun. Comput. Technol.*, pp. 825–829, 2024, doi: 10.1201/9781003559085-141.

[19] M. M. Hafez, M. Mohamed, Y. Hesham, M. M. Gomaa, G. Sadek, and A. Amer, "Deepfake Image Forensics: Towards Efficient and Reliable Detection," *Inf. Sci. Lett.*, vol. 13, no. 2, pp. 341–349, 2024, doi: 10.18576/isl/130212.

[20] L. Zhang *et al.*, "X-Transfer: A Transfer Learning-Based Framework for GAN-Generated Fake Image Detection," *Proc. Int. Jt. Conf. Neural Networks*, 2024, doi: 10.1109/IJCNN60899.2024.10650566.

[21] A. S. Al-Qazzaz, P. Salehpour, and H. S. Aghdasi, "Robust DeepFake Face Detection Leveraging Xception Model and Novel Snake Optimization Technique," *J. Robot. Control*, vol. 5, no. 5, pp. 1444–1456, 2024, doi: 10.18196/jrc.v5i5.22473.

[22] N. Qazi and I. Ahmed, "Enhancing Authenticity Verification with Transfer Learning and Ensemble Techniques in Facial Feature-Based Deepfake Detection," *2024 14th Int. Conf. Pattern Recognit. Syst. ICPRS 2024*, 2024, doi: 10.1109/ICPRS62101.2024.10677831.

[23] S. A. Khan and D. T. Dang-Nguyen, "Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms," *IEEE Access*, vol. 12, no. December 2023, pp. 1880–1908, 2024, doi: 10.1109/ACCESS.2023.3348450.

[24] N. N. Zanje, A. M. Bongale, and D. Dharrao, "Detecting facial image forgeries with transfer learning techniques," *Int. J. Adv. Appl. Sci.*, vol. 13, no. 1, pp. 93–105, 2024, doi: 10.11591/ijaas.v13.i1.pp93-105.

[25] M. Khalid *et al.*, "Novel Sentiment Majority Voting Classifier and Transfer Learning-Based Feature Engineering for Sentiment Analysis of Deepfake Tweets," *IEEE Access*, vol. 12, no. February, pp. 67117–67129, 2024, doi: 10.1109/ACCESS.2024.3398582.

[26] W. A. Asha, N. Saba, S. M. Huq, and M. I. Hossain, "Deepfake Detection Using Neural Networks," *2024 27th Int. Conf. Comput. Inf. Technol. ICCIT 2024 - Proc.*, vol. 9, no. 3, pp. 3523–3528, 2024, doi: 10.1109/ICCIT64611.2024.11022076.

[27] S. Ali Raza, U. Habib, M. Usman, A. Ashraf Cheema, and M. Sajid Khan, "MMGANGuard: A Robust Approach for Detecting Fake Images Generated by GANs Using Multi-Model Techniques," *IEEE Access*, vol. 12, no. April, pp. 104153–104164, 2024, doi: 10.1109/ACCESS.2024.3393842.

[28] M. Usama Tanveer, K. Munir, B. Rathore, A. Alabdulatif, R. H. Jhaveri, and M. Fatima, "Neuro-VGNB: Transfer Learning-Based Approach for Detecting Brain Stroke," *IEEE Access*, vol. 12, no. September, pp. 178862–178874, 2024, doi: 10.1109/ACCESS.2024.3490693.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2026) 16(2), 377-386 | ISSN:2251-6727

[29] R. M. Alnafea, L. Nissirat, and A. Al-Samawi, "CNN-GMM approach to identifying data distribution shifts in forgeries caused by noise: a step towards resolving the deepfake problem," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.1991.

[30] E. Şafak and N. Barışçı, "Detection of fake face images using lightweight convolutional neural networks with stacking ensemble learning method," *PeerJ Comput. Sci.*, vol. 10, pp. 1–20, 2024, doi: 10.7717/PEERJ-CS.2103.

[31] A. Lopez Pellcier, Y. Li, and P. Angelov, "PUDD: Towards Robust Multi-modal Prototype-based Deepfake Detection," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 3809–3817, 2024, doi: 10.1109/CVPRW63382.2024.00385.

[32] G. Naskar, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar, "Deepfake detection using deep feature stacking and meta-learning," *Heliyon*, vol. 10, no. 4, p. e25933, 2024, doi: 10.1016/j.heliyon.2024.e25933.

[33] U. Kosarkar, G. Sarkarkar, and S. Gedam, "Revealing and Classification of Deepfakes Video's Images using a Customize Convolution Neural Network Model," *Procedia Comput. Sci.*, vol. 218, pp. 2636–2652, 2022, doi: 10.1016/j.procs.2023.01.237.

[34] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, *Deepfake video detection: challenges and opportunities*, vol. 57, no. 6. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10810-6.

[35] S. Karim, X. Liu, A. A. Khan, A. A. Laghari, A. Qadir, and I. Bibi, "MCGAN—a cutting edge approach to real time investigate of multimedia deepfake multi collaboration of deep generative adversarial networks with transfer learning," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-80842-z.

[36] V. Dudykevych, S. Yevseiev, H. Mykytyn, K. Ruda, and H. Hulak, "Detecting Deepfake Modifications of Biometric Images using Neural Networks," *CEUR Workshop Proc.*, vol. 3654, pp. 391–397, 2024.

[37] S. Boovaneswari, A. Divya, and A. Harisri, "Efficiently Identifying Fake Audio and Images Using Transfer Learning," *2024 Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2024*, 2024, doi: 10.1109/ICSCAN62807.2024.10894602.

[38] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 14, no. 2, pp. 1–45, 2024, doi: 10.1002/widm.1520.

[39] H. Song, S. Huang, Y. Dong, and W.-W. Tu, "Robustness and Generalizability of Deepfake Detection: A Study with Diffusion Models," pp. 1–8, 2023, [Online]. Available: http://arxiv.org/abs/2309.02218

[40] S. Suratkar and F. Kazi, "Deep Fake Video Detection Using Transfer Learning Approach," *Arab. J. Sci. Eng.*, vol. 48, no. 8, pp. 9727–9737, 2023, doi: 10.1007/s13369-022-07321-3.

[41] A. Aghasanli, D. Kangin, and P. Angelov, "Interpretable-through-prototypes deepfake detection for diffusion models," *Proc. - 2023 IEEE/CVF Int. Conf. Comput. Vis. Work. ICCVW 2023*, pp. 467–474, 2023, doi: 10.1109/ICCVW60793.2023.00053.

[42] Y. Patel *et al.*, "An Improved Dense CNN Architecture for Deepfake Image Detection," *IEEE Access*, vol. 11, no. February, pp. 22081–22095, 2023, doi: 10.1109/ACCESS.2023.3251417.

[43] Y. Patel *et al.*, "Deepfake Generation and Detection: Case Study and Challenges," *IEEE Access*, vol. 11, no. December, pp. 143296–143323, 2023, doi: 10.1109/ACCESS.2023.3342107.

[44] W. H. Abir *et al.*, "Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods," *Intell. Autom. Soft Comput.*, vol. 35, no. 2, pp. 2151–2169, 2023, doi: 10.32604/iasc.2023.029653.

[45] A. Saxena *et al.*, "Detecting Deepfakes: A Novel Framework Employing XceptionNet-Based Convolutional Neural Networks," *Trait. du Signal*, vol. 40, no. 3, pp. 835–846, 2023, doi: 10.18280/ts.400301.

1.