Journal of Computer Virology and Hacking Techniques (2026) 22:2
https://doi.org/10.1007/s11416-025-00584-y

RESEARCH

A comprehensive system for detecting and verifying counterfeit images using deep neural networks

Praharsh Ajit PaiaAjit Paia[1] · Rizwan Ur Rahman[1] · Deepak Singh Tomar[2]

Received: 11 March 2025 / Accepted: 21 October 2025

Abstract

The widespread adoption of advanced image morphing techniques, including deepfakes, has posed significant challenges to cybersecurity, digital forensics, and biometric authentication. Morphed images are increasingly exploited for identity theft, misinformation campaigns, cyber extortion, and other fraudulent activities. The rapid advancement of generative AI technologies, such as large language models (LLMs) like ChatGPT and image generation platforms like MidJourney, has further exacerbated these threats. While these technologies have transformed creative and professional domains, they are also misused to produce highly convincing fake content, including deepfakes and morphed images.

In response to these emerging challenges, this paper proposes a Comprehensive System for Detecting and Verifying Counterfeit Images using Deep Neural Networks. The system integrates metadata analysis, AI-driven anomaly detection, and advanced deep learning models such as EfficientNet, DenseNet, ResNet, and VGG to improve detection accuracy. Extensive testing on diverse real-world datasets containing both authentic and morphed images demonstrates the system's ability to effectively identify morphing-related inconsistencies. By leveraging cutting-edge neural network architectures, this solution aims to strengthen digital trust and play a pivotal role in combating cybercrimes involving counterfeit images.

Keywords: Image Morphing · Image Crime · LLM · Face Morphing · Deepfake Generation · EfficientNet · DenseNet · ResNet

1 Introduction

In today's digital landscape, advancements in image editing tools have made manipulating visual content easier than ever, raising significant concerns about privacy, security, and ethics. One of the most alarming forms of manipulation is facial image morphing—a technique that blends features from two or more faces to create a convincing composite. While this technology has legitimate applications in art and media, it is increasingly being exploited for malicious purposes, including cybercrimes and security breaches. Facial morphing poses a severe threat to biometric systems, which rely heavily on facial recognition for identity verification in critical areas such as passport issuance, driver's licenses, and secure access control. By combining distinguishable features from multiple individuals, these morphed images can bypass authentication processes, facilitating unauthorized access, identity theft, and fraud.

Image morphing's threat has been further heightened by the fast expansion of generative artificial intelligence technologies including image generating tools like MidJourney and large language models (LLMs) like ChatGPT. Although they are transforming the creative and business sectors, these tools have also been used to produce quite convincing false content including synthetic identities, deepfakes, and manipulated photos. Recent instances of abuse have underlined its capacity for evil intent including biometric fraud, cyber extortion, identity theft, and disinformation operations. The dual-edged character of generative artificial intelligence emphasizes how urgently strong forensic

✉ Rizwan Ur Rahman
rizwan.rahman12@gmail.com

Praharsh Ajit PaiaAjit Paia
praharsh.24mcf10008@vitbhopal.ac.in

Deepak Singh Tomar
deepaktomar@manit.ac.in

[1] School of Computing Science Engineering and Artificial Intelligence (SCAI) , VIT Bhopal, University, Bhopal, India

[2] Department of Computer Science, Maulana Azad National Institute of Technology, Bhopal, India

Published online: 13 November 2025

systems are needed to properly identify and counteract such risks. Recent data clearly show the fast acceptance of these artificial intelligence tools. With around 100 million users by January 2023 [1], ChatGPT, for example, is the fastest-growing consumer app in history and has seen hitherto unheard-of expansion. ChatGPT exceeded 180 million users by August 2023 [2], hence this figure kept increasing. With approximately 300 million weekly active users by December 2024 [3], the platform indicated consistent and quick user involvement. MidJourney has similarly seen notable user base increase. With daily active users ranging between 1.2 million and 2.5 million [4], MidJourney had amassed over 16 million overall as of November 2023. Reflecting a development of around 2.86 million users within just four months [5], the number of registered users rose to 19.26 million by March 2024. The fast spread of AI-driven content production platforms has reduced the obstacle for creating complex synthetic media, thereby raising the possibility of their application in different hostile operations. The broad availability and easy-to-use interfaces of these tools help people with low technical knowledge create plausible fake information, therefore aggravating the difficulties experienced by experts in digital forensics and cybersecurity.

The increasing number of cybercrimes involving morphed images in India serves as a stark warning about the risks of this technology. In 2023, Odisha reported 273 cases [6] where women's photographs were altered and distributed online, predominantly to humiliate and harass them. In Mumbai, a sextortion case [7] connected to a fraudulent loan app tragically ended in a victim's suicide after their doctored image was shared among their contacts. Other cities have also witnessed alarming incidents of this nature: extortion schemes in Delhi [8], blackmail through morphed images in Bengaluru [9], matrimonial scams in Ahmedabad [10], revenge porn cases in Kolkata [11], and cyberstalking in Jaipur [12]. Similarly, Hyderabad and Chennai [13] reported multiple instances where women's photos were manipulated and misused on social media for harassment. These cases highlight the profound emotional, social, and reputational harm caused by morphing-related offenses, further eroding trust in digital platforms.

The detection of manipulated images is becoming increasingly difficult due to the advanced capabilities of modern editing tools, which can create highly realistic alterations. Traditional forensic methods are often unable to detect these subtle manipulations, necessitating the adoption of advanced artificial intelligence solutions. However, developing effective AI-based frameworks presents its own set of challenges, including the lack of standardized datasets and evaluation criteria for training and testing detection models. Compounding this issue is the absence of clear legal provisions in India to address crimes involving morphing. Although the Information Technology Act, 2000, addresses general cyber offenses, it does not explicitly cover cases of morphing, leaving victims unprotected and complicating legal proceedings against offenders.

Addressing the misuse of image morphing demands a holistic approach that combines technological innovation, legal measures, and ethical considerations. Developing advanced AI-driven detection tools capable of accurately identifying morphed images is crucial. Equally vital is the establishment of legal policies to penalize offenders and protect victims from image-based exploitation. Public awareness initiatives and digital literacy campaigns can further mitigate the misuse of image manipulation technologies while promoting responsible online behavior. By integrating these efforts, it becomes possible to enhance biometric security, protect individual privacy, and build trust in digital environments, fostering a safer and more ethical cyberspace.

To the best of our knowledge, there is currently no fully developed system capable of effectively detecting morphed images, especially those involving human faces. This gap poses a significant challenge for cybersecurity professionals and law enforcement agencies tasked with combating crimes involving image manipulation. Without such a system, the timely identification and prevention of offenses that exploit advanced image editing tools for malicious purposes remains hindered. To address this issue, we propose a six-phase approach leveraging modern artificial intelligence techniques to ensure accurate detection of altered images. These phases include: Identification (tracing the image's origin and analyzing its context), Preservation (maintaining image integrity and documentation), Collection (gathering metadata and relevant environmental details), Analysis (applying AI/ML algorithms to detect alterations), Documentation (systematically recording findings and methodologies), and Presentation (clearly presenting results with supporting evidence and actionable recommendations). This proposed system not only focuses on identifying manipulations with high precision but also considers the broader technical, legal, and ethical aspects of image morphing. By addressing these dimensions, the approach aims to enhance digital security, protect privacy, and establish stronger ethical standards in a world increasingly influenced by sophisticated image editing technologies.

As part of this initiative, our proposed system has been rigorously tested against real-world scenarios to validate its effectiveness. To demonstrate its capabilities, the proposed system is applied to a well-documented case involving a widely circulated morphed image used in cybercrime. By analysing this image, the proposed system successfully detected key markers of manipulation, including inconsistencies in pixel patterns, lighting discrepancies, and

anomalies in facial features that were seamlessly blended. This validation process not only highlighted the proposed system's accuracy but also underscored its potential for practical application in addressing real-life crimes involving morphing.

The results of this test case reinforce the importance of integrating AI-driven solutions with cyber security expertise to develop a robust tool capable of mitigating the risks posed by morphing-related offenses. Through such advancements, this system represents a pivotal step toward securing digital platforms, protecting individual privacy, and fostering public trust in the authenticity of visual content.

2 Related work

The rapid advancement of digital technologies has led to an unprecedented rise in the sophistication of manipulated images and deepfakes, raising significant concerns about the trustworthiness of visual content. From subtly altered identification photos to highly convincing synthetic faces, these manipulations leverage advanced algorithms to deceive viewers. In response to this growing challenge, researchers have developed various methods to distinguish between authentic and tampered images. This section delves into the key advancements, methodologies, and ongoing challenges in this critical area of study, shedding light on the continuous efforts to combat the rise of digital deception.

The ongoing research on detecting tampered or manipulated images, particularly in the context of deepfake detection, has led to significant advancements in various techniques and methods. However, challenges persist, especially with new forms of image manipulation. The work by Miki Tanaka, Sayaka Shiota and Hitoshi Kiya which proposed a method using robust hashing to detect tampered images [14], exemplifies the traditional approach of using hash-based comparisons for manipulation detection. While their method can identify tampering even under compression and resizing, it struggles to keep up with emerging and more sophisticated manipulation techniques, such as those generated by Generative Adversarial Networks (GANs), which are designed to avoid traditional detection methods.

Dilip Kumar Sharma and Bhuvanesh Singh's survey on forensic, machine learning, and deep learning techniques for fake image detection on social media [15] provides a comprehensive overview of the state of the art, focusing on the growing importance of multimodal systems. Such systems combine different types of data, such as visual, auditory, and contextual information, to improve the accuracy of fake detection. Despite this, their research lacks practical implementation strategies that can be applied in real-world social media scenarios, where challenges such as the speed of detection and the adaptability of the models are critical factors for success.

In the realm of deep learning, Chandra Bhushana Rao Killi, Narayanan Balakrishnan and Chinta Someswara Rao explored the use of VGG-19 for classifying deepfake images [16], achieving 96% accuracy through the application of regularization techniques to prevent overfitting. However, the lack of clarity regarding dataset diversity raises concerns about how well this model can generalize to unseen data. Chih-Chung Hsu, Yi-Xiu Zhuang and Chia-Yen Lee's work on a pairwise learning approach with a Channel Feature Fusion Network (CFFN) [17] demonstrated superior precision and recall in detecting GAN-generated images. Nonetheless, their evaluation is limited to a smaller set of GANs, which makes their method less robust to newer generations of GANs.

Human visual cognition has also played a role in the detection of image manipulations. The work of Giuseppe Cartella, Vittorio Cuculo, Marcella Cornia and Rita Cucchiara focused on gaze patterns [18] to differentiate real from fake images, introducing an innovative approach based on human behavior. While this method has shown potential, the reliance on manual gaze data collection is a major limitation, as it severely restricts scalability, making it difficult to implement in large-scale or automated systems. In a similar vein, Huang and Juefei-Xu developed FakeLocator [19], which uses grayscale fakeness maps for localizing manipulations. Though this method shows high accuracy in identifying tampered regions, it becomes less effective when images are subjected to heavy compression, which is often the case for social media images that are compressed before being uploaded.

Transformers, which have gained prominence in recent years for various image-processing tasks, have also been adapted for synthetic image detection. Huan Liu1, Zichang Tan and Chuangchuang Tan's FatFormer, a forgery-aware adaptive transformer [20], shows excellent performance in detecting both GAN and diffusion model-generated images. The model's accuracy, however, is tempered by its high computational cost, which can make real-time detection in practical applications, such as video calls or live streaming, infeasible. Similarly, Tessa R. Flack and Kay L. Ritchie's research into face morph detection using collaborative training [21] leverages shared features between paired images, improving the model's ability to detect face morphing manipulations. However, while the model performs well in controlled environments, its scalability in real-world applications, such as border control systems, remains unverified.

Compact neural networks, such as MesoNet developed by Darius Afchar, Vincent Nozick, Junichi Yamagishi and Isao Echizen [22], have also been explored for video-based manipulations. MesoNet, a lightweight architecture,

achieves over 98% detection rates for Deepfake videos, offering an efficient solution for video-based fake detection. Despite its success, the model faces significant limitations when it comes to highly compressed video content, which is common in video-sharing platforms where compression algorithms reduce file sizes. In a similar context, Luca Bondi1 and Silvia Lameri proposed a tampering detection method that clusters CNN-based camera features [23] to detect manipulated regions, providing a robust solution for detecting manipulations in videos. However, the method struggles to generalize to videos captured from unfamiliar camera models, highlighting a limitation of CNN-based methods when working across diverse camera types.

Frequency domain analysis has been another avenue of research. Muhammad S. Mandisha and Mohamed A. Hussien's work, which combined wavelet transforms with gradient boosting for tampered image detection [24], outperforms traditional Fourier-based methods, achieving a high accuracy of 95%. However, it faces challenges when applied to adversarial manipulated images, as such images may intentionally be altered to avoid detection by frequency-based methods. Lucy Chai, David Bau, Ser-Nam Lim and Phillip Isola also contributed to this area with patch-based classifiers [25] that demonstrate superior generalization across different datasets. However, similar to Muhammad S. Mandisha and Mohamed A. Hussien's method, these classifiers struggle to detect images manipulated by fine-tuned adversarial generators.

