International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-3, May 2022
https://doi.org/10.55524/ijircst.2022.10.3.7
Article ID IRP1261, Pages 29-39
www.ijircst.org

Digital Forensics Triage Classification Model using Hybrid Learning Approaches

Mohmed Afridhi L[1], and Palanivel K[2]

[1]M. Tech Scholar, Department of Computer Science, Pondicherry University, Puducherry, India
[2] Systems Analyst, Department of Computer Science, Pondicherry University, Puducherry, India

Correspondence should be addressed to Mohmed Afridi; md.afridhi@gmail.com

ABSTRACT- The Internet and the accessibility of gadgets with connectivity have resulted in the global spread of cyber threats and cybercrime, posing significant hurdles for digital forensics. Consequently, the volume of information that may need to be investigated is growing, necessitating the development of new forensic technologies and methods. Those now in use are, in fact, old-fashioned, as they are more focused on complete device extraction for case-relevant device identification. A practical approach, a Digital Forensics Triage, tries to quickly collect facts and give essential insight into this circumstance, which could be described as data-rich but information-poor. In time-sensitive scenarios, digital forensics triage approaches can prioritize some electronic gadgets over others based on their significance to the criminal case. The Digital Forensic Laboratories (DFS) make it easier to identify essential gadgets in criminal proceedings when time, significant accumulations, and the accused's confidentiality are critical considerations. Consequently, digital forensics and machine learning techniques allow for the rapid classification of appropriate gadgets despite dipping the quantity of information that has to be adequately studied. This study presents a digital forensic model that may be utilized to build a robotic digital device categorization tool employed in real-world criminal investigations.

KEYWORDS- Cyber security, cyber threats, cybercrimes, digital forensics, digital triage, multimedia forensics.

I. INTRODUCTION

Mobile infrastructures and multimedia devices are unique and predicted to continue in the following years in today's world. Intelligent mobile devices, such as mobile phones and tablets with Wi-Fi/3G connectivity, are widely used worldwide, and this trend is projected to continue shortly. Criminals and terrorists exploit cell phones and smartphones as their practical tools, as they benefit from increased operational flexibility and adaptability due to near-instantaneous connections. Consequently, smartphones and tablets are more likely than desktop computers or laptops on today's crime scene, making them valuable sources of insight and proof in several crimes.

The shreds of evidence, such as images/photos, audio, video recordings, etc., gathered from the smart devices are crucial in today's criminal investigations. As a result, multimedia forensics [19] has become increasingly important to find evidence in multimedia data. The related terms of multimedia forensics are content-aware gadget categorization, fake image detection, video streaming classification, and source camera authentication. Consequently, the advancement of technology is likely to have a detrimental impact on criminal investigations due to the wide range of devices, operating systems, data formats, ubiquitous encryption, and cloud computing that make law enforcement's job more difficult.

The present digital forensics methodologies and tools for data investigation [24] cause an interruption in time-critical experiments [37]. They are unsuitable for growing storage density and processing vast amounts of data at Digital Forensic Laboratories [15], and they may potentially violate legal constraints on search and seizure [19]. When it refers to action like a criminal offense or assault, death, abduction, missing person, etc., a rapid detection, investigation, and presentation of accessible multimedia content, particularly on the spot, is regarded as a critical task for the entire study [28]. Furthermore, the amount of multimedia data required for DFL analysis is rapidly expanding due to the increasing storage densities. The current forensic approaches do not have the resources to keep up with that speed. The seized devices requiring forensic analysis are not unusual for DFLs to have a backlog of several months. To complicate things worse, legal restrictions on unreasonable investigation and seizure of evidence, and the necessity to safeguard the criminal's identity, render evidence acquired outside the limits of the warrant issued inadmissible [19] as it is in the Republic of Korea.

