Al-Rafidain Journal of Engineering Sciences

Journal homepage https://rjes.iq/index.php/rjes

ISSN 3005-3153 (Online)

A Review of Machine Learning Classifiers for Feature-Based Image Forgery Detection

Maryam Kareem Khudair, Omar Munthir AlOkashi

University of Anbar, College of Computer Science and Information Technology, Department of Computer Science

ARTICLE INFO

Article history:
Received 6 March 2026
Revised 6 March 2026
Accepted 9 April 2026
Available online 10 April 2026

Keywords:
Digital image forensics,
Deepfake images,
Diffusion-based synthetic images,
Machine learning,
GAN-based forgery

ABSTRACT

The increasing use of digital image manipulation and the generation of synthetic images has led to serious concerns regarding the authenticity and trustworthiness of images. The recent development of advanced generative models like Style-GAN, StyleGAN2, and diffusion models has the potential to generate very realistic synthetic images that are very difficult to distinguish from real images, thus leading to serious challenges in digital image forensics. This paper presents a comprehensive review of image forgery detection techniques with a primary focus on machine learning-based techniques. Various image forgery techniques, including traditional image forgery and AI-created synthetic images, are presented to understand their nature and challenges in image forgery detection. The paper presents a review of machine learning-based methods, popular datasets, and evaluation metrics for image forgery detection. Furthermore, the paper presents the challenges of existing methods, particularly their poor generalization capability for diffusion-based forgeries and the lack of representative benchmarks, and concludes with open challenges and future research avenues for developing generalized machine learning methods for authentic detection of image forgeries.

1. Introduction

Digital images have greatly replaced conventional photographs over the past ten years and have become an essential tool for information dissemination in the communication of today, including newspapers, websites, and social media [1]. The use of digital images has also become more apparent in the field of computer forensics. Although there have been major developments in digital image processing, which have resulted in the emergence of various forensic methods, they have also made image manipulation simpler and more accessible. Because of this, image security has become a major concern in every industry that uses digital photographs. Tampered images, including criminal images, crime scenes, biometric data, and other forensic details, have been employed in forensic analysis for a considerable period of time [2][3]. A digital image can provide a quantitative representation of real-world phenomena; However, the manipulation of digital images has become a simple process for anyone, even without technical expertise, because of the availability of image editing software on common devices such as smartphones and tablets. In this case, visual information can be easily manipulated or merged to create very convincing forged images that can fool even expert viewers [4].

The process of altering the components of an image with the aim of deceiving or misleading is referred to as digital image manipulation [5].

Corresponding author E-mail address: omar.alokashi@uoanbar.edu.iq
https://doi.org/10.61268/wwmhw250

Digital image manipulation is also known by other names, such as image tampering, image manipulation, or forgery. Image modification has been used for ages and is not a fresh occurrence. The government and authorities have been made aware of numerous instances of image alteration in the past that have caused public concern [6]. At present, there are many image manipulation software packages available in the market, such as Adobe Photoshop, Paint-Slinger, and the GNU Image Manipulation Program (GIMP). Although some of these software packages are free and others are commercial, they are relatively inexpensive and simple to use. Furthermore, the images manipulated by these software packages also go through several post-processing and image enhancement tasks, making the manipulated images appear very realistic. Consequently, the human visual system is often unable to distinguish between original and manipulated images with the naked eye. This makes the digital image highly vulnerable, thus making digital images less trustworthy [7]. This field of research has been able to attract a large number of scholarly publications around the world. Between 2001 and 2019, an analysis was conducted to use two of the biggest academic databases to determine the annual number of publications Regarding digital image forensics: Elsevier (ScienceDirect) and IEEE (IEEE Xplore), as illustrated in Figure 1. The gradual increase in the number of publications indicates the growing significance of research in image forgery detection.

This work aims to give a thorough and organized evaluation of machine learning classifiers for feature-based image fraud detection. This paper will specifically concentrate on the analysis and classification of existing image forgery techniques, such as picture splicing, retouching, and copy-move image counterfeiting, and review the existing handcrafted feature extraction methods used in forgery detection.

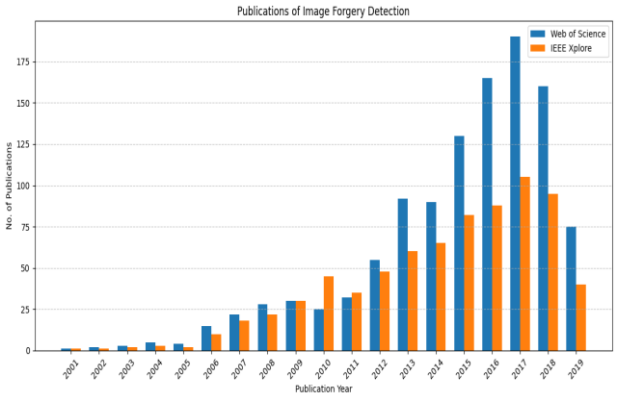

Figure 1. The quantity of articles in digital image forensics that occur per year [9]

In addition, this paper will also attempt to review and compare the commonly used machine learning classifiers, such as Support Vector Machines, k-Nearest Neighbors, Random Forest, and Artificial Neural Networks, based on their performance, robustness, and computational complexity. The publicly available datasets and metrics will also be reviewed to ensure a consistent comparison framework. Through this review, the paper hopes to provide useful insights and future research directions for researchers and practitioners in the forensics of digital images sector.