A systematic literature review by El-Sayed Atlam and Malik Almaliki explored the research gaps in the deepfake dynamics [26], advocating for multidisciplinary collaboration in addressing the issue. While their review is comprehensive, it lacks empirical evaluations of the proposed solutions, making it difficult to gauge the practical impact of their recommendations. Ramesh Gorle and Anitha Guttavelli combined Error Level Analysis (ELA) with Convolutional Neural Networks (CNNs), achieving 96.21% accuracy in detecting tampered images. Despite the high accuracy, their method is constrained by the limited size of evaluated datasets, raising concerns about the model's generalizability [27].

Face morphing detection has also attracted significant attention. Chalini G R and K. V. Kanimozhi proposed Differential Morphing Attack Detection (D-MAD) [28], which is particularly useful for detecting morphing attacks in border control scenarios. However, its effectiveness across diverse datasets remains uncertain, limiting its real-world applicability. A. Ramesh, Bheema Sri Lakshmi, Dasari Narendar, Moinuddin Mohammed Najeeb and Vudaru Sai employed pre-trained deep learning models like VGG19 for morphing attack detection [29], but the lack of standardized benchmarks in their study makes it challenging to compare their approach with others. Padmaja Kadiri and Palagati Anusha's exploration of machine learning algorithms like Support Vector Machines (SVM) and Photo Response Non-Uniformity (PRNU) for detecting morphed images [30] emphasizes the need for diverse datasets, as their approach struggles with real-world variability in image manipulations [31].

The comparison of the literature papers discussed above is provided in the table below for better clarity and understanding (Table 1).

Table.1. Comparison table of all related/background work.

Author and Year	Methodology	Focus	Datasets	Key Findings
Luca Bondi, Silvia Lameri, David Guera, Paolo Bestagini, Edward J. Delp, Stefano Tubaro (2017)	CNN feature extraction and clustering	Tampering detection based on camera model features	Dresden Image Database	High accuracy in detecting tampering, even with unknown camera models; effective localization of forged regions.
Darius Afchar, Vincent Nozick, Junichi Yamagishi, Isao Echizen (2018)	Deep learning networks with low layer count	Detection of face tampering in videos	Custom dataset (175 forged videos), FaceForensics dataset	Achieved over 98% detection rate for Deepfake and 95% for Face2Face; challenges noted with video compression.
Chih-Chung Hsu, Yi-Xiu Zhuang and Chia-Yen Lee (2020)	Pairwise learning with Common Fake Feature Network (CFFN)	Detecting GAN-generated images	CelebA (202,599 aligned face images, plus GAN-generated fake images)	Outperformed existing methods in precision and recall. The approach generalizes well to new GAN-generated images.
Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola (2020)	Patch-based classifier with limited receptive fields	Identifying detectable artifacts in fake images	CelebA-HQ, FFHQ, FaceForensics	Patch-based classifiers outperform full-image classifiers for out-of-domain synthetic images; even adversarially fine-tuned image generators leave detectable artifacts.

Author and Year	Methodology	Focus	Datasets	Key Findings
Miki Tanaka, Sayaka Shiota and Hitoshi Kiya (2021)	Robust hashing for feature extraction	Detecting fake images under compression and resizing	Image Manipulation Dataset, UADFV, CycleGAN, StarGAN	Effective detection of tampered images using hash values from reference and query images; superior to state-of-the-art methods. Applicable for monitoring unauthorized manipulation.
Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, Geguang Pu (2021)	Localization using grayscale fakeness maps	GAN-based face manipulation localization	CelebA, FFHQ	High localization accuracy; robust to degradations like JPEG compression. Integrated attention mechanism improves universality.
Dilip Kumar Sharma, Bhuvanesh Singh, Saurabh Agarwal, Lalit Garg, Cheonshik Kim and Ki-Hyun Jung (2023)	Review of forensic, ML, and deep learning methods	Social media misinformation and fake image detection	PGGAN, FNC, Medi-aeval, CASIA, CelebA	Highlights deep learning's effectiveness in fake image detection and emphasizes the need for multimodal approaches and interdisciplinary research.
Chandra Bhushana Rao Killi, Narayanan Balakrishnan, Chinta Someswara Rao (2023)	Deep learning (VGG-19)	Classification of deepfake images	Large dataset of natural images (specific datasets not detailed)	Achieved 96% accuracy; highlights the effectiveness of VGG-19 with regularization techniques like Dropout and Batch Normalization.
Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei1, Yao Zhao, Jingdong Wang (2023)	FatFormer Transformer with forgery-aware adapter	Detection of GAN and diffusion-based fake images	ProGAN, StyleGAN, BigGAN, CycleGAN, StarGAN, GauGAN, DALL-E, LDM, Glide	Achieved 98.4% accuracy for GANs and 95% for diffusion models; outperformed existing methods in detection generalizability.
Tessa R. Flack, Kay L. Ritchie, Charlotte Cartledge, Elizabeth A. Fuller, Robin S. S. Kramer (2023)	Collaborative training experiments	Face morph detection	Data from experiments with 166 participants	Pairs training effect significantly improves detection accuracy; real-world application potential in enhancing border control.
Giuseppe Cartella, Vittorio Cuculo, Marcella Cornia, Rita Cucchiara (2024)	Eye-tracking experiments	Human visual perception of real vs. fake images	COCO, ADE20K, LHQ	Identified distinct gaze patterns for real vs. fake images; potential to enhance fake detection through semantic knowledge and gaze behavior.
Muhammad S. Mandisha, Mohamed A. Hussien, Amr K. Shalaby, Omar M. Fahmy (2024)	Wavelet transform with gradient boosting classifier	Detecting manipulated images	FaceForensics	Achieved 95% accuracy, outperforming Fourier transform-based methods; computationally efficient.
Ramesh Gorle, Anitha Guttavelli (2024)	ELA combined with CNN	Image tampering detection	CASIA v2.0	Achieved 96.21% detection accuracy, surpassing established models like VGG16, VGG19, and ResNet101; highlights the effectiveness of integrating ELA with CNN.
Chalini G R, K. V. Kanimozhi, (2024)	Differential Image-Based Morphing Attack Detection (D-MAD)	Detection of face morphing attacks	Custom dataset with 143 data subjects, public facial datasets	D-MAD demonstrates strong performance in border control scenarios, surpassing traditional methods; effective across multiple datasets.

Author and Year	Methodology	Focus	Datasets	Key Findings
Mr. A. Ramesh, Bheema Sri Lakshmi, Dasari Narendar, Moinuddin Mohammed Najeeb, Vudaru Sai (2024)	Deep convolutional neural networks	Face morph detection	Datasets not specified	Pretrained networks like VGG19 outperform those trained from scratch; emphasizes the need for standardized benchmarks and robust detection algorithms.
Padmaja Kadiri, Palagati Anusha, Madhav Prabhu, Rolito Asuncion, Voonna Sainath Pavan, Jami Venkata Suman (2024)	Machine learning (SVM, PRNU analysis, deep learning)	Detection of morphed images	Custom dataset for morphed images	Highlights vulnerabilities in face recognition systems to morphing attacks; emphasizes the importance of diverse datasets for improving detection accuracy.
El-Sayed Atlam, Malik Almaliki, Ghada Elmarhomy, Abdulqader M. Almars, Awatif M.A. Elsiddieg, Rasha ElAgamy (2025)	Systematic literature review	Research focus on detection, dynamics, and prevention of deepfakes	No datasets specified	Identifies gaps in research on deepfake dynamics and prevention; advocates for multidisciplinary collaboration and digital intervention strategies to combat misinformation.

Despite significant progress in detecting altered images, several critical challenges remain. Many existing algorithms struggle to identify subtle changes, particularly in complex scenarios involving multiple adjustments or real-world distortions such as noise, compression artifacts, or low resolution. These limitations highlight the inability of current methods to detect fine-grained or minor alterations, which are often intentionally designed to evade detection. While deep learning-based techniques have proven effective in certain cases, they require large, labeled datasets for training. This dependency restricts their adaptability to new manipulation techniques that may not be well-represented in the training data.

Localizing tampered regions, particularly for minor or intricate modifications, continues to be a significant challenge. While most existing models can detect image manipulation, they often fail to pinpoint the exact locations of tampering, limiting their effectiveness in forensic investigations. Multi-modal techniques that integrate visual and contextual information (such as metadata or textual cues) hold promise for enhancing detection, but their scalability and applicability in diverse real-world situations have yet to be thoroughly explored. Additionally, the interpretability of many advanced systems remains a critical issue. Most models operate as "black boxes," making them unsuitable for high-stakes applications like legal proceedings or law enforcement, where trust, transparency, and explainability are essential.

Collaboration and human-centered approaches, such as gaze monitoring and cooperative training, present promising opportunities for detection, but their practicality and integration into existing workflows still require further testing. Additionally, cross-platform adaptation and generalization continue to pose challenges. While current methods perform well on specific datasets, they often struggle to maintain accuracy when applied to diverse datasets with different modification strategies. This inconsistency across platforms and conditions limits their broader applicability. These gaps emphasize the need for scalable, interpretable, and adaptable techniques that can effectively detect manipulated images in real-world environments.

3 Proposed system for detecting and verifying counterfeit images

With the advancement of sophisticated image editing tools, identifying manipulated images, particularly morphed and deepfake content, has become a significant challenge in digital forensics. How can experts ensure the authenticity of images when modern editing techniques allow for nearly seamless alterations? Current detection methods often lack the necessary AI-driven solutions and structured approaches to effectively tackle this issue.

To address this challenge, we propose a system for detecting and verifying counterfeit images, specifically focusing on morphed and deepfake content. The system is designed to systematically identify, assess, and authenticate manipulated visual content. It incorporates both traditional techniques and AI/ML-based detection methods to improve accuracy and effectiveness. Key components of this system include secure evidence handling, metadata extraction, anomaly detection, and systematic documentation. The model diagram of the proposed system is shown in Fig. 1.

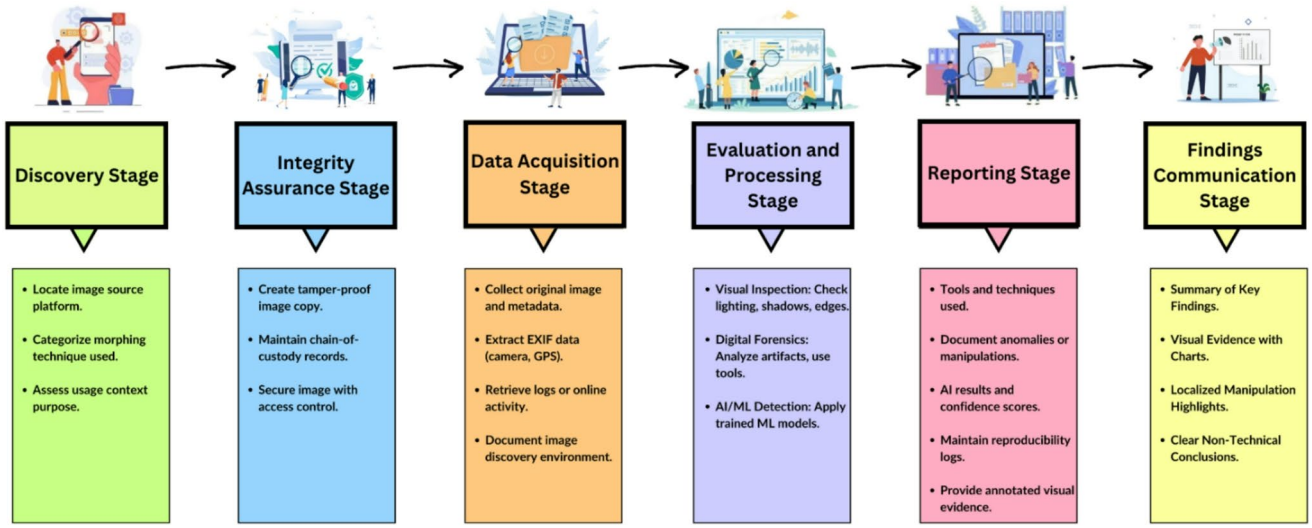

Fig. 1 Comprehensive System for the Detection and Verification of Counterfeit Images.

By addressing gaps in current detection practices, this approach enhances biometric security, reduces cyber threats, and provides a scalable solution for detecting manipulated images, thereby strengthening the overall capacity to combat digital deception.