A. Manual Detection

In forensic evidence, deepfake detection approaches rely on image and video modification identification. They are detailed in the image identity verification best practices handbook. Because the forensic investigator should carefully uncover signs of tampering, those approaches are generally rigid and momentarily. Morphing and image fabrication are the most critical aspects of human deepfake detection. Using creative techniques, the generation of virtual faces involves training with a real faces database. It establishes whether a clip or image is a deepfake. The forensic experts must examine the image structure to see if any characteristics indicate whether the image includes modification artifacts.

Examples of image structure analyses are video file type, noise detection, and image data evaluation. Furthermore, the findings of these approaches should be utilized carefully because they can be faked or fabricated using commercial software. Artifact attributes and physical aspects of image content may be examined. Besides these issues with focus,

depth of field, sharpness, blurring, perspective, noise, and lens distortion should all be considered. The investigator must analyze the chain of evidence and determine enough time to create a deepfake. It is possible to modify it between the time the video was recorded and when it is seen on a desktop.

B. Automated Detection

Most deepfake detection systems, such as Face-Forensics++, depend on Deep Neural Networks (DNN) to locate images and videos. DNN utilizes a limited amount of data to enhance its detection of deepfakes. Furthermore, a big issue concerning deepfake detection is that once a detection technique for any deepfake generation method is revealed to the public, it is easy for a designer to evade the approach. A growing number of crimes are being committed via digital technologies. The digital gap, the complexities of the evidence, and the judiciary's ability to interpret it all demand fresh thinking. Consequently, deepfakes will likely become an essential aspect of modern digitalization, criminology, and crime investigation.

Prospective forensics experts may find evidence of a deepfake on gadgets such as desktops and laptops or smartphones utilizing specialist software. Neither computers nor humans should detect the similarity between deepfake and the actual image, footage, or audio sample if it has been carefully manipulated. Rapid data abstraction has been established for various objectives, such as selecting appropriate hard drives in fraudulent activity based on precise credit card numbers and recovered e-mail accounts. Despite selective statistics extraction and correlating functions, identifying digital indication as significant to a criminal probe is still primarily a manual procedure that relies heavily on personal understanding. More subsequently, research has concentrated on the necessity for automation of such a process by applying criminal patterns [13], ML supervised classification [16], and clustering algorithms, illustrating inspiration from the early Digital Triage model [9].

C. Motivations and Objectives

According to digital forensics frameworks [24], the seizing of a whole hard disc drive or the memory of a smartphone, followed by the development of a forensic picture, seems to have developed the best training embraced by law enforcement authorities to maintain proofs consistency. As a result, DFLs receive sized items and captured photos for analysis. DFLs are then entrusted with reconstructing the incident timeframe and extracting proof of the criminal's guilt using time-consuming and technologically complex processes. The outlined process eventually causes delays in investigating and prosecuting crimes, increasing bottleneck, and even breaching the confidentiality of suspects due to the intrinsic structure of the potentially incriminating evidence that may be uncovered, which typically reflects only a minor part of the confiscated goods.

This article highlights such digital forensics challenges in light of this. It investigates these issues by establishing an intelligent device's reputation in a criminal process without in-depth forensic examination or expert human effort. The purpose is to develop a digital forensic paradigm that optimizes suspicious investigation categorization.

The recommended digital forensic model involves examining and correlating multimedia data from obtained intelligent devices in various criminal cases. This concept attempted to search and investigate devices at DFLs while respecting suspects' privacy, a challenge that had hitherto gone unaddressed. This model can be customized to identify machines suspected of being used for child pornography distribution or copyright laws. It may be expanded to accommodate an unlimited number of offense types. The suggested model is based on a dataset collected from real-life criminal proceedings and uses machine learning (ML) or neural networks (NN) methods. Consequently, it should deliver better consistent and repeatable results with genuine and observable accuracies.

This article confers the subsequent research problems: The research questions addressed in this article are:

- How can the device categorization models assist the investigator in determining if a gadget is relevant to a criminal investigation?
- Is the model based on machine learning (ML) or neural networks (NN) algorithms?
- How can the categorization of mobile phones affect the efficacy of criminal investigations?