2. Overview of Image Forgery Techniques

The more general term "image forgery" refers to the process of manipulating or changing digital photographs in order to produce misleading information, fraud, deceit, or misinformation. Recently, picture forging has become more complex and difficult to detect due to the rising use of sophisticated picture editing programs and models for image synthesis powered by AI [10]. From basic copy-move picture forgery to incredibly lifelike deep-fake images that can accurately replicate visual content seen in the actual world, the level of forgery varies. Since most forged photos have few visual abnormalities and may successfully fool both human vision and traditional detection techniques, this has

significantly increased the challenges faced by digital image forensics. Depending on the type of manipulation and the technology used, image forgeries can be categorized into two major types, which are quite challenging and require different approaches for detection. The first type of image forgery is the traditional manipulation-based forgery, where existing images are manipulated using operations such as copying, splicing, retouching, filtering, or compressing. In the second kind of picture forgeries, which are created using artificial intelligence (AI) technology, Deep generative models, such diffusion models and Generative Adversarial Networks (GANs), are used to create or modify images. The images created using this technology are very realistic and do not have the usual signs of manipulation, making them very difficult to detect. Figure 2 shows the broad categories of image forgery, emphasizing the difference between conventional image manipulation and the generation of synthetic images using AI. The below classification offers a systematic approach to understanding the various types of image forgery and will be used as a reference point for the challenges and approaches involved in image forgery detection.

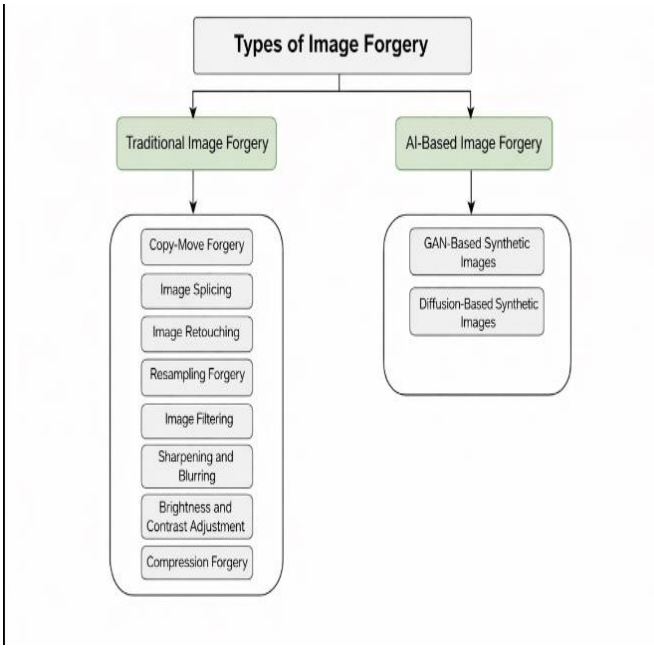

Figure2. Types of Image Forgery

2.1 Copy and Move Forgery

This type includes copy-move forgery, or CMF for short, where a part of the image is copied to another part of the image with the aim of concealing or altering that particular sensitive information. This includes removing unnecessary objects and replicating features [11]. Since the modified area is derived from the same image, it preserves the same noise patterns, textures, and lighting. To make the detection more challenging, the modifications frequently include adjustments like compression, blurring, scaling, or rotation. Social media, news media, and forensic document changes all frequently use it. Furthermore, when forgers apply alterations to their work, including rotation, compression, or rescaling, conventional techniques become difficult [12]. To understand fine-grained features of manipulation, deep learning-based detection techniques often need a lot of data [13]. SIFT and SURF are feature-based techniques that identify comparable patterns in various regions of an image [14]. Using block-matching techniques, by splitting the image into small chunks and comparing the pixel similarities, duplicate areas can be found [15]. For more accurate detection automation, CNNs and ViTs have been used [16]. When working with pattern-based images like those in Figure 3, it is clear how copy-move forgery can be used for forgeries that have faint or undetectable traces.

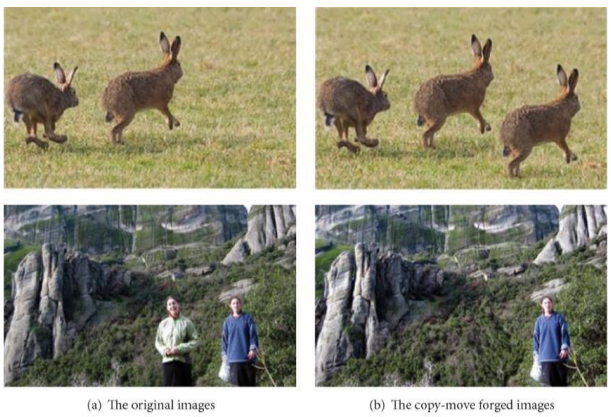

Figure3. The Forgery of Copy-Move [16].