3.1 Discovery stage

In the Discovery Stage, the primary objective is to recognize digital images or datasets that potentially contain morphed images. This step is crucial as it sets the foundation for subsequent forensic analysis. The process begins by determining the possible sources where morphed images may exist. These sources can include social media platforms, where altered images are frequently shared, cloud storage services that may host manipulated files, digital devices such as smartphones or computers, and archived data repositories. Identifying these sources helps to narrow the scope of the investigation.

Additionally, understanding the different morphing techniques used is critical. Morphing methods may involve face-swapping using artificial intelligence, deepfake generation, or manual image editing through graphic software. Each technique leaves distinct traces or artifacts that can aid in detection. For instance, face-swapping might introduce mismatched facial features, while deepfake images may exhibit inconsistencies in lighting or unnatural textures.

To facilitate this phase, clear criteria must be defined for selecting suspect images. Indicators of morphing include low-resolution images, inconsistent or missing metadata (e.g., EXIF data showing unusual editing history), and visual anomalies such as blurred edges, mismatched lighting, or distorted features. Furthermore, relevant datasets containing a mix of authentic and morphed images should be identified for training and testing purposes in cases where automated detection techniques like machine learning are applied. Datasets with labelled examples of both types of images, such as those sourced from research repositories or generated for forensic purposes, are invaluable for building and validating detection models.

This systematic approach ensures that all potential sources and signs of image morphing are considered, forming a comprehensive basis for effective forensic analysis.

3.2 Integrity assurance stage

Integrity Assurance is a critical stage in ensuring the integrity and authenticity of the identified images before subjecting them to forensic analysis for morph detection. The primary objective is to secure the original images and prevent any alterations or tampering during the investigation. To achieve this, forensic tools are employed to create bitwise copies of the images. These copies replicate the exact binary data of the original files, ensuring that the evidence remains intact while allowing analysts to work on duplicates. This process safeguards the originals as admissible evidence in potential legal proceedings.

To further enhance security, the collected images are stored in encrypted formats or secured repositories, restricting access to authorized personnel only. Additionally, maintaining a chain of custody log is essential. This log meticulously records details such as the source of the image, the date and time of acquisition, and the device or platform from which it was obtained. Such documentation ensures transparency and accountability throughout the forensic process.

To verify the authenticity of the images during and after analysis, hashing techniques like MD5 or SHA-256 are

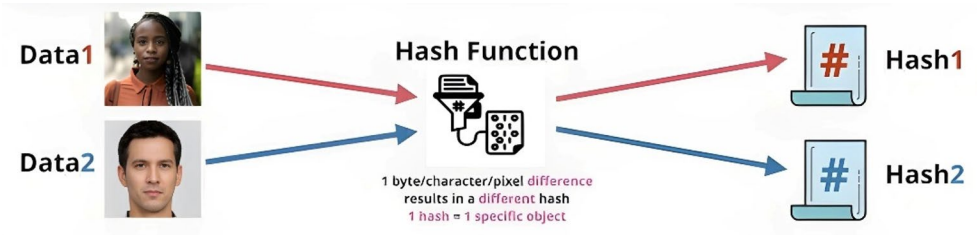

Fig. 2 Hashing method to Preserve and Verify the original state of Image.

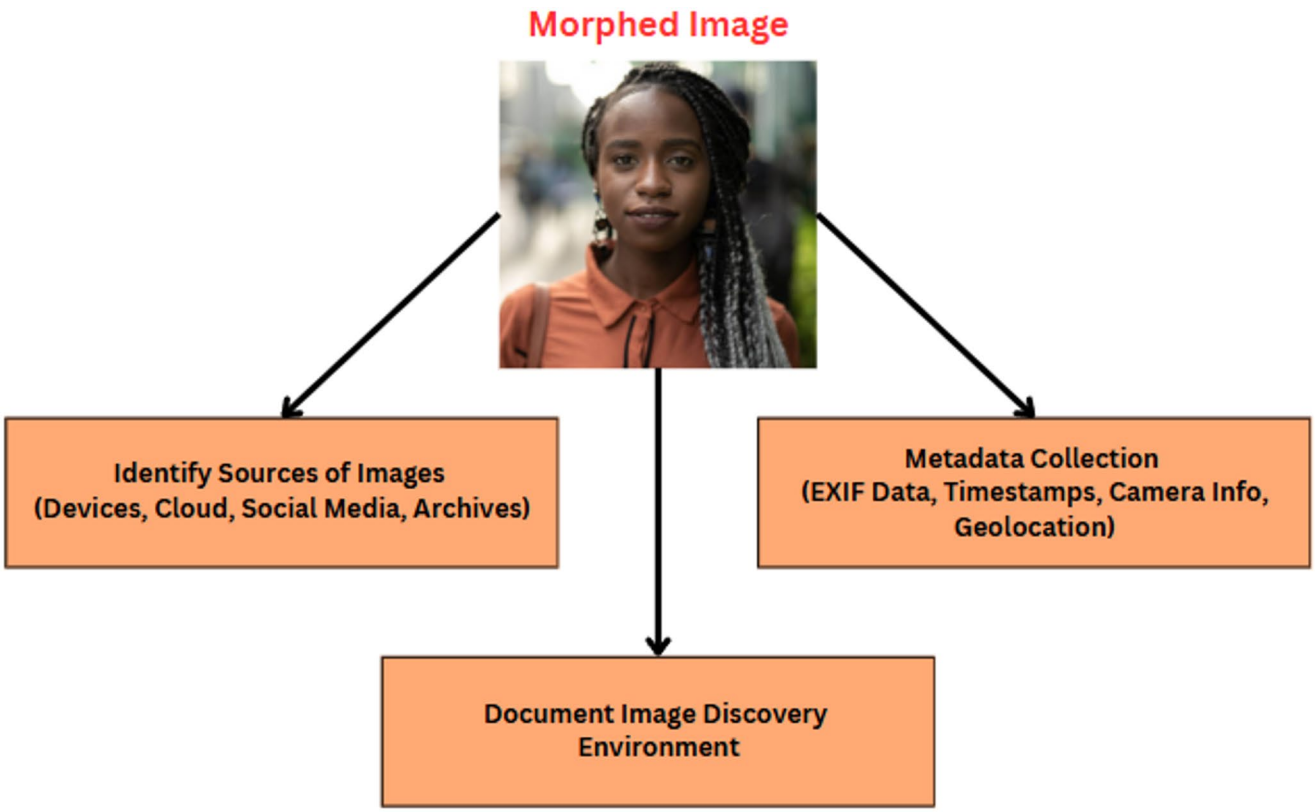

Fig. 3 Collection Phase.

applied. The Hashing method to Preserve and Verify the original state of Image is shown in Fig. 2. These methods generate a unique digital fingerprint for each image. Any subsequent alteration in the image, no matter how minor, will result in a different hash value, immediately alerting investigators to potential tampering. This step is particularly significant in the context of detecting morphed images, as discrepancies in the hash values can indicate unauthorized modifications.

By preserving the original state of the images and maintaining rigorous documentation, forensic analysts can proceed with confidence, ensuring that the evidence is both reliable and legally defensible while focusing on identifying signs of image morphing.

3.3 Data acquisition stage

The Data Acquisition Stage involves systematically gathering images and their associated metadata to ensure comprehensive analysis for detecting morphed images as shown in Fig. 3. This step is critical, as it provides the raw data required for both manual inspection and automated detection techniques. Images can be sourced from various repositories, such as personal devices (hard drives and smartphones), cloud storage platforms, social media, or online archives. Along with the images, it is equally important to collect metadata, such as EXIF (Exchangeable Image File Format) data, timestamps, camera model information, geolocation data, and other attributes embedded within the file. This metadata often holds critical clues to inconsistencies that could indicate image morphing, such as

mismatched timestamps, unusual editing history, or discrepancies in camera make and model.

To extract image files and their metadata effectively, open-source forensic tools like FTK Imager or EnCase can be employed. These tools ensure that data is collected without altering the original files, maintaining the integrity of the evidence. For large-scale analysis, custom scripts can be developed to automate the extraction and organization of data from various sources. Additionally, publicly available datasets containing both authentic and morphed images can be acquired to train and validate machine learning models designed for morph detection. These datasets provide a controlled environment for analyzing features that differentiate morphed images from genuine ones, such as inconsistencies in edges, lighting, and compression artifacts.

By gathering high-quality data and metadata, the collection phase lays the foundation for reliable analysis and detection. Proper documentation of the sources, collection methods, and tools used ensures the reproducibility of the results and maintains the chain of custody, a critical factor in forensic investigations.

3.4 Evaluation and processing stage

The Evaluation and Processing Stage serves as a vital segment of the detection system, encompassing a detailed evaluation of suspected counterfeit images to ascertain their authenticity. This phase combines various methodologies, ranging from conventional visual analysis to advanced digital forensic strategies and state-of-the-art artificial intelligence (AI) and machine learning (ML) techniques, as shown in Fig. 4. By integrating these diverse approaches, this phase ensures a thorough and accurate assessment. Each step in the process contributes uniquely, leading to a holistic understanding of the potential alterations present in the image.

3.4.1 AI/ML-based detection

Artificial intelligence and machine learning have revolutionized forensic image analysis, providing automated, efficient, and scalable solutions for detecting tampered images. Pre-trained models such as convolutional neural networks (CNNs) are utilized to classify images as either authentic or manipulated. These models, trained on diverse and extensive datasets, excel at identifying intricate patterns that signal potential tampering. To enhance detection capabilities, visualization techniques like Grad-CAM and heatmaps are employed to pinpoint specific regions of alteration. These visual tools assist investigators in focusing on areas of concern for further detailed analysis. By extracting subtle, complex features from images and cross-referencing them with reference datasets, AI models provide accurate, reliable results. This AI-driven approach not only improves the speed of analysis but also enhances detection precision, making it indispensable in modern forensic frameworks.

In the analysis of morphed images, EfficientNet-B0, DenseNet50, ResNet50 and VGG-16 were utilized to harness their unique architectures and capabilities.

The EfficientNet model, illustrated in Fig. 5, represents a cutting-edge deep learning architecture specifically engineered to optimize the trade-off between accuracy and computational efficiency, making it particularly well-suited for image classification and other related tasks. The architecture begins with an input layer, where each image is introduced as a multidimensional feature map that accurately reflects the underlying pixel intensity values.

The core computational strength of EfficientNet arises from its sophisticated hidden layers, which are composed of two advanced building blocks: Squeeze-and-Excitation (SE) blocks and Mobile Inverted Bottleneck (MBConv) blocks. The SE block enhances feature discrimination by performing global average pooling, followed by a set of fully connected layers and a sigmoid activation function; this mechanism allows the model to selectively emphasize the most relevant feature channels while suppressing those that are less important. In parallel, the MBConv

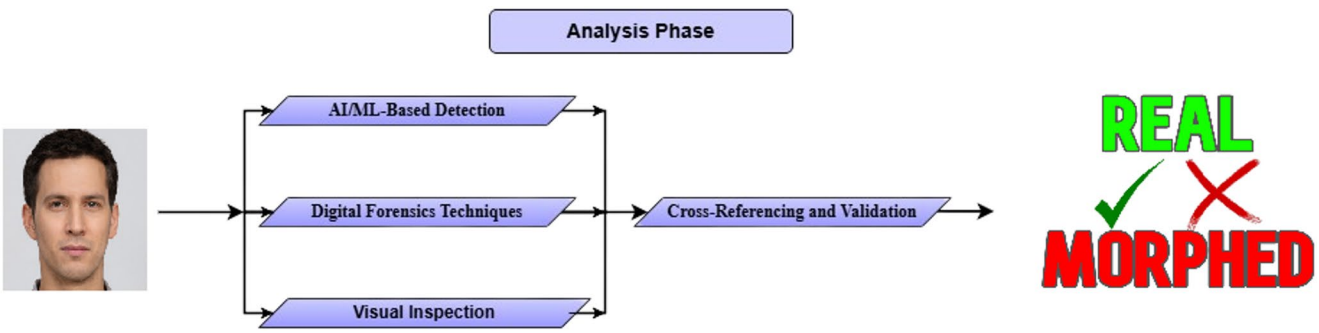

Fig. 4 Analysis and Examination Phase.