This article offers a digital forensic model for selective pre-examining and categorizing digital devices, which takes inspiration from Digital Triage to automate data recovery in criminal investigation and even from DFLs by utilizing ML-supervised classification approaches. When time is of the core, the digital forensic model strives to offer detectives timely, business insights on the criminal investigation, reducing massive bottlenecks at DFLs and protecting suspect identity where the legal system prohibits random unreasonable searches and seizures.

D. Contribution

The proposed digital triage investigation model is designed explicitly for deepfake detection tasks. The significant components of the proposed model are data extraction, data processing, data classification (or classifier), and data presentation. It reviews recent digital triage models for using specific offense categories for automated device classification, relying on different learning algorithms.

The rest of the article is structured as follows: Chapter 2 reviews the literature and introduces the paper's background information in digital forensics, digital triage, and machine learning. The digital triage model for device classification, the primary focus of this paper, is presented in chapter 3. Final observations and further discussion are discussed in chapter 4, with this study's limits and future efforts in Chapter 5.

II. LITERATURE REVIEW & RELATED WORKS

A comprehensive literature review on digital evidence, multimedia forensic work, media triage, and ML follows modern concepts and technologies. They provide a theoretical background for the proposed method. An evaluation system must be employed to establish the feasibility and efficacy of the proposed model and review and discuss the outcomes. This section contains a literature review on digital evidence, media forensics, digital triage, and learning algorithms.

A. Deepfakes – An Introduction

Deepfakes (*also known as face-swapping*) technologies are made using methodologies that incorporate a person's actual face images onto a source person's video to create the target person's video and ensure the source person's doing. Face swap is a form of deepfake that falls into this category. Deepfakes are artificial intelligence-generated videos classified as either lip-sync or puppet-master. Lip-sync deepfakes have altered their lips' expressions to match an audio recording. The target person's videos (puppet), which are animated after the person's attitudes and the motions of someone else (master) seated before a camera, are used in puppet-master deepfakes [48].

Traditional visual effects or computer-graphics approaches can create deepfakes. In addition, Deep learning (DL) approaches like Autoencoders, and Generative Adversarial Networks (GANs) have also been frequently used in computer vision subsequently [30]. These models are used to survey facial expressions and activities of a person and synthesize facial images of another person making equivalent expressions and movements.

To train methods to make photo-realistic images and movies, deepfake processes often take up the multiple-image and video data. They're utilized in pornographic photographs and films to replace public figures' appearances with bodies. They can have imaginative effects on images, interactive media, virtual reality, feature films, recreation, illustration, original video translating of foreign movies, education through to the resurrection of historical characters, virtual checking on garments when retail, and so on [1]. DeepNude has already shown enough alarming risks, as it can alter a person into non-consensual pornographic [46]. Consequently, discovering the digital domain's reality has become increasingly important. Dealing with deepfakes is considerably more difficult because they are utilized for nefarious intentions. A conflict between positive and negative uses of deep learning algorithms has erupted due to the several approaches to discovering deepfakes [41].

B. Deepfake Creation

Deepfakes Apps are generally created using deep learning algorithms [22]. Autoencoders are deep networks commonly used for dimensionality reduction and image compression. Various works, including DeepFaceLab, DFaker, and DeepFake_tf, use a deepfake generation process.

C. Deepfake Detection

Deepfake detection is typically a classifier, with algorithms distinguishing between genuine and altered videos. A vast repository of actual and false videos is required to develop classification techniques. *Pavel Korshunov* [36] created a well-known deepfake dataset built on the GAN model using the open-source Faceswap-GAN. Deepfake detection methods are divided into fake image and video detection techniques. The latter subgroups include visual artifacts inside real-time video frame-based approaches and temporal features across frames-based procedures. Though most spatial and motion techniques use DL recurrent classification models, the systems have used visual artifacts within video frames, which shallow classifiers may implement.

D. Fake Image Detection

Deepfakes are harmful to confidentiality, safety, and governance [42]. As long as this vulnerability was identified, deepfakes approaches were developed. Traditionally, faked image creation algorithms relied on customized characteristics and discrepancies. Current methods, such as [32, 20], have frequently used DL methods to retrieve prominent and discriminatory aspects to detect deepfakes automatically.