2.2 Splicing Forgery

Splicing is the process of joining fragments of two or more distinct images to create a new image. Splicing is frequently used in advertising, social media manipulation, and political disinformation [17]. It results in uneven perspective, lighting, texture, and shadow which, usually calls for additional post-processing steps like filtering and blending to reduce visible seams and hide the manipulation. This type of forgery can be produced by AI utilizing GAN-based synthesis or manually (by a human editor). Boundary identification is challenging since spliced photos typically go through a retouching process. Because GAN-based splicing forgeries produce nearly flawless blending, they are nearly invisible to the naked eye [18]. By using edge filtering algorithms, edge and boundary detection techniques are able to detect abrupt changes. includes changes in an image's color histogram, shadow, and light [19]. Deep learning segmentation: FCNs and GANs have been tested to isolate parts of an image that have been altered [20]. Figure 4 provides an example of image splicing, with (a) representing the original image and (b) representing the spliced image. Two zebras from a third image are added to the spliced image. The spliced image has irregularities of some kind, like unusually sharp edges, as a result of the image manipulation procedure. According to this finding, feature extraction in image splicing detection methods need to be strong enough to spot issues in spliced images that aren't visible in typical Figure 4.

2.3 Morphing and Retouching

While morphing is combining two faces or things to create a hybrid image, retouching takes-into-account altering an image for aesthetic qualities like skin smoothing and object enhancement. Applications for this are numerous, ranging from identity fraud to the advertising and cosmetics sectors [22]. Malicious retouching is linked to concealing identity or flaws, while non-malicious retouching is used for enhancement purposes like smoothing the skin and contrast enhancement. Morphing-based forgeries are employed to forge documents, deceive biometric systems, and steal identities. Subtle variations alone make detection difficult without reference images. Facial morphing attacks have the potential to mislead biometric authentication systems and undermine security verification [23]. High-frequency characteristics, texture smoothing, and irregularities in color distribution can all be found via texture analysis [24]. FaceNet and CNN-based biometric models are used in AI-based face verification to distinguish between real and altered identities [21]. The authentic (original) image in Figure 5 is the one that is highlighted with a blue box in each row; the other photos in the same row are altered (attack) images.

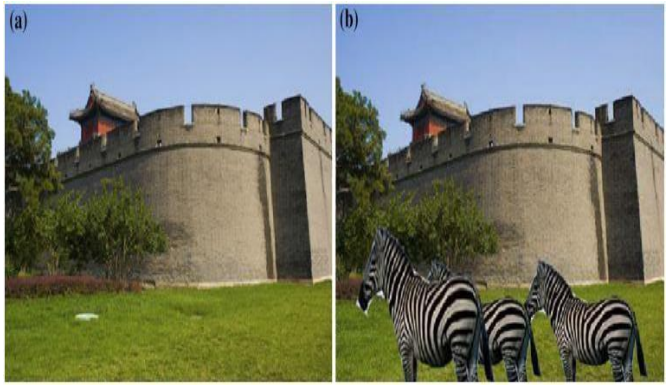

Figure4. An example of image splicing (a) original image and (b) altered image [21].

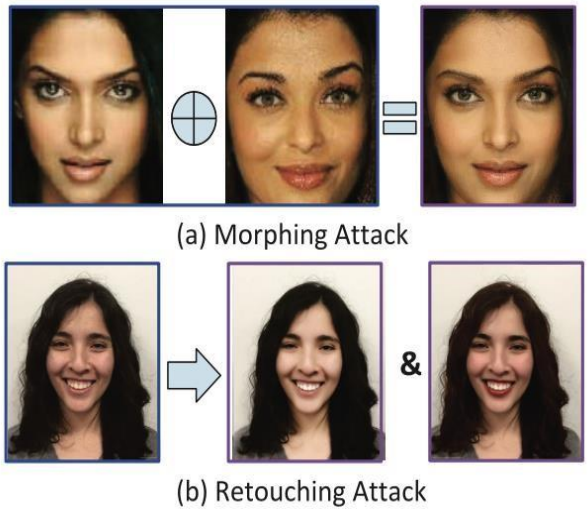

Figure5. Attacks using morphing and retouching [21].

2.4 GAN-Based Synthetic Images (Deepfake)

The main concept behind deep-fake technology is based on creating incredibly lifelike synthetic images and films that closely resemble actual individuals using Generative Adversarial Networks (GANs) and autoencoders [25]. These methods can be used to produce lifelike images and videos of real and non-existent individuals. Deep-fake technology can be used to manipulate facial expressions, body movements, and even voice in videos. Deep-fakes have been linked to various forms of cybercrimes such as identity theft, creation of fake social media profiles, spreading misinformation, and political propaganda. The ever-improving nature of deep-fake generation models has made traditional image and video forensics less effective, posing a challenge to digital image and video forensics [26]. GAN models [27] are particularly effective at producing hyper-realistic synthetic images—commonly referred to as *deep-fakes*—due to their adversarial architecture, in which the generator network is iteratively refined until the discriminator network can no longer distinguish between generated and real data [28]. Recent advancements in training strategies, network architectures, and data quality have further narrowed the perceptual gap between GAN-generated and authentic images, making detection increasingly difficult. Nevertheless, existing studies suggest that static images generated by GANs still retain subtle digital fingerprints related to their generative processes, which can be exploited for forensic detection.