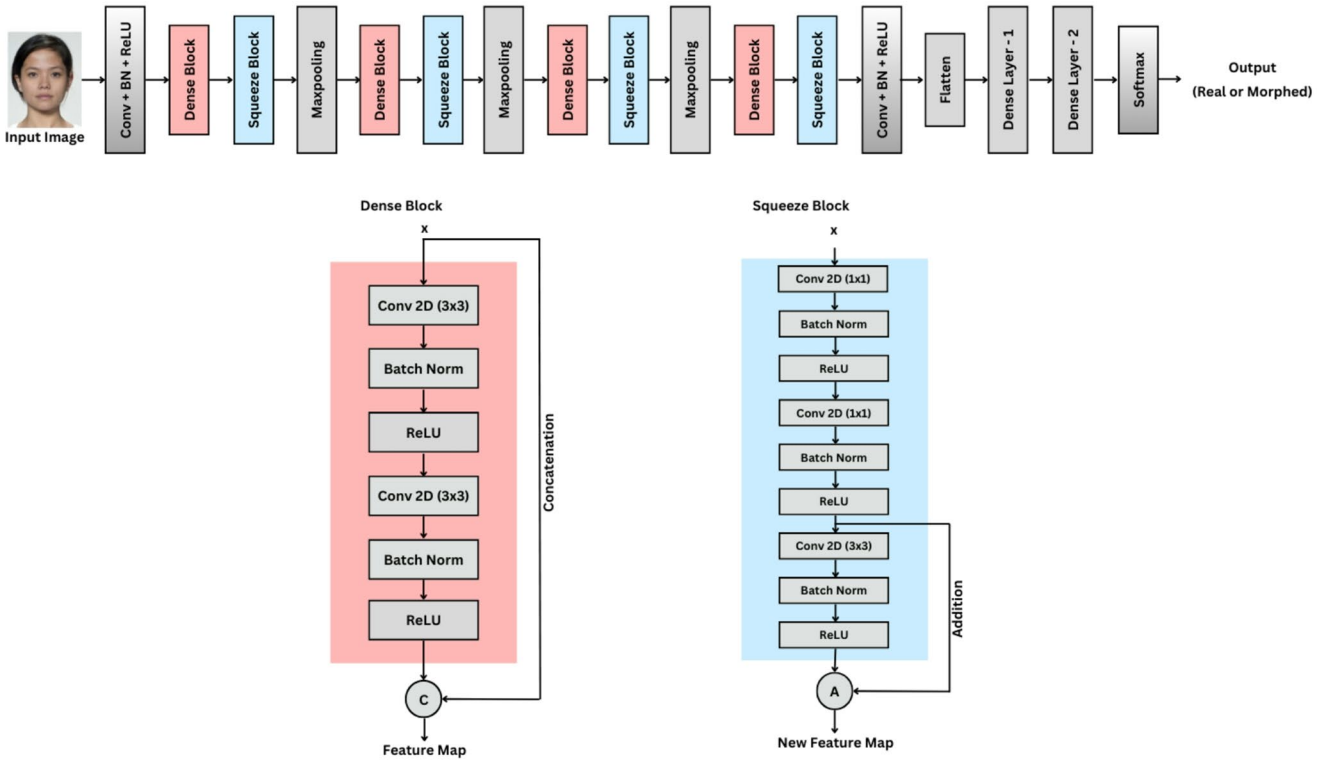

Fig. 5 Block Diagram for EfficientNet model.

block promotes efficient feature extraction using depthwise convolutions, batch normalization for stable training, and dropout layers to prevent overfitting. These blocks incorporate SE modules for further refinement and utilize skip connections to preserve essential information and facilitate improved gradient flow during backpropagation. Multiple MBConv blocks are stacked to form what are known as Efficient Blocks, which progressively capture increasingly high-level feature representations as data flows through the network. The final output layer of EfficientNet consolidates these deeply refined features using global average pooling or, alternatively, fully connected layers, before passing them through a softmax or sigmoid activation function to produce the final classification outputs.

In the context of this study, the EfficientNet-B0 variant was employed due to its ability to provide a strong balance between computational speed and prediction accuracy. EfficientNet-B0 employs a unique compound scaling method, which uniformly increases network depth, width, and input resolution according to fixed scaling coefficients, thereby achieving high accuracy even under resource constraints. Experimental configurations for EfficientNet-B0 in this study included the use of the Adam optimizer, an initial learning rate set to 0.0001 with a ReduceLROnPlateau scheduler (factor = 0.5, patience = 2), and a batch size of 32, with training conducted over 10 epochs. In addition, a weight decay of 0.00001 and dropout regularization with a rate of 0.3 were incorporated to further mitigate overfitting. All input images were resized to 224 × 224 pixels, and the model was trained using binary cross-entropy as the loss function. Furthermore, data augmentation techniques such as random horizontal flips and random rotations were applied to improve the robustness and generalization capability of the model during training.

The VGG model, depicted in Fig. 6, is a convolutional neural network (CNN) architecture renowned for its straightforward design and strong performance in image classification tasks. It starts with an input layer that receives images as multidimensional feature maps representing pixel intensities. The network is structured into five sequential convolutional blocks, each progressively extracting increasingly complex and abstract features from the input. Within each block, multiple convolutional layers utilize 3×3 filters to effectively capture spatial patterns such as edges, textures, and shapes. These convolutional layers are followed by Rectified Linear Unit (ReLU) activation functions, which introduce non-linearity to help the model learn complex representations. At the end of each block, a max-pooling layer reduces the spatial dimensions, preserving the most salient features while lowering computational demand. The number of filters increases from 64 in the initial block up to 512 in the deeper blocks, enabling the extraction of finer details. After the final convolutional block, the resulting feature maps are flattened using global pooling layers, and

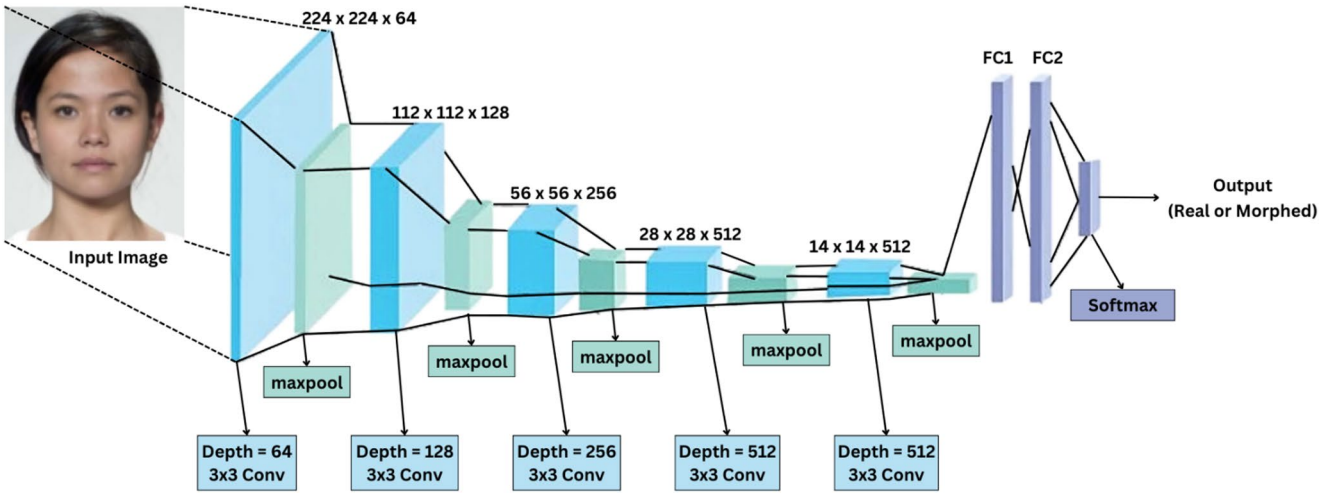

Fig. 6 Block Diagram for VGG model

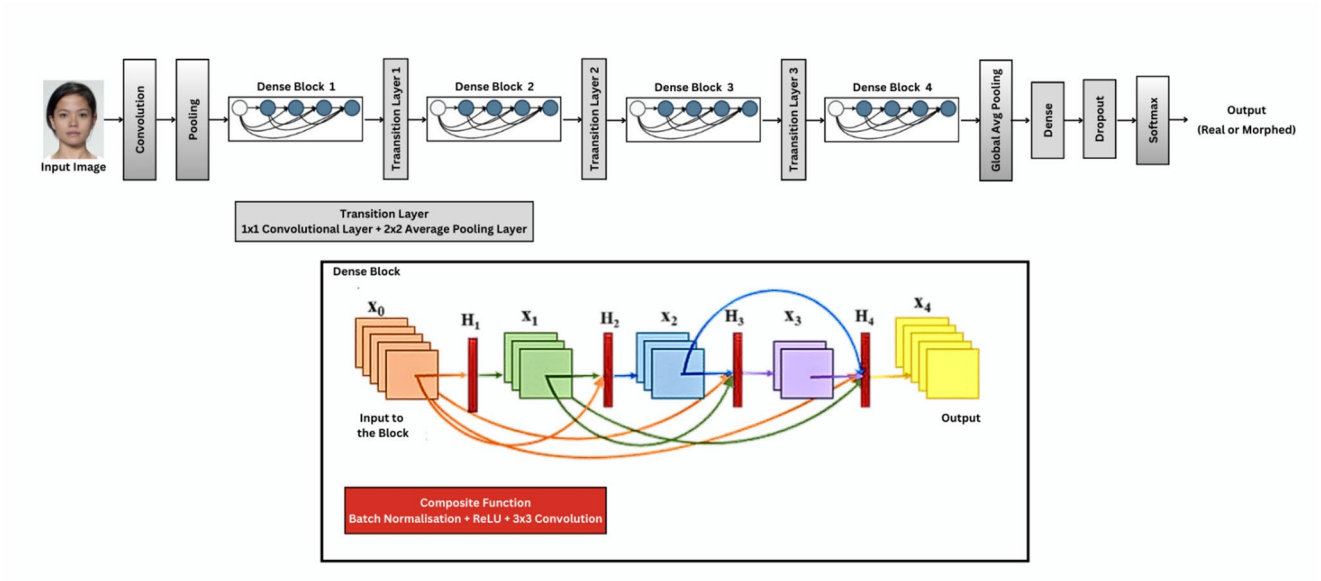

Fig. 7 Block Diagram for DenseNet model

fully connected layers then aggregate these extracted features. The output layer employs a softmax activation function to classify the image into predefined categories. This combination of simplicity, uniform design, and hierarchical feature learning makes VGG a powerful yet computationally intensive framework for image recognition.

In this study, the VGG16 variant was chosen as the baseline model due to its well-known uniform architecture of stacked convolutional layers. ImageNet pre-trained weights for VGG16 were transferred and partially unfrozen for fine-tuning, allowing the model to adapt to the specific task. Experiments utilized the Adam optimizer with a learning rate of 0.0001, running for 20 epochs with a batch size of 32. To reduce overfitting, a weight decay of 0.00001 was applied along with a relatively higher dropout rate of 0.5 in the fully connected layers. The input images were resized to 224×224 pixels, and binary cross-entropy served as the loss function. Additionally, data augmentation techniques such as random rotations and horizontal flips were employed to enhance model robustness and generalization during training.

The DenseNet-121 model, depicted in Fig. 7, employs a densely connected architecture to efficiently extract hierarchical feature representations from input images. The initial feature maps (F_{enc}) are enhanced using a Convolutional Block Attention Module (CBAM), which selectively focuses on important spatial and channel-wise information to generate refined features (F). These features are further processed through a Context Refinement Module (CRM), integrating broader contextual information and producing

enriched representations (F_{CRM}). A Global Average Pooling (GAP) layer then reduces the spatial dimensions, ensuring a compact yet informative feature vector. This vector is passed through a dense layer with a sigmoid activation function, enabling the model to output probabilities for binary or multi-class classification. Integrating DenseNet's dense connectivity with attention and context refinement modules results in robust feature extraction and improved image recognition performance.

For experimental evaluation, DenseNet-121 was implemented with pre-trained ImageNet weights and fine-tuned on the target dataset. The model was trained using the Adam optimizer (learning rate 0.0001, ReduceLROnPlateau scheduling), a batch size of 32, and 10 epochs. Regularization measures included a weight decay of 0.00001 and a dropout rate of 0.3 in the classification head. The model received inputs resized to 224×224 pixels and used binary cross-entropy as the loss function. Data augmentation techniques such as random horizontal flips and rotations were applied, consistent with those used for EfficientNet-B0, to ensure a fair comparison across models.

Table.2. Hyperparameters used for the Models

Parameter	Efficient-Net-B0	DenseNet121	VGG16	ResNet50
Optimizer	Adam	Adam	Adam	Adam
Initial Learning Rate	0.0001	0.0001	0.0001	0.0001
Learning Rate Schedule	ReduceLROnPlateau (factor=0.5, patience=2)	ReduceLROnPlateau (factor=0.5, patience=2)	ReduceLROnPlateau (factor=0.5, patience=2)	ReduceLROnPlateau (factor=0.5, patience=2)
Epochs	10	10	20	10
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy
Input Image Size	224x224	224x224	224x224	224x224

Table.2 summarizes the key hyperparameters used for training the four deep learning models evaluated in this study: EfficientNet-B0, DenseNet121, VGG16, and ResNet50. It includes details such as optimizer choice, learning rate, epochs, loss function and input image size. The table provides a clear and comprehensive overview to facilitate reproducibility and comparison across models. This standardized presentation ensures transparency in the experimental setup and supports the validation of reported results.