- Handcrafted Features-based Method. Deepfakes that use GAN-generated images ignore the detecting methods' generalization capability. Xinsheng Xuan [5] used a pre-processing images stage to provide the investigative classifier with basic and essential properties, allowing it to generalize further than earlier image forensics methods. To distinguish changed face images from the actual, Ying Zhang [4] employed the bag of words approach to produce a collection of brief features, which he had already put into classifiers like SVM, random forest (RF), and multilayer perceptrons (MLP). The GAN-based deepfake detection approach [44] is a hypothesis testing task, offering a statistical model for proper authentication research.
- Deep Features-based Method. Face swapping has a lot of practices in video mixing and image transformation, notably for ensuring protection. It can replace faces in pictures with stock images from a database. Because deep learning (such as CNN and GAN) can keep the stance, body language, and brightness of the images, changed face images have become far more problematic for forensics systems [21]. Tianchen Zhao [38] has presented a deepfake detection technique focused on local source characteristics and content-independent, spatially-local image information.

E. Fake Video Detection

Due to the significant loss of frame data following video compression [6]. Videos include temporal properties that vary between frames, making it difficult for algorithms to identify solely still fraudulent images to detect them. Deepfake, video detection techniques, are divided into temporal information and visual artifacts among frames.

- Temporal Features across Video Frames. Because video editing is done frame by frame, low-level anomalies generated by facial expression alterations have been hypothesized to manifest as temporal aberrations involving inconsistencies among frames. Ekraam Sabir [10] employed spatiotemporal characteristics of streaming video to detect deepfakes.
- Visual Artifacts within Video Frame. The other method divides videos into frames and searches for distinguishing features in visual artifacts. These characteristics are bundled into either a deep or shallow classifier to differentiate between fraudulent and authentic films. This section separated the techniques into two groups - shallow and deep classifiers.
- Deep classifiers. To mimic the original setting, deepfake videos are often generated with poor quality, demanding the employment of an inverse face warping algorithm. Due to the quality disparity between the warped face region and local context, based approaches produce artifacts. This demonstrates the capsule network's ability to construct a universal detection technique capable of detecting a wide range of counterfeit pictures and video assaults.

- Shallow classifiers. Most deepfake detection algorithms seek artifacts or inconsistencies among fake and real images or videos. Blockchain technology and intelligent contracts [18] could be used to identify deepfake videos since videos are only genuine when their sources can be traced. This technology is multipurpose and can be used with a variety of digital content types, such as photographs, audio, and periodicals.

Matthew Groha [29] analyzed human and artificial performance across video footage and the usefulness of initially disclosed randomized disturbances on deepfake identification to discover the strengths and weaknesses of humans and computers as deepfake detectors. It suggested looking at the machine vision surroundings of digital downloads to detect quality. The recommended method verifies whether the movie has been modified by collecting the difference rate in neighboring frames' computer vision layers. In the literature, Thanh Thi Nguyena [39] examined the techniques used to generate deepfakes and the approaches predicted to detect deepfakes. They conducted a detailed examination of deepfake practices, allowing the novel's progress, and added reliable methods for dealing with more creative deepfakes.s

F. Digital Forensics

Digital forensics [14] is an essential technique for solving cybercrimes, such as hacking and financial crimes, and crimes against humans when proof may be found on digital devices, such as money laundering and youth mistreatment. As a result, digital forensics may be primarily concerned with obtaining digital fraud evidential data to comprehend an occasion better and obtain consistent evidence submission in court. In this case, digital forensic methodologies can be used to evaluate suspect systems, capture and protect the information, and recreate events and timelines in criminal probes. According to the NIST definition [24], digital forensics, known initially as computer and network forensics, maybe apparently the "application of science to the collection, examination, analysis, and reporting of data while conserving the reliability of the data and keeping a strict chain of custody for the data."