A wide variety of GAN architectures have been proposed for face and body manipulation and forgery tasks. RSGAN [29] and StarGAN [30] focus on face replacement by manipulating facial and hair features, and facial attributes and expressions, respectively. GANimation [31] enables facial expression synthesis while preserving the original identity by transferring facial contours without distortion. PGGAN [32] is capable of generating high-resolution, hyper-realistic synthetic images, whereas BiGAN [33] produces forgeries that are particularly difficult to detect due to the incorporation of a random noise variable through an encoder alongside the standard GAN framework. To address generalization limitations and reduce computational overhead, CBiGAN [34] was introduced, achieving improved learning efficiency and lower power consumption. Furthermore, WGAN [35] has been shown to generate higher-quality outputs by enhancing training stability and improving the relationship between generator convergence and loss functions when compared to conventional GAN models [36]. Examples of faces produced using InterfaceGAN, ProGAN, StyleGAN3, and an actual image from the CelebA dataset are displayed in Figure 6.

Figure6. Examples of faces produced using InterfaceGAN, ProGAN, StyleGAN3, and an actual image from the CelebA dataset are shown from left to right [36].

2.5 Diffusion-Based Synthetic Images

Diffusion-based deepfakes are very different from previous deepfake generation methods in a number of key ways. Firstly, they are able to produce a level of visual fidelity that is very high, with face images containing a level of realistic details which is difficult to distinguish from real images, removing typical artifacts such as edge distortion, smearing, and biometric irregularities such as asymmetrical eyes or ears. Secondly, diffusion models are very diverse in the outputs they produce. When trained on a massive dataset such as LAION-5B, which consists of billions of real-world images sourced from a variety of online locations, they are able to produce faces in a variety of contexts and styles [37]. This

diversity allows for the production of artificial images that accurately depict the range of situations and appearances seen in the real world. Thirdly, diffusion-based deep-fakes have become very accessible, making it easier to create convincing deep-fakes. This is because diffusion-based deep-fake generation models do not require much technical knowledge to create convincing deep-fakes. This is opposed to previous deep-fake models that required a lot of technical knowledge to create convincing deep-fakes. Some of the popular diffusion-based models include Stable Diffusion [38] and MidJourney [39]. However, despite these developments, the fast-paced progress of diffusion-based deep-fake generation has led to a gap in the development of deep-fake detection research, resulting in several key challenges. The first challenge is the absence of specific datasets for the evaluation of deep-fake detection techniques against state-of-the-art diffusion-based deep-fakes. Current datasets, such as FaceForensics++ (FF++) [40] and CelebDF [41], were developed a few years ago and used outdated facial manipulation methods, making them inadequate for the evaluation of deep-fake detection techniques against diffusion-based deep-fakes. Another important challenge is related to the generalization ability of the existing deep-fake detection methods. Most of the existing deep-fake detection methods are evaluated in a controlled experimental setup, where the training and testing datasets are generated using the same set of manipulation techniques and domains. Although these methods are able to detect deep-fakes with high accuracy in a closed-world setup, they are not effective when used for diffusion-generated deep-fakes, which cover unseen domains and distributions of content. Recent works [42, 43] have shown that the existing deep-fake detection methods are not able to generalize well to unseen manipulation techniques and domains. Although some methods like domain adaptation and transfer learning have been proposed to handle this problem [44], their effectiveness is still limited and not sufficient for real-world applications. Figure 7 shows an example of a diffusion-generated synthetic face image that showing a high realism and consistency achieved by the modern diffusion models.

Figure 7. Compared to earlier deepfake datasets (c–f), diffusion-based deepfake samples (a–b) exhibit more visual realism and background content diversity [45].

3. Methods for Detecting Image Forgeries Based on Machine Learning

The traditional approach of forgery detection, which relies on statistical analysis and manual examination, is unable to identify complex image forgeries due to the rapid advancement of image manipulation technologies. As a result, methods based on machine learning and deep learning have been developed that may use intricate patterns and artifacts in the faked photos to automate the forgery detection process. Digital forensics, media authentication, and cybersecurity applications place a great deal of attention on machine learning approaches because they offer greater accuracy and flexibility than traditional methods [10]. Machine Learning and Deep Learning Techniques such as: CNNs, ViTs, and their combination are used in deep learning-

based forgery detection to automate the detection of faked images [25]. They are used for forensic analysis, deep-fake detection, and AI-generated picture authentication. aids in the identification of AI-generated material and GAN-generated photos. Large volumes of label data are required for deep learning techniques, which makes their practical implementation challenging. Forgery detection methods become outdated over time due to the ongoing evolution of AI-generated forgeries [26]. The high-frequency elements of AI-generated images are identified via Fourier Transform-based GAN Detection [46]. According to [47], traditional machine learning-based picture fraud detection techniques often employ handcrafted feature extraction to highlight the distinctions between authentic and manipulated images. These techniques often entail a two-step procedure. Finding unique or statistically unusual characteristics in the image is the first step in the feature extraction process. In the second step, photos are categorized as either authentic or fake using machine learning classifiers.

Several feature-based methods have been extensively used in this regard. Features based on the frequency domain, such as the Discrete Cosine Transform (DCT), have been used to examine the compression artifacts created during image forgery [11]. Features based on texture, such as Local Binary Patterns (LBP), have been used to examine the texture anomalies created during forgery or artificial image merging [12]. Wavelet transform features have also been used to detect sudden changes in frequency, which may be indicative of image forgery [13]. In copy-move image forgery detection, keypoint detection features such as SIFT and SURF are extensively used to detect copied areas in an image [14].