3.4.2 Digital forensics techniques

The digital forensics stage delves deeper into the image's composition, using scientific and algorithmic techniques to uncover hidden signs of manipulation. This step focuses on analyzing both the content and metadata of the image to identify discrepancies that may not be visible during visual inspection. One effective technique involves examining compression artifacts, as edited areas of an image often display varying compression levels compared to untouched regions. Similarly, pixel-level analysis is conducted to detect subtle inconsistencies, while metadata, including timestamps, GPS data, and editing history, is scrutinized for evidence of tampering.

Error Level Analysis (ELA) is a crucial tool used in this phase, highlighting differences in compression by visualizing regions of the image that are more likely to have been altered. Another method, Photo Response Non-Uniformity (PRNU), analyzes variations in the camera's unique sensor pattern noise, a digital fingerprint that can reveal whether parts of the image have been replaced or edited. By combining these techniques, digital forensics offers objective evidence of manipulation, enhancing the reliability of the findings and providing a deeper understanding of the alterations.

3.4.3 Visual inspection

The initial step in the analysis phase involves visual inspection, where human expertise is employed to identify visible or subtle irregularities in the image. This step emphasizes evaluating visual elements for signs of tampering. For example, discrepancies in lighting and shadows, such as incorrect shadow directions or inconsistent brightness levels, can point to possible manipulations, as tampered sections often deviate from the natural lighting observed in the rest of the image. Additionally, edge and boundary inconsistencies are carefully examined, as abrupt transitions or uneven blending may signal alterations where elements have been introduced or modified.

When dealing with facial morphing, particular attention is given to features such as the eyes, nose, and mouth, assessing their alignment, proportionality, and symmetry. Irregularities in these features often indicate tampering. Moreover, differences in texture and resolution across the image are scrutinized, as incorporating elements from different sources often results in mismatched clarity or inconsistent textures. By carefully analyzing these visual cues, this step lays the foundation for identifying potential manipulations, enabling the application of more technical tools in the subsequent stages.

3.4.4 Cross-referencing and validation

The final step of the analysis phase is dedicated to cross-referencing and validation, which is essential for ensuring the accuracy, consistency, and reliability of results obtained from different assessment methodologies. This process involves a systematic comparison of findings drawn from visual inspection, digital forensic techniques (such as PRNU and metadata analysis), and AI/ML-based detection models. The objective is to identify concordances and resolve discrepancies, ultimately synthesizing these insights to produce a comprehensive and well-rounded evaluation of the image's authenticity.

A critical component supporting this process is the use of trusted image databases and reference datasets, which provide benchmark comparisons to validate the investigation's outcomes and verify the origin and integrity of the images under analysis. The framework also employs iterative cross-validation techniques, repeating the analysis using multiple folds or rounds to reinforce the reliability of the conclusions. This iterative approach to testing guarantees a thorough investigation by minimizing risks of false positives and negatives and ensuring the robustness of the derived results.

Within this multi-method framework, findings from high-confidence AI models are not viewed in isolation. Instead, they are corroborated by traditional forensic analyses such as PRNU, a physics-based approach that inspects unique sensor noise patterns characteristic of individual imaging devices. PRNU analysis provides crucial physical-level evidence that complements the pattern recognition capabilities of AI models. When AI predictions and forensic signals such as PRNU or metadata analysis produce conflicting results—e.g., an AI model suggests the image is authentic while PRNU detects tampering—the human analyst weighs these multiple lines of evidence critically. This complementary relationship leverages the strengths of each methodology, with AI providing rapid, large-scale pattern recognition, and forensics supplying fine-grained, device-specific validation cues. Moreover, metadata analysis serves as another vital pillar of validation, revealing anomalies or inconsistencies in image creation timestamps, device information, or editing history that may escape automated AI detection. Cross-referencing metadata with AI and PRNU results amplifies the investigative rigor, ensuring that discrepancies are systematically flagged and explored.

Another key feature of the validation process is the application of visual inspection by human experts. Despite advances in AI and forensic algorithms, human intuition and domain expertise remain invaluable, especially for complex or borderline cases. Visual inspection helps contextualize automated findings and detect subtle artifacts, lighting inconsistencies, or unnatural blending that might not be fully captured by computational methods. Ultimately, the human expert synthesizes all inputs—visual cues, AI outputs, PRNU signals, and metadata consistency—applying professional judgment to prioritize and resolve conflicting evidence. This ensures that the final authenticity assessment is transparent, high-confidence, and legally robust, particularly for applications requiring rigorous evidentiary standards such as forensic investigations and identity verification.

By integrating traditional expertise, rigorous scientific methodologies, and state-of-the-art AI technologies within a cross-referenced, validated workflow, the framework achieves a trustworthy and comprehensive forensic evaluation. This multi-pronged analysis phase serves as an indispensable foundation for counterfeit image detection, delivering stakeholders high confidence in the veracity of its conclusions and supporting informed decision-making. The iterative nature of cross-referencing and validation also facilitates continuous improvement of the detection models. Insights gained from discrepancies and human expert reviews feed back into model refinement, enabling the AI to learn from forensic evidence and expert knowledge progressively. This feedback loop ensures that both forensic processes and AI models evolve alongside emerging image manipulation techniques, maintaining relevance and efficacy against increasingly sophisticated counterfeit content.

Thus, cross-referencing and validation form a critical nexus within the forensic framework that balances computational power with physical evidence and human expertise. This comprehensive approach mitigates risks posed by relying solely on a single detection modality, thereby enhancing the overall reliability and trustworthiness of image authenticity determinations. Such rigor is vital to support the growing need for dependable verification in areas spanning cybersecurity, digital forensics, content moderation, and legal proceedings.

3.5 Reporting stage

The Reporting Stage is a critical aspect of the forensic framework, ensuring that all findings and processes from the analysis phase are comprehensively recorded for transparency, reproducibility, and use in subsequent legal or investigative contexts. This phase plays a pivotal role in preserving the integrity of the investigation and providing a structured record that can be revisited or scrutinized by stakeholders.

In this phase, every tool and technique employed during the analysis is meticulously documented. This includes detailing the software, digital forensic methodologies, and AI or ML models used. By recording this information, the

investigation ensures that all methods are replicable and that other investigators can validate the processes and outcomes. Such thorough documentation also supports transparency, which is essential when presenting findings in court or other formal settings.

A comprehensive record of observations made during the analysis is another vital component. This involves detailing any anomalies, inconsistencies, or manipulations identified through various techniques. For example, discrepancies in lighting, irregularities in edges, or metadata inconsistencies must be described with precision. Each finding is tied to the specific methods or tools used to uncover it, providing a clear trail of evidence that links the observations to the analytical processes.

The results generated by AI-based detection models are also carefully recorded in this phase. These results include not only the classification outcomes, such as whether an image is authentic or counterfeit, but also the confidence scores or probabilities associated with these determinations. These metrics offer a quantitative foundation for the analysis, making the findings objective and defensible. When paired with detailed explanations of the AI model's decision-making process, this information strengthens the credibility of the investigation.

Visual evidence forms a key part of the documentation. Annotated images, such as those produced through techniques like Grad-CAM heatmaps or Error Level Analysis (ELA), are created to highlight areas of interest or suspected manipulation. These visual aids are invaluable for communicating findings to non-technical stakeholders, such as legal professionals or investigators, as they provide an intuitive understanding of the evidence. The visualizations not only support the analysis but also enhance the presentation of the investigation's results.

Additionally, detailed logs of all processes are maintained to ensure that every step of the investigation is reproducible. These logs include information about data inputs, parameter settings, intermediate outputs, and the progression of each analytical step. This level of detail is especially important in forensic investigations, where the ability to reproduce results can significantly impact the credibility of the findings. The logs also provide a reference for improving or refining future analyses [34].

Overall, the documentation phase ensures that the entire forensic process is transparent, verifiable, and defensible. By capturing the tools, techniques, observations, AI results, and visual evidence in a structured manner, it creates a robust and trustworthy record of the investigation. This thorough and well-organized documentation serves not only as a reference for stakeholders but also as critical evidence in contexts that demand a high degree of rigor and accuracy.

3.6 Findings communication stage

The Findings Communication Stage serves as the final and equally critical component of the forensic framework, focusing on organizing and conveying the findings of the analysis in a structured and comprehensible manner. This phase is essential for translating technical insights into actionable information for stakeholders, including legal teams, investigators, or decision-makers, who may not possess technical expertise in forensic methodologies. The process involves summarizing the results, providing visual aids to enhance understanding, and presenting conclusions and recommendations in a clear and concise format.

A comprehensive report is prepared to document the investigation, ensuring that all findings are systematically detailed and supported with evidence. This report includes a thorough summary of the analysis process, outlining the methods used, the results obtained, and the conclusions drawn. To make the findings more accessible, the report incorporates visual elements such as charts, graphs, and annotated images that highlight localized regions of manipulation. For instance, heatmaps generated by AI/ML models or outputs from tools like Error Level Analysis (ELA) can be used to visually demonstrate the specific areas of the image where tampering has occurred. These visual aids not only enhance clarity but also provide compelling evidence to support the investigation's conclusions.

The conclusions are presented in a straightforward, non-technical language to ensure they are easily understood by stakeholders who may not have a background in digital forensics or AI-based techniques. This section synthesizes the findings into actionable insights, outlining the authenticity of the analyzed image and the methods used to reach this determination. The goal is to present the results in a manner that is both clear and persuasive, enabling stakeholders to make informed decisions based on the evidence provided.

Additionally, the report includes recommendations to address the broader implications of the findings. These may involve suggesting measures to prevent future cases of image morphing, such as adopting stricter authentication protocols or implementing advanced verification systems in image workflows. The recommendations might also highlight areas where current systems can be improved, such as increasing the use of AI/ML-based tools to enhance detection capabilities or integrating metadata validation processes to identify tampering early. These suggestions aim to not only resolve the current case but also mitigate similar risks in the future.

In conclusion, the presentation phase acts as the bridge between technical analysis and actionable outcomes, ensuring that the investigation's findings are effectively communicated to all relevant stakeholders. By combining detailed

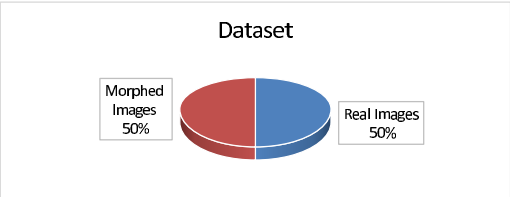

Fig. 8 Sample Images from Dataset Used.

reporting, clear visualizations, and practical recommendations, this phase plays a pivotal role in supporting legal and investigative efforts while fostering improvements in image verification and security practices.

4 Experimental setup and result analysis for detecting and verifying morphed images

To evaluate the proposed system, we applied it to an image taken from a real case involving the dissemination of a morphed image of an Indian politician on social media. This incident had significant public and political repercussions, making it a critical example for analysis. The framework was employed to scrutinize the morphed image for inconsistencies and verify its authenticity.

The proposed system integrated a combination of advanced forensic techniques, including metadata analysis, error level analysis (ELA), and machine learning-based methods to detect morphing. In particular, AI-powered models played a pivotal role in the detection process, employing convolutional neural networks (CNNs) and deep learning architectures trained to identify subtle artifacts introduced during image morphing. These models were able to detect minute discrepancies in texture, lighting, and spatial coherence that are often imperceptible to the human eye.

Key findings from the analysis revealed deliberate alterations in facial features and background elements, consistent with common manipulation patterns observed in synthetic and edited media. This case study highlights the robustness of the proposed framework and its ability to leverage AI models for precise and efficient detection of digital forgeries. These results underscore its potential application in combating misinformation, safeguarding authenticity, and protecting individuals' reputations in sensitive contexts.

After the morphed image mentioned above was identified and collected, first we analysed it using advanced AI models such as EfficientNet, DenseNet, ResNet, and VGG. These deep learning models were specifically selected for their ability to effectively extract and analyze intricate image features. To prepare these models for the analysis, they were trained on a dataset as shown in Fig. 8, comprising 140,000 images, equally divided between 70,000 real images and 70,000 morphed images as shown in pie chart below.

The morphed images in the dataset were generated using various sophisticated manipulation techniques designed to mimic real-world scenarios, ensuring the models would perform effectively across diverse cases. The techniques included:

1. Face Morphing: A process where facial features from two individuals are blended to create a seamless composite image, often used in identity fraud or spoofing attempts.
2. Deepfake Generation: Using generative adversarial networks (GANs) to manipulate facial expressions, swap faces, or create entirely artificial yet highly realistic images and videos.
3. Image Splicing: Combining sections from multiple images into a single composite image, altering the original context and creating a fabricated narrative.
4. Region Duplication (Cloning): Copy-pasting elements within an image to either remove unwanted details or replicate specific features for manipulation.
5. Attribute Editing: Altering visual properties such as skin tone, background, lighting, or texture using advanced editing tools or AI-based techniques to subtly manipulate the image while maintaining realism.

The dataset, which merges sources from publicly available datasets including the "140k Real and Fake Faces" dataset from Kaggle [32] and the Deepfake Detection Dataset from Michigan State University [33], was carefully sampled. A total of 140,000 images—70,000 real and 70,000 fake—were randomly selected from the combined pool. This dataset was chosen for training AI models due to its diverse and sophisticated image manipulation techniques, including face morphing, deepfake generation, image splicing, region duplication, and attribute editing. These realistic and complex manipulations provide a challenging yet realistic scenario, allowing models to effectively learn to detect subtle image forgeries and enhance their performance in real-world applications such as identity verification and digital forensics. The dataset used was designed to encompass a wide range of scenarios, including variations in lighting, resolution, facial angles, and backgrounds, to ensure that the models were robust and capable of identifying manipulations under diverse conditions.

To ensure effective training and evaluation of the models, the dataset was then split into three subsets:

- Training Set (100,000 images): The majority of the dataset was used for training the models, enabling them to learn distinct patterns, features, and anomalies that differentiate real images from morphed ones.

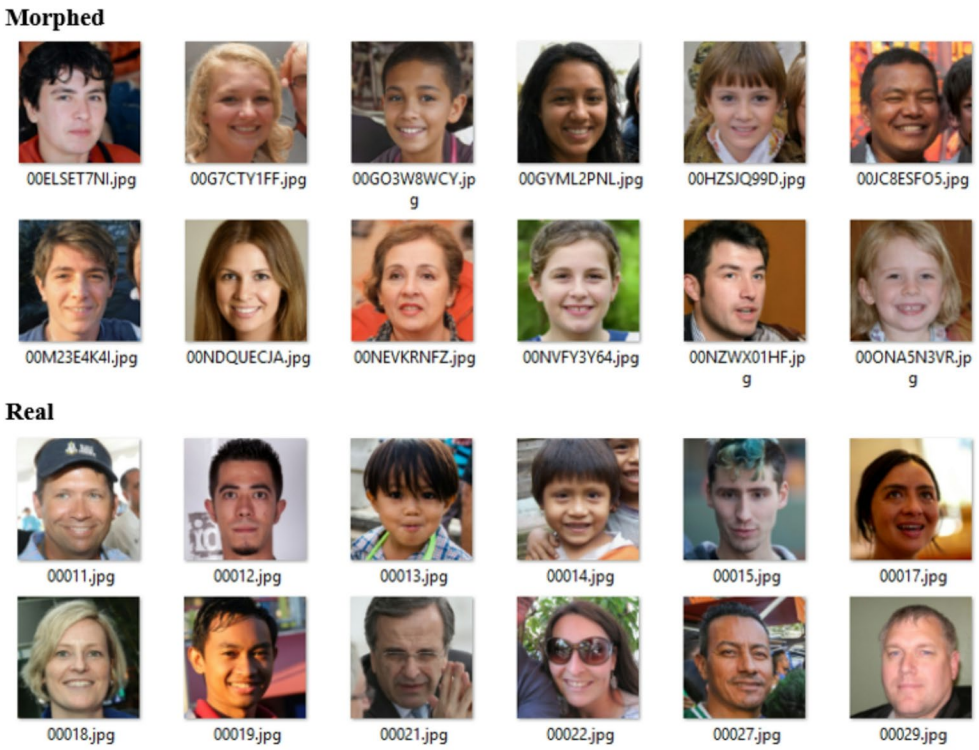

Fig. 8 (continued)

- Validation Set (20,000 images): A portion of the dataset was used to fine-tune the models, allowing for optimization of hyperparameters and prevention of overfitting during training.
- Test Set (20,000 images): The remaining dataset was set aside for unbiased evaluation of the models on previously unseen data, providing insights into their real-world applicability and accuracy.

This structured training process allowed the models to achieve high accuracy in identifying manipulations, even those involving subtle changes that might evade human detection. The inclusion of diverse and realistic morphing techniques ensured the framework's capability to handle a broad spectrum of manipulation methods.

The graphs below illustrate the accuracy and loss metrics of the trained models.

The presented graphs in above figure (Fig. 9) showcase the training and validation performance metrics for the EfficientNet model over 10 epochs. In the accuracy graph, both training and validation accuracies demonstrate a rapid improvement in the initial epochs, followed by stabilization

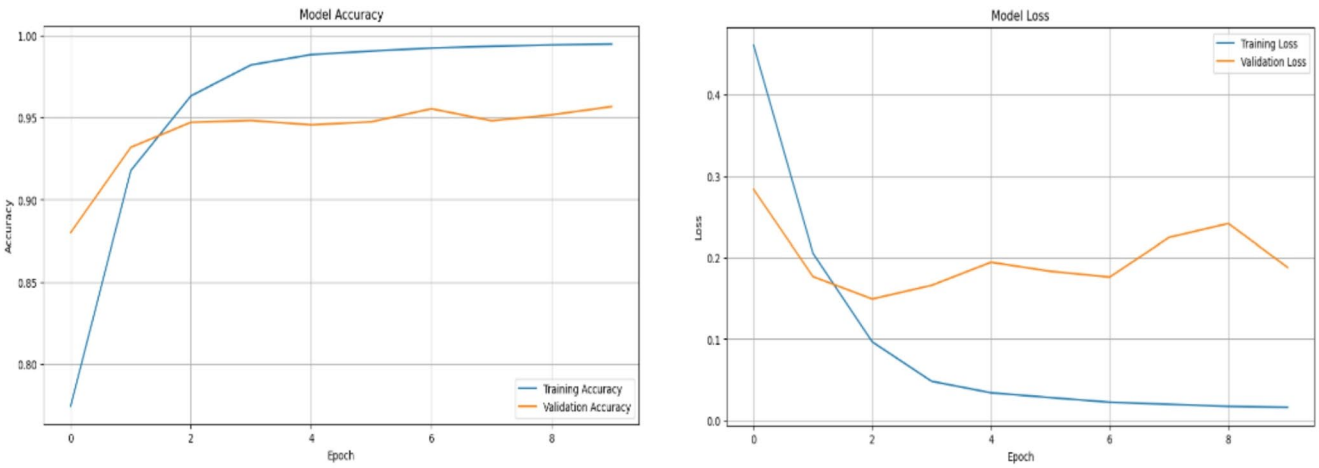

Fig. 9 Accuracy and Loss Graph for EfficientNet Model

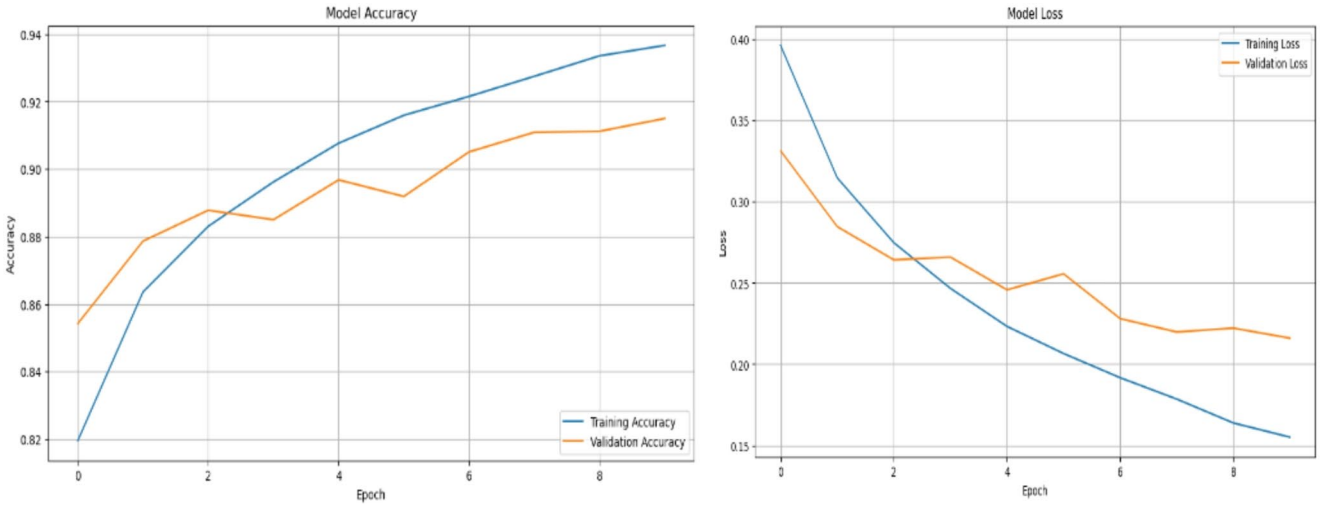

Fig. 10 Accuracy and Loss Graph for DenseNet121 Model

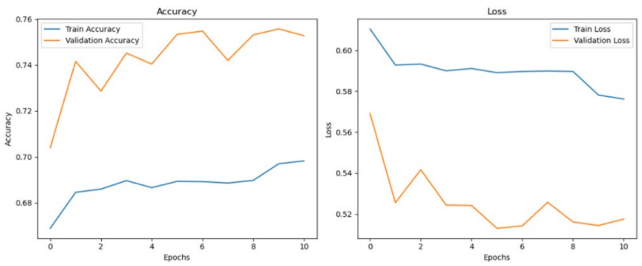

Fig. 11 Accuracy and Loss Graph for ResNet50 Model

at high accuracy levels. The training accuracy approaches 100%, while the validation accuracy plateaus slightly below, reflecting the model's strong ability to generalize to unseen data.

The loss graph indicates a significant and consistent reduction in both training and validation losses during the early epochs, highlighting effective optimization and convergence of the model. The training loss approaches near-zero values, while the validation loss stabilizes at a slightly higher value. This alignment of accuracy and loss trends confirms the model's robust learning capability and its efficiency in capturing the underlying patterns of the dataset, making EfficientNet a highly effective choice for the classification task.

The graphs (in Fig. 10) above illustrate the training and validation performance of the DenseNet121 model over 10 epochs. In the accuracy graph, both training and validation accuracies show a consistent increase, with training accuracy steadily climbing toward 94%. The validation accuracy also improves progressively, reaching a slightly lower value, which indicates that the model is effectively learning from the data and generalizing well to unseen samples.

The loss graph depicts a downward trend in both training and validation losses throughout the training process. The training loss decreases consistently, reflecting the model's optimization and its ability to minimize prediction errors. Similarly, the validation loss exhibits a steady decline, although at a slightly slower rate compared to the training loss. These trends confirm that DenseNet121 demonstrates strong learning capabilities and is well-suited for the classification task, achieving a balance between accuracy and loss minimization.

The graphs (in Fig. 11) illustrate the performance of the ResNet50 model during its training and validation phases over 10 epochs. In the accuracy graph, both the training and validation accuracies exhibit a steady upward trend. Training accuracy gradually increases as the model learns, while validation accuracy follows a similar trajectory, indicating effective learning and good generalization to unseen data.

The loss graph highlights a consistent reduction in both training and validation losses throughout the epochs. Training loss decreases progressively, showcasing the model's ability to optimize its predictions and minimize errors. Validation loss also declines over time, though at a slightly slower pace, suggesting that the model retains its generalization capabilities without overfitting. Overall, ResNet50 demonstrates reliable learning dynamics, achieving a balance between improving accuracy and reducing losses for the classification task.

The above figure (Fig. 12) depicts the training and validation performance metrics for the VGG-16 model across 20 epochs. In the accuracy graph, both training and validation accuracies exhibit a steady increase as the number of epochs progresses. This indicates that the model's ability to correctly classify inputs improves consistently over time. The validation accuracy remains slightly higher than the training accuracy, demonstrating robust learning and effective adaptation to unseen data.

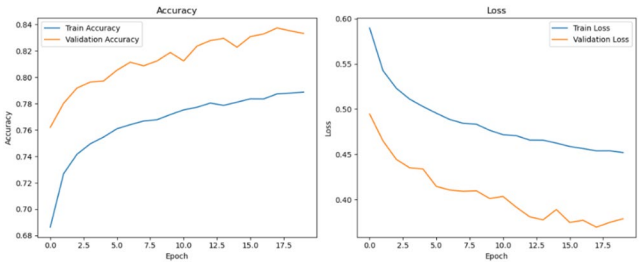

Fig. 12 Accuracy and Loss Graph for VGG-16 Model

In the loss graph, both training and validation losses show a consistent decline, confirming the model's optimization as it minimizes the error between predicted and true outputs. The gradual reduction in loss indicates stable learning dynamics, with the model improving its predictions iteratively. The alignment of trends in accuracy and loss validates the model's reliability and its effectiveness in learning complex patterns in the dataset. These results confirm the suitability of the VGG-16 architecture for the classification task, reflecting its capacity for high performance over iterative training.

The confusion matrices (in Fig. 13) illustrate the binary classification performance of four deep learning models: EfficientNet, DenseNet-121, VGG16, and ResNet-50. EfficientNet achieves the highest accuracy, with a large number of true positives (9424 for "fake" and 9662 for "real") and minimal misclassifications, indicating robust performance.