After extracting the features, different machine learning classifiers are used for forgery detection. Support Vector Machines (SVM) are used to classify images by finding the best hyperplanes to separate real and forged images [15]. k-Nearest Neighbors (k-NN) is used to detect forged images based on their similarity to existing patterns [16]. Decision Trees (DT) and Random Forests (RF) are used to classify manipulated images using hierarchical decision-making based on multiple feature criteria [17]. However, these machine learning-based methods have several limitations. The extracted features are dataset-dependent, which may result in poor generalization performance on different datasets and forgeries. Moreover, feature extraction is computationally expensive and requires manual tuning. In addition, traditional feature-based machine learning methods may fail to detect highly sophisticated forgeries, especially GAN-based and diffusion-based synthetic images, as there are no distinctive handcrafted features [18].

4. Related Work

To identify various forms of picture alteration and the unapproved spread of falsified photos, a number of methods have been put forth. Finding duplicated areas within the same image is specifically the main goal of copy-move forgery detection. Accurately identifying and matching visually comparable parts is the primary problem, particularly when those regions have undergone post-processing techniques like noise addition, compression, rotation, or scaling. A strategy to increase the DWT-based technique's detection efficiency was put forth by Zhang et al. [48]. The input images were first subjected to the block Discrete Wavelet Transform (DWT); then the relationship between wavelet coefficients across positions was described by the Markov features. Lastly, to differentiate between the real and spliced images, (SVM) was used. Experiments showed that, given the ideal block size and feature count, the features acquired in DWT had an 89% detection efficiency. Jaiswal et al. [49] used a combination of handcrafted features: (LBP), Discrete Wavelet Transform (DWT), and Histogram of Oriented Gradients (HOG), to create a feature vector from grayscale photos. A logistic regression model was then used to classify the collected feature vectors. To assess performance and guarantee accurate result estimation, a 10-fold cross-validation approach was used. Using logistic regression, the images were divided into two classes: spliced and non-spliced. 99.5%

accuracy was suggested in experimental findings on the CASIA II.0 dataset. To identify image splicing, Habibi and Hassanpour [50] suggested a method based on edge pixel color distribution analysis. To increase localization efficiency and reduce overall measurement time, a segmentation technique was employed. An accuracy of 97% was demonstrated by the experimental findings obtained using the Columbia Picture Splicing dataset. A new forgery detection method is presented in the publication by A. Parnak et al. [51]. By applying the classic Benford's rule. The Mean Absolute Deviation (MAD) characteristic is extracted using the proposed method. Benford's law is also applied to the expanded mantissa distribution feature vector. The final feature vector was created using a combination of Benford's law-based features and other statistical characteristics. Finally, to differentiate between real and fake photos, a (SVM) with three distinct kernel functions was used.

To address the identification of such manipulations, Solaiyappan and Wen [52] carried out an organized case study. In order to identify between modified and unmanipulated photos, they assessed how well a number of machine learning methods performed, including Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT). Every pre-trained model was adjusted to enhance classification performance before feature extraction. The experimental findings showed that tumor injection and tumor removal operations may be detected with almost perfect precision. A forensic machine learning framework for identifying copy-move frauds was presented by Archana and Jamkhandikar [53]. To improve detection accuracy and robustness, their method combines an effective classification strategy with appropriate feature selection. The proposed approach enhances the performance of a Multi-Support Vector Machine (M-SVM) classifier through the application of the Golden Jackal Optimization (GJO) algorithm to identify the optimal feature subset. The MICC-F2000 dataset, which contains 2000 images with 1300 genuine and 700 forged images, was applied to validate the proposed method. Based on the experimental results, the proposed approach demonstrated a detection performance with an accuracy of 99.47%, sensitivity of 97.01%, specificity of 96.51%, precision of 99.62%, and MCC of 96.39%, outperforming the existing methods of SSDAE-GOA-SHO, ConvLSTM, CNN, and two-branch CNN. Taking benefit of the anomalous facial biometric anomalies caused by face swap and lip sync methods, Norman and Farid [54] proposed a forensic machine learning approach to detect deep-fake talking head videos. The Deep-Speak v1.0 dataset, consisting of authentic and deep-fake videos without any overlap in identity between the method and the dataset, was employed to test the proposed approach by splitting the data into the training and test sets. XGBoost, a classification algorithm chosen for its ability to deal with features of different scales and identify non-linear relationships, was employed to identify and classify the 9-dimensional facial biometric features of the videos. The experimental results demonstrated high detection accuracy of 94.9% on the combined datasets and 99.1% on the Celeb-DF-v2 dataset. For the classification of handwritten digits, Singh and Bansal introduced a hybrid generative and classification model based on Generative Adversarial Networks (GANs) and conventional machine learning approaches [55]. The model was tested on the MNIST database, where a GAN was used to extract the latent features and generate new samples of handwritten digits. The generated samples were classified with a 96.67% classification accuracy using an SVM classifier that was trained on the original MNIST database. Table 1 below demonstrate a summary of the main works related to image forgery classification methods.