DenseNet-121 performs slightly worse, with increased false positives and false negatives but still maintains strong classification ability. VGG16 shows moderate performance, with more significant misclassification errors, likely due to its simpler architecture. ResNet-50 exhibits the weakest performance, with the highest number of false positives and false negatives, particularly struggling to classify "fake"

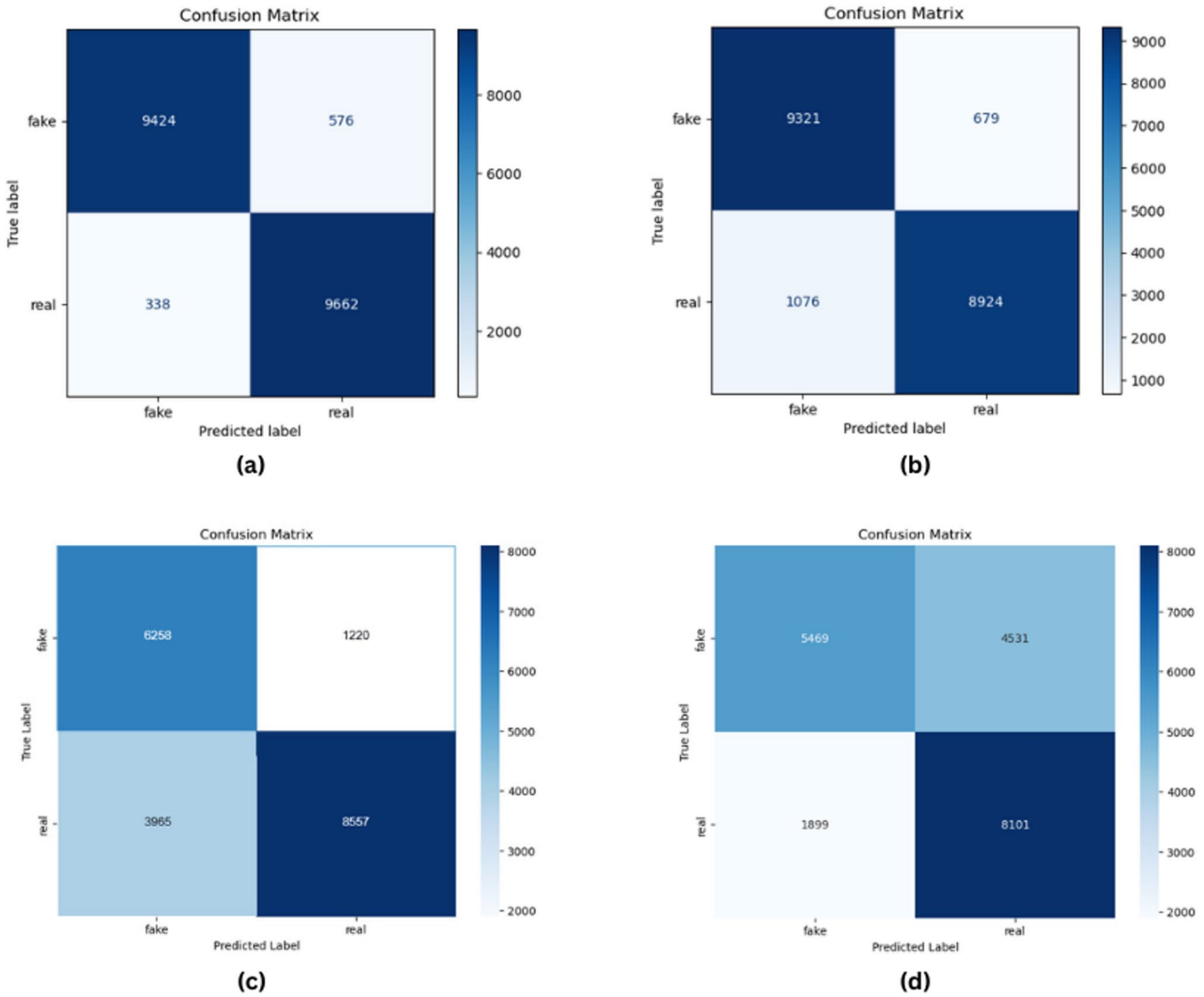

Fig. 13 Comparison of Confusion Matrices: (a) EfficientNet, (b) DenseNet-121, (c) VGG16, and (d) ResNet-50 in Binary Classification.

Table 3 provides a comparative analysis of the performance of four deep learning models—VGG-16, EfficientNet, DenseNet121, and ResNet50—based on their training, validation, and test accuracies. Among the models, EfficientNet demonstrates the highest overall performance, achieving a training accuracy of 99.47%, validation accuracy of 95.96%, and test accuracy of 95.42%. This indicates its superior ability to generalize across unseen data. DenseNet121 also performs well, with training, validation, and test accuracies of 93.80%, 91.74%, and 91.22%, respectively, showcasing its strong learning capabilities. VGG-16, though older, achieves competitive results with a training accuracy of 78.02%, validation accuracy of 84.64%, and test accuracy of 83.74%, highlighting its reliability in simpler scenarios. On the other hand, ResNet50 shows comparatively lower performance, with training accuracy at 69.73%, validation accuracy at 77%, and test accuracy at 75%, suggesting challenges in fully adapting to the dataset complexities. Overall, the table underscores the superiority of modern architectures like EfficientNet and DenseNet121 for tasks requiring robust and accurate classifications. Refer to the bar chart below for a clearer comparison. Figure 14 shows comparison of Training Accuracy, Validation Accuracy and Test Accuracy for EfficientNet, DenseNet-121, VGG16, and ResNet-50

Model	Training Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)
VGG-16	78.02	84.64	83.74
EfficientNet	99.47	95.96	95.42
DenseNet121	93.80	91.74	91.22
ResNet50	69.73	77	75

instances accurately. Overall, EfficientNet outperforms the others, followed by DenseNet-121, with VGG16 and ResNet-50 lagging behind.

The following table presents a comparison of the accuracies achieved by various AI models trained on the dataset:

Table.3. Comparison of Model Accuracies

Fig. 15 Image showing AI model prediction

After successfully training the AI models, image taken from a real-world case, featuring a morphed image of a politician, was analyzed using the trained AI models. This image, manipulated to blend facial features or alter attributes for misinformation purposes, was evaluated by all four models. The predictions generated provided critical insights into the models' ability to detect and accurately classify such real-world manipulations, demonstrating their practical applicability in identifying altered images.

The image shown above (in Fig. 15) represents comprehensive analysis, combining the predictions from all four trained AI models, indicating that the image is "fake" with an average confidence level of 86.43% and "real" with an average confidence level of 13.57%. This collective assessment demonstrates the models' ability to effectively detect manipulated content, even in scenarios involving highly realistic alterations. The results highlight the importance

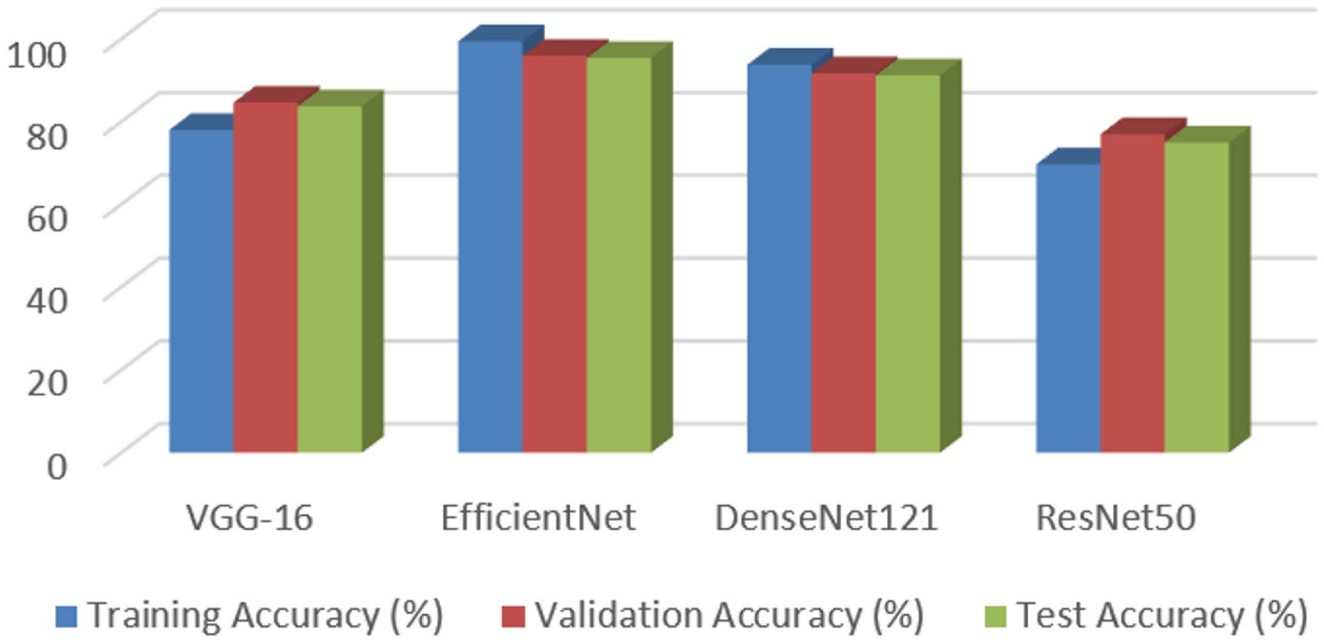

Fig. 14 Comparison of Training Accuracy, Validation Accuracy and Test Accuracy for EfficientNet, DenseNet-121, VGG16, and ResNet-50

of leveraging multiple AI models to improve accuracy and reliability in identifying morphed images in real-world applications.

In addition to AI models, forensic tools were employed to analyze the image, and the results corroborated the findings that the image had been morphed. Error Level Analysis (ELA) and noise analysis were conducted to detect inconsistencies and anomalies within the image, providing further evidence of manipulation.

The ELA highlighted areas with uneven compression levels, indicating alterations, while the noise analysis revealed discrepancies in texture and pixel patterns that are typically present in morphed images. These forensic techniques, shown in the figure above (Fig. 16), reinforced the conclusion that the image is not authentic, further validating the predictions made by the AI models.

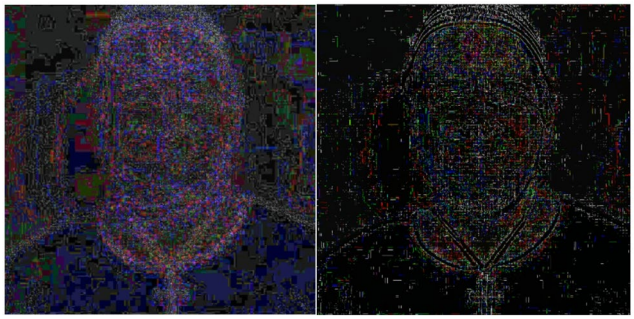

Fig. 16 Error Level Analysis and noise analysis of Morphed Image

5 Discussion and limitations

The proposed system for the detection and verification of counterfeit images offers a structured methodology to identify and analyze manipulated images, particularly morphed and deepfake content. By incorporating AI/ML-based detection techniques alongside traditional methods, this system enhances the reliability of image authentication. The six-stage process—identification, preservation, collection, analysis, documentation, and presentation—ensures a comprehensive approach to handling evidence. Implementing this system can help mitigate threats such as identity fraud, cyber extortion, online harassment, and biometric security breaches.

Despite its effectiveness, certain challenges remain. AI-driven detection models heavily depend on extensive and diverse datasets for training. However, the lack of standardized and high-quality datasets for detecting morphing manipulation poses a significant hurdle. Additionally, the effectiveness of detection tools can vary due to differences in image sources, editing techniques, and the sophistication of morphing methods. The computational feasibility of deploying deep learning models such as EfficientNet-B0, DenseNet-121, and VGG16 in real-time forensic or law enforcement settings depends on several factors, including model complexity, hardware availability, and latency requirements. While these models demonstrate high accuracy and robustness in detecting sophisticated image manipulations like morphing and deepfakes, their computational demands—both in terms of memory and processing power—can pose significant challenges for real-time applications, especially in resource-constrained environments typical of many operational forensic and law enforcement contexts. EfficientNet-B0 is designed with compound scaling principles that balance network depth, width, and input resolution to achieve an optimal trade-off between accuracy and efficiency. It is more lightweight compared to deeper variants and models like DenseNet-121 and VGG16, making it more appropriate for near real-time processing. However, even EfficientNet-B0 may require GPU acceleration or specialized hardware to meet strict latency requirements in operational settings. DenseNet-121, with its dense connectivity pattern, offers improved feature propagation and accuracy at the cost of increased computational complexity and memory consumption, which may limit its usability in real-time deployments without hardware acceleration. VGG16, although simpler and widely adopted as a baseline, is relatively heavy due to its large number of parameters and computational operations, making it less suitable for time-sensitive or resource-constrained forensic tasks without significant optimization. For deployment outside controlled lab conditions, where real-time processing and resource limitations are critical, lightweight or optimized model variants are recommended. Examples include EfficientNet-lite versions, MobileNet, ShuffleNet, or SqueezeNet, which are specifically architected for efficiency with smaller model sizes and reduced computational costs. These models maintain competitive accuracy with a fraction of the inference time and memory footprint, enabling deployment on edge devices or standalone forensic workstations. Additionally, techniques such as model pruning, quantization, and knowledge distillation can further reduce the size and complexity of standard deep learning architectures without substantial loss in performance. These optimizations can facilitate faster inference, lower energy consumption, and broader accessibility in real-world law enforcement scenarios.