Table1. Summary of Image Forgery detection works

Ref	Target of Tampering	Methodology	Dataset	Warrants /Drawback	Accuracy
[48]	Spliced Images	SVM-RFE	CASIA v1.0, CASIA v2.0	Warrants:- Method to boost the detection efficiency of the DWT-based technique. Drawback:- Low achieved accuracy compared to the methods that are published at the same time.	89% accuracy
[49]	Spliced Images	(HoG LTE, DWT, and LBP)	CASIA II.0	Warrants:- Logistic regression of a machine learning classification technique was used to divide images into two classes, spliced and nonspliced images. An accuracy Drawback:- It does not give highly accurate results with all the data sets that are used.	Accuracy of 99.5% is gained with CASIA II.0 dataset. This method was implemented and gave wrong results with 59% accuracy
[50]	Spliced Images	Interquartile Range (IQR)	Columbia Image	Warrants:- Reliably distinguishing original from tampering edges increase localization efficiency and reduce scores on measures of time. Drawback:- The used dataset was a small one with no underlying data and also a very apparent spliced area. This high accuracy with such a small dataset may be due to fitting problems. In addition, the design may not ever work on highly qualified forgery	Accuracy of 97% with the Columbia dataset
[51]	Splicing Image	SVM	CASIA V1.0 and CASIA V2.0	Warrants:- Presents a novel forgery detection algorithm using combined features. Drawback:- Numerous databases were not used. Other types of forgery were not detected.	98.45
[52]	Copy-Move	Support Vector Machine, Random Forest, and Decision Tree	LIDC-IDRI (untampered) and CT-GAN (tampered)	Warrants: - Reaching a high accuracy rate Drawback:- The system has not been tested for various situations such as noise	99.0%
[53]	Copy-Move	Golden Jackal Optimization (GJO) for feature selection combined with Multi-Support Vector Machine (M-	MICC-F2000 (1300 authentic, 700 forged images)	Warrants: Optimized feature selection improves discriminative capability and classification accuracy; robust performance compared to deep learning methods. Drawbacks: Requires handcrafted feature extraction and optimization overhead; performance may depend on feature quality.	99.47%
[54]	Copy-Move	Extraction of 9D facial biometric anomaly features followed by XGBoost classification to distinguish authentic and deepfake videos	DeepSpeak v1.0, Celeb-DF-v2	Warrants: High robustness to compression and resolution changes; strong generalization to unseen deepfake generators; interpretable feature importance. Drawback: Performance drops when evaluated across different datasets (cross-dataset generalization).	Up to 99.1% (Celeb-DF-v2), 94.9% (combined data)
[55]	GAN-Based Synthetic	GAN used for synthetic data generation; SVM classifier trained on original MNIST and used to classify GAN-generated samples	MNIST	Warrants: GAN effectively augments data without explicit probability modeling; SVM provides good performance with low computational cost. Drawbacks: Performance is lower than deep learning models and limited to simple datasets such as handwritten digits.	96.67%

4. Image Forgery Detection Performance Metrics

There are a number of metrics that have been used by researchers to assess how well machine learning-based models detect image forgeries. The metrics can be used to evaluate the efficiency of the model in detecting original and forged images. Below, an explanation of the main used metrics will be outlined.

- **Accuracy**

One of the most widely used evaluation criteria for picture forgery detection performance is accuracy, which is an essential indicator of the model's capability. to distinguish between authentic and fake images. It is calculated by dividing the total number of predictions generated by the model by the number of accurate predictions for both faked and original photos. It can be stated numerically as:

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \quad (1)$$

However, accuracy is not always a good measure, especially in imbalanced datasets where the number of real images is much higher than the fraudulent ones. In such cases, the model can be accurate by predicting most images as real, which does not measure the actual detection performance of the model [12]. Accuracy alone is still insufficient for trustworthy performance evaluation, especially in imbalanced datasets, even though certain deep learning approaches trained on large-scale forgeries datasets have shown accuracy surpassing 95%, outperforming handcrafted feature-based methods [13]. In general, traditional machine learning techniques like Support Vector Machines and k-Nearest Neighbors produce lower accuracy than deep learning models like CNNs and Vision Transformers. Nevertheless, to obtain a comprehensive evaluation of model performance, accuracy should be considered alongside additional measures, such as F1-score, precision, and recall, particularly for unbalanced datasets [14].

- **Precision and Recall**

Since the real and fake images are typically very different, the accuracy of identifying image forgeries is actually insufficient. The precision and recall value can provide a clue about the ability of the model in the detection of actual forgeries while reducing the false positives to a great extent [10]. This can be mathematically represented as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall that this is also referred to as sensitivity or true positive rate (TPR), can be represented mathematically as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

In many instances, there may be an inverse relationship between the two measures. Recall may decline in order to achieve more precision, and vice versa. In order to strike a compromise between minimizing false alarms and ensuring that no counterfeit photos are overlooked, this is a crucial trade-off in the forensic field [47]. Depending on the objective, the majority of forgery detection systems would like to maximize either recall or precision: In Copy-Move Forgery Detection for example, to identify every copied area, the majority of models created for this purpose strive for high recall values [11]. Additional post-processing procedures, including context-based filtering, might be necessary as a result of the potential rise in false alarms. On the other side, for Deep-fake and Splicing Detection: most GAN-based deep-fake detection algorithms strive to maximize precision values to reduce the number of false alarms because it can be illegal to mistakenly classify real photos as forgeries [13].

5. Challenges and Limitations

Despite the effectiveness of machine learning-based image forgery detection techniques, they

are also suffering from some challenges and limitations. Handcrafted feature extraction is highly dependent on the dataset and the forgery type, resulting in poor generalization capabilities for different forgery methods. Moreover, feature design and selection are often manually tuned and computationally expensive. Moreover, traditional feature-based machine learning techniques are not very effective in identifying highly realistic synthetic images created using advanced models such as StyleGAN, StyleGAN2, and diffusion-based models, as these forgeries contain fewer artifacts. The absence of comprehensive and updated benchmarks, especially for diffusion-based forgeries, makes it difficult to effectively evaluate the existing forgery detection techniques.

6. Future Research Directions

The rising level of realism in the synthetic face images produced by models such as StyleGAN, StyleGAN2, and diffusion models makes it a challenge for the current image forgery detection techniques. The current benchmarks and methods are primarily focused on GAN-based forgeries and do not represent diffusion-based synthetic faces well. Future research should aim at developing comprehensive datasets that include StyleGAN-based and diffusion-based synthetic faces, and also develop feature-based machine learning methods that can utilize the discriminative features of the images effectively to identify forgeries of synthetic faces.

7. Conclusion

This paper has provided a thorough analysis of methods for detecting image forgeries with a specific focus on machine learning-based approaches. Various types of image forgeries, starting from traditional manipulation techniques to more contemporary AI-generated synthetic images such as GAN-based and diffusion-based deep-fakes, have been highlighted to stress the ever-evolving nature of digital image forgery. The machine learning approaches reviewed in this paper employ handcrafted feature extraction and traditional classifiers to distinguish between authentic and forged images, which are interpretable and computationally efficient. However, it has been noted that the existing approaches for detection via machine learning are currently faced with some challenges, particularly in the generalization and robustness aspects when it comes to highly realistic images generated by contemporary generative models. The lack of appropriate benchmarks for diffusion-based forgeries further limits the scope of the review. To sum up, this review article concludes the need for improving feature representation and enhancing the generalizability of machine learning models for efficient image forgery detection.

References

[1] A. F. H. Sewan and M. S. M. Altaei, "Copy Move Forgery Detection Using Forensic Image," *Iraqi Journal of Science*, vol. 62, no. 9, pp. 3167-3181, 2021.

[2] L. D. Griffin, M. Caldwell, J. T. A. Andrews, and H. Bohler, "Unexpected item in the bagging area": Anomaly detection in X-ray security images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1539-1553, June 2019, doi: 10.1109/TIFS.2018.2881700.

[3] M. Vafadar and H. Ghassemian, "Hyperspectral anomaly detection using combined similarity criteria," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4076-4085, 2018.

[4] M. Ning, P. Yu, W. Shaojun and G. Wei, "A weight SAE based hyperspectral image anomaly targets detection," in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, 2017, pp. 511-515.

[5] M. G. Alex, C. Rajalakshmi, and R. Balasubramanian, "Study of image tampering and review of tampering detection techniques," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, pp. 963-967, 2017.

[6] H. Farid, "Image forgery detection," *IEEE Signal processing magazine*, vol. 26, no. 2, pp. 16-25, 2009.

[7] P. Korus, "Digital image integrity–a survey of protection and verification techniques,"

Digital Signal Processing, vol. 71, no. 2017, pp. 1-26, 2017.

[8] S. Ozturk and E. Gul, "A novel hash function based fragile watermarking method for image integrity," *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 17701-17718, 2019.

[9] Z. Zhang, Y. Ren, X. Ping, Z. He and S. Zhang, "A survey on passive-blind image forgery by doctor method detection," *in 2008 international conference on machine learning and cybernetics*, 2008, pp. 3463-3467.

[10] Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2020). Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5074–5083).

[11] Bai, R. (2025). Weakly-Supervised Cross-Contrastive Learning Network for Image Manipulation Detection and Localization. Available at: ScienceDirect.

[12] Chen, W., Zhang, C., & Xu, B. (2023). Hybrid CNN-ViT Model for Image Forgery Detection. Neural Information Processing Systems (NeurIPS).

[13] Chaudhuri, A., Bhunia, A. K. (2024). Self-Supervised Learning for Robust Forgery Detection Models. IEEE International Conference on Computer Vision (ICCV).

[14] Dixit, A., Sharma, S. K., & Dhaka, M. (2024). Image and Video Retrieval and Authentication using AI-Driven Techniques for Secure Media Management. ICTACT Journal on Image and Video Processing.

[15] Duan, H., Jiang, Q., Xu, X. (2025). Adversarial Samples Generated by Self-Forgery for Face Forgery Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

[16] Gaikwad, S., & Mizwan, Z. (2025). Detection of Image Forgery using Deep Learning. Available at: AIP.

[17] Gonzalez, S., & Tapia, J. E. (2025). Forged Presentation Attack Detection for ID Cards on Remote Verification Systems. Available at: ScienceDirect.

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Courville, A. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS).

[19] Han, R., Wang, X., Bai, N. (2024). HDF-Net: Capturing Homogeny Difference Features to Localize the Tampered Image. IEEE Transactions on Machine Intelligence.

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[21] Zahra, Hamid.A.Jalab & Rafidah Md. Noor (2019). Image splicing forgery detection based on low-dimensional singular value decomposition of discrete cosine transform coefficients. Neural Computing and application 31(1)

[22] Jin, X., Wozniak, M., Duan, H., Jiang, Q. (2025). Mf-net: Multi-Feature Fusion Network Based on Two-Stream Extraction and Multi-Scale Enhancement for Face Forgery Detection. Springer Complex & Intelligent Systems.