Furthermore, adaptive model selection based on available hardware and required throughput can be beneficial—deploying heavier models like DenseNet-121 when sufficient resources are available and switching to lightweight alternatives when constraints are tighter. This tiered approach, combined with continual updates to keep pace with evolving image manipulation techniques, can help

maintain model effectiveness while respecting operational limitations.

Furthermore, as digital manipulation techniques advance, attackers continuously refine their methods to evade forensic detection, making it difficult to develop a foolproof detection system.

One of the major obstacles in tackling crimes related to image morphing and deepfake content is the absence of well-defined legal provisions in India. While existing laws under the Information Technology (IT) Act, 2000, and the Indian Penal Code (IPC) cover broader cyber offenses, they do not explicitly address the forensic detection or criminalization of deepfake and morphed images. For instance, Section 66D of the IT Act penalizes identity fraud, but it does not specify guidelines for the forensic validation of manipulated images. Similarly, Sections 67 and 67A focus on prohibiting the distribution of obscene content but do not provide forensic frameworks or legal guidelines for AI-based evidence verification. Additionally, India currently lacks strict regulations that specifically address the creation, distribution, or misuse of deepfake technology, leaving law enforcement agencies with limited resources to prosecute offenders.

To bridge this legal gap, we must introduce dedicated legislation that explicitly defines and criminalizes offenses related to deepfake and morphed images. Additionally, forensic guidelines should be developed to establish the credibility of AI-based image authentication in legal proceedings. Strengthening the legal framework requires collaboration between policymakers, cybersecurity experts, and forensic professionals to ensure the effective prosecution of cybercriminals while safeguarding individual privacy and security.

Another notable limitation is the reliance on human expertise in forensic investigations. While AI-powered detection tools significantly enhance efficiency, expert forensic analysts are still needed to interpret results, verify anomalies, and ensure accuracy. This underscores the need for interdisciplinary collaboration between forensic analysts, cybersecurity specialists, and legal professionals to develop a well-rounded approach to combating the misuse of image manipulation technologies.

By addressing these limitations, future research can focus on improving detection accuracy, reducing computational costs, and advocating for stronger legal frameworks to effectively combat cybercrimes involving counterfeit images.

6 Conclusion

The increasing prevalence of digital image manipulation, particularly through morphing and deepfake techniques, presents significant challenges in cybersecurity and biometric security. The proposed system addresses these challenges by providing a structured methodology that combines traditional techniques with advanced AI/ML-based detection methods. This approach enhances investigations by ensuring systematic identification, preservation, analysis, and documentation of manipulated images, thereby strengthening the credibility and reliability of digital evidence. While the system offers significant improvements over existing methods, challenges such as dataset availability, evolving adversarial techniques, computational demands, and legal considerations must be addressed for effective real-world implementation.

While this research primarily utilizes well-established deep learning architectures such as EfficientNet, DenseNet, and VGG, it offers a valuable contribution by systematically evaluating these models within the context of counterfeit image detection in digital forensic applications. The study advances the field by integrating sophisticated modules like Convolutional Block Attention Modules (CBAM) and Context Refinement Modules (CRM) to enhance feature extraction and classification performance specifically for morphed and deepfake images. Moreover, the construction of a large and diverse dataset tailored to morphing detection, combined with rigorous preprocessing and augmentation strategies, addresses critical gaps in existing resources and supports robust model training and evaluation.

Beyond model performance, this work acknowledges and addresses the practical challenges of deploying these AI-driven solutions in real-time forensic and law enforcement environments, such as computational resource constraints and the evolving tactics of malicious actors. The inclusion of discussions on lightweight model alternatives and optimization techniques further bridges the gap between the theoretical efficacy and operational feasibility. Additionally, by highlighting the current legal shortcomings and proposing the need for dedicated forensic guidelines and legislation, the study situates technical advancements within the broader socio-legal framework essential for effective cybercrime prosecution.

Overall, this research not only validates the applicability of proven deep learning methods to forensic image authentication but also paves the way for future innovation focused on creating more specialized, resource-efficient, and legally compliant solutions that can keep pace with the rapidly evolving landscape of digital image manipulation.

Future research should focus on developing more resilient AI-based detection models, standardizing methodologies,

and collaborating with legal and regulatory bodies to establish robust policies against digital image manipulation crimes. By addressing these aspects, this system can serve as a critical tool in the fight against cybercrimes involving counterfeit images, ensuring enhanced security, privacy, and trust in digital environments.

Author contributions 1.Conceptualization: Rizwan ur Rahman, Praharsh.2.Methodology: Rizwan ur Rahman, Praharsh, Deepak Singh Tomar3.Investigation: Rizwan ur Rahman, Praharsh.4.Formal Analysis: Rizwan ur Rahman, Praharsh, Deepak Singh Tomar5.Writing: Rizwan ur Rahman, Praharsh.6.Review & Editing: Rizwan ur Rahman, Praharsh, Deepak Singh Tomar

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Vaghasiya, K.: *The latest ChatGPT statistics and user trends (2024–2025)*. URL: (2024)., November 28 https://wisernotify.com/blog/chatgpt-users/?utm_source=chatgpt.com
2. Nerdynav: 107 Up-to-Date ChatGPT Statistics & User Numbers [Nov 2024]. URL: (2024)., November 4 https://nerdynav.com/chatgpt-statistics/?utm_source=chatgpt.com
3. Singh, S.: *Number of ChatGPT users (January 2025)*. URL: (2025)., January 2 https://www.demandsage.com/chatgpt-statistics/?utm_source=chatgpt.com
4. Broz, M., Broz, M.: *Midjourney statistics (2025)*. URL: (2024)., December 20 https://photutorial.com/midjourney-statistics/?utm_source=chatgpt.com
5. Kumar, N.: *Midjourney Statistics 2025*: URL: (2024)., December 31 https://www.demandsage.com/midjourney-statistics/?utm_source=chatgpt.com
6. Mohapatra, D.: Morphed photos of women: State records most cases. URL: (2023)., December 5 https://timesofindia.indiatimes.com/city/bhubaneswar/morphed-photos-of-women-odisha-records-highest-cases-ncrb-report-reveals/articleshow/105741454.cms
7. Badgeri, N.N.G.: Families struggle to cope after some sextortion, loan app victims end lives. URL: (2023)., December 3 https://timesofindia.indiatimes.com/city/mumbai/families-struggle-to-cope-after-some-sextortion-loan-app-victims-end-lives/articleshow/105689716.cms
8. TIMESOFINDIA.COM. Delhi Police arrests UP man for trying to extort money from woman by using her morphed pictures. URL: (2023)., April 27 https://timesofindia.indiatimes.com/city/delhi/delhi-police-arrests-up-man-for-trying-to-extort-money-from-woman-by-using-her-morphed-pictures/articleshow/99813099.cms
9. Prajwal, D.S., Prajwal, D.S.: *Bengaluru: Housewife blackmailed with morphed photos after downloading loan app*. URL: (2024)., July 31 https://www.deccanherald.com/india/karnataka/bengaluru/bengaluru-housewife-blackmailed-with-morphed-photos-after-downloading-loan-app-3131508
10. Tnn: Gujarat: Ex-fiancé uploads morphed pictures. URL: (2019)., September 4 https://timesofindia.indiatimes.com/city/ahmedabad/ex-fianc-uploads-morphed-pictures/articleshow/70967592.cms
11. Ghosh, D.: Kol logs most photo-morph cases in India in '22: NCRB. URL: (2023)., December 5 https://timesofindia.indiatimes.com/city/kolkata/kolkata-tops-photo-morph-cases-in-india-in-2022-ncrb-report/articleshow/105740314.cms
12. Tnn: Rajasthan: Cyberbullying, e-frauds at record high. URL: (2019b)., October 23 https://timesofindia.indiatimes.com/city/jaipur/cyberbullying-e-frauds-at-record-high/articleshow/71712887.cms
13. Express News Service, & Express News Service: *2 post morphed pics of women online, arrested*. URL: (2023)., July 17 https://www.newindianexpress.com/states/tamil-nadu/2023/Jul/17/2-post-morphed-pics-of-women-online-arrested-2595640.html
14. Tanaka, M., Shiota, S., Kiya, H.: A detection method of operated fake-images using robust hashing. J. Imaging. **7**(8), 134 (2021)
15. Sharma, D.K., Singh, B., Agarwal, S., Garg, L., Kim, C., Jung, K.H.: A survey of detection and mitigation for fake images on social media platforms. Appl. Sci. **13**(19), 10980 (2023)
16. Killi, C.B.R., Balakrishnan, N., Rao, C.S.: Deep Fake Image Classification Using VGG-19 Model. Ingénierie des. Systèmes d'Information, **28**(2). (2023)
17. Hsu, C.C., Zhuang, Y.X., Lee, C.Y.: Deep fake image detection based on pairwise learning. Appl. Sci. **10**(1), 370 (2020)
18. Cartella, G., Cuculo, V., Cornia, M., Cucchiara, R.: Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. IEEE. Signal. Process. Lett. (2024)
19. Huang, Y., Juefei-Xu, F., Guo, Q., Liu, Y., Pu, G.: FakeLocator: Robust localization of gan-based face manipulations. IEEE Trans. Inf. Forensics Secur. **17**, 2657–2672 (2022)
20. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10770–10780). (2024)
21. Flack, T.R., Ritchie, K.L., Cartledge, C., Fuller, E.A., Kramer, R.S.: Improving face morph detection with the pairs training effect. Appl. Cogn. Psychol. **37**(6), 1158–1166 (2023)
22. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE. (2018), December
23. Bondi, L., Lameri, S., Güera, D., Bestagini, P., Delp, E.J., Tubaro, S.: Tampering detection and localization through clustering of camera-based CNN features. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1855–1864). IEEE. (2017), July
24. Mandisha, M.S., Hussien, M.A., Shalaby, A.K., Fahmy, O.M.: Wavelet Transform-based Methods for Forensic Analysis of Digital Images. J. Adv. Res. Appl. Sci. Eng. Technol. **44**(1), 46–54 (2025)
25. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16* (pp. 103–120). Springer International Publishing. (2020)
26. Atlam, E.S., Almaliki, M., Elmarhomy, G., Almars, A.M., Elsiddieg, A.M., ElAgamy, R.: SLM-DFS: A systematic literature map of deepfake spread on social media. Alexandria Eng. J. **111**, 446–455 (2025)
27. Gorle, R., Guttavelli, A.: Enhanced Image Tampering Detection using Error Level Analysis and CNN. Eng. Technol. Appl. Sci. Res. **15**(1), 19683–19689 (2025)
28. Chalini, G.R., Kanimozhi, K.V.: Evaluation Techniques to Detect Face Morphing Vulnerabilities for Differential Images. In *2024 5th International Conference on Smart Electronics and*

Communication (ICOSEC) (pp. 1532–1537). IEEE. (2024), September
29. Seibold, C., Samek, W., Hilsmann, A., Eisert, P.: Detection of face morphing attacks by deep learning. In *Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23–25, 2017, Proceedings 16* (pp. 107–120). Springer International Publishing. (2017)
30. Kadiri, P., Anusha, P., Prabhu, M., Asuncion, R., Pavan, V.S., Suman, J.V.: Morphed Picture Recognition using Machine Learning Algorithms. In *2024 Second International Conference on Advances in Information Technology (ICAIT)* (Vol. 1, pp. 1–6). IEEE. (2024), July
31. ur Rahman, R., Tomar, D.S., Das, S.: Dynamic image based captcha. In *2012 International Conference on Communication Systems and Network Technologies* (pp. 90–94). IEEE. (2012), May
32. Xhlulu: *140k Real and Fake Faces* [Dataset]. Kaggle. (2020). https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
33. Dolhansky, B., Palmer, J., Howes, R., Wang, M., Ferrer, C.C.: *Deepfake Detection Dataset (DFDD)* [Dataset]. Michigan State University CV Lab. (2019). https://cvlab.cse.msu.edu/dffd-dataset.html
34. Rahman, R.U., Kumar, P., Mohan, A., Aziz, R.M., Tomar, D.S.: A Novel Technique for Image Captioning Based on Hierarchical Clustering and Deep Learning. SN Comput. Sci. **6**(4), 360 (2025)