[23] Khalaf, L. I., Jumaili, M. L. F., & Mahmood, M. T. (2025). Image Forgery Detection using Convolutional Neural Networks and Blockchain Technology. Available at: ACM.

[24] Kim, H., & Park, J. (2024). Improving Generalization in Image Forgery Detection with Adversarial Training. IEEE Transactions on Artificial Intelligence (TAI).

[25] Liu, R., Zhang, S., Xu, Y. (2025). High-Resolution Network-Based MultiFeature Fusion for Generalized Forgery Detection. Multimedia Systems, Springer. Link

[26] Liang, H., Leng, Y., Luo, J. (2024). A Face Forgery Video Detection Model Based on Knowledge Distillation. IEEE Transactions on Artificial Intelligence (TAI). Link

[27] Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.

[28] Wang, X.; Guo, H.; Hu, S.; Chang, M.C.; Lyu, S. Gan-generated faces detection: A survey and new perspectives. *arXiv* **2022**, arXiv:2202.07145.

[29] Natsume, R.; Yatagawa, T.; Morishima, S. RSGAN: Face swapping and editing using face and hair representation in latent spaces. *arXiv* 2018, arXiv:1804.03447.

[30] Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.

[31] Pumarola, A.; Agudo, A.; Martinez, A.M.; Sanfeliu, A.; Moreno-Noguer, F. GANimation: One-shot anatomically consistent facial animation. *Int. J. Comput. Vis.* **2020**, *128*, 698–713.

[32] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In ICLR

2018, Vancouver, BC, Canada, 30 April–3 May 2018.

[33] Donahue, J.; Krähenbühl, P.; Darrell, T. Adversarial feature learning. *arXiv* 2016, arXiv:1605.09782.

[34] Carrara, F.; Amato, G.; Brombin, L.; Falchi, F.; Gennaro, C. Combining GANs and autoencoders for efficient anomaly detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3939–3946.

[35] Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 214–223.

[36] Chaudhari, P.; Agrawal, H.; Kotecha, K. Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Comput.* **2020**, *24*, 11381–11391.

[37] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. vol. 35, pp. 25278–25294. Curran Associates, Inc. (2022).

[38] Stability ai. https://stability.ai/

[39] Midjourney discord server. https://discord.com/invite/midjourney

[40] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11 (2019) Supplementary Material.

[41] Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging datasetfor deepfake forensics. In: CVPR (2020)

[42] Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B.: Deepfakebench: A comprehensivebenchmark of deepfake detection. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)

[43] Cozzolino, D., Thies, J., Ro¨ssler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018).

[44] Aneja, S., Nießner, M.: Generalized zero and few-shot transfer for facial forgerydetection. arXiv preprint arXiv:2006.11863 (2020).

[45] C. Bhattacharyya, H. Wang, F. Zhang, S. Kim, and X. Zhu, "Diffusion Deepfake," arXiv:2404.01579v1 [cs.CV], Apr. 2024. doi: 10.48550/arXiv.2404.01579.

[46] Moghaddasi, Z., Jalab, H. A., & Noor, R. M. (2018). Image splicing forgery detection based on low-dimensional singular value decomposition of discrete cosine transform coefficients. Neural Comput. Appl. 31(11), 7867–7877.

[47] Zhu, Y., Li, Q., Wang, J., Xu, C. Z., & Sun, Z. (2021). One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4834–4844).

[48] Zhang, Q. et al. "Digital image splicing detection based on Markov features in block DWT domain", *Multimedia Tools and Applications*, vol. 77, no 23, DODOI10.1007/s11042-018, 2018.

[49] A. K. Jaiswal, and R. Srivastava, "A technique for image splicing detection using hybrid feature set," *Multimedia Tools and Applications*, vol. 79, no. 17, pp. 11837-11860, 2020.

[50] M. Habibi and H. Hassanpour, "Splicing image forgery detection and localization based on color," *International Journal of Engineering*, vol. 34, no. 2, pp. 443-451, 2021.

[51] A. Parnak, Y. D. Baleghi and . S. Kazemitabar, "A Novel Image Splicing Detection Algorithm Based on Generalized and Traditional Benford's Law," *International Journal of Engineering*, vol. 35, no. 4, pp. 626-634, 2022.

[52] S. Solaiyappan and . Y. Wen, "Machine learning based medical image deep fake detection: A comparative study, *Machine Learning with Applications*, vol. 8, no. 2, pp. 2666-8270, 2022.

[53] M. R. Archana, D. Jamkhandikar, and D. N. Biradar, "Image copy-move forgery detection and classification using golden jackal optimization based multi-support vector machine," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 6, pp. 123–133, 2023, doi: 10.22266/ijies2023.1231.11.

[54] J. D. Norman and H. Farid, *"Detecting Deepfake Talking Heads from Facial Biometric Anomalies,"* arXiv preprint arXiv:2507.08917, Jul. 2025, doi: 10.48550/arXiv.2507.08917.

[55] A. Singh, A. Bansal, N. Chauhan, and S. P. Sahu, "Image generation using GAN and its classification using SVM and CNN," in *Proceedings of Emerging Trends and Technologies on Intelligent Systems*, Advances in Intelligent Systems and Computing, vol. 1365, Singapore: Springer, 2022, pp. 89–100, doi:10.1007/978-981-16-3097-2_8.