Apart from criminal instances, forensic information might be used for many purposes, including acquiring evidence for judicial process and internal disciplinary measures, managing against malware and viruses, and other strange technological activities. The recommended 4-phases paradigm [24] for investigative work, based on the capture, inspection, examination, and presentation of data shown in Fig.1, has become the de-facto standard.

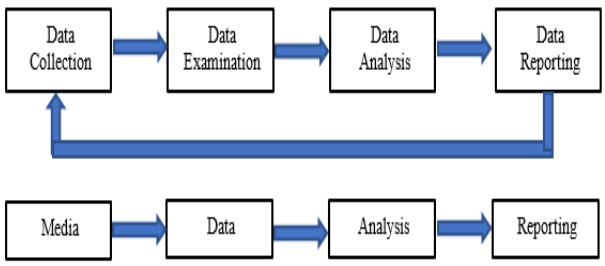

Figure 1: Forensic Model (Kent 2006)

- Collection. Information about a particular crime event is identified, labeled, captured, and gathered during collection, and its integrity is maintained.
- Investigation. During the investigation phase, investigative procedures and technologies specific to the information collected are used to find and retrieve relevant material without retaining its authenticity. In an appraisal, automated technologies and manual processes may be employed.
- Analysis. The analysis step entails analyzing the test results to obtain valuable information that provides an explanation that prompted the collecting and assessment in the first place.
- Reporting. Reporting analytical results, including summarising the measures taken, identifying what more tasks need to be done, and making suggestions for better rules, standards, processes, tools, and other parts of the forensic process is part of the final step.

As shown in Fig.1, if the evidence is required for enforcement agencies or internal use, the forensic procedure turns material into evidence. When data is reviewed, the conversion takes place, which removes information from media and converts it into a style that forensic tools can understand. Through analysis, the obtained data is converted into knowledge. When applying the ideas provided by the investigation in even more strategies during the reporting phase, the data conversion into witness is equivalent to converting information into action. These behaviors might be used as proof and insight to assist halt, minimizing, or producing new opportunities.

G. Multimedia Forensics

Digital Forensics is classified into Multimedia, network, database, and cloud. Examples of multimedia forensics are image fabrication detection, image and video content classification, content-aware devices categorization, and massive data analysis. Digital image forgery detection is about digital image changes. It includes copy-move forgery, retouching, filtering, partial deletion of objects, mounting and image splicing, luminance, colors or contrast manipulation, and geometry manipulation. Digital image tampering detection [33] approaches can be divided into several levels based on their complexities.

The data must be evaluated in an individual incident frequently equates to many terabytes due to the massive availability of high-tech gadgets and cheap storage capabilities. The content categorization and content-aware gadget classification are evolving as possible solutions to such substantial data loads. They are intended to reduce bottlenecks at DFLs and slowdowns in time-sensitive inquiries.

- The digital content categorization is primarily concerned with retrieving offense-relevant data (for example, nudity) from metadata (for example, video footage and image) using specific procedures (for example, skin detection) and detecting the devices on location correspondingly primarily in child pornography criminal matters.
- The content-aware devices categorization of the digital triage is based on ML algorithms and emphasizes detecting devices related to illegal cases. Such relation can be evaluated statistically using a database of multimedia material that has already been collected and

identified as related or unrelated to the crime under consideration.
- When it comes to image and video media analysis, handling content classification may be significant. In terms of visual information integration, efforts to locate inappropriate videos included identifying areas of skin in an image using geometric grouping criteria or the assumption of a pixel skin likelihood' using its color. In the case of video media, Jansohn [23] developed a novel approach to detecting and recognizing video databases with sexually explicit content that incorporates analysis of movement and systematic detection methods, and key-frame research.

Rogers [43] presented reliable proof of content-aware data from applications and Digital Triage using the Computer Forensics Field Triage Process Model (CFFTPM). Digital triage is linked to the categorization of content-aware systems [19] and bulk data analysis [14]. Luisa Verdoliva [26] designed an analysis approach to examine visual media integrity confirmation of manipulated images and videos. The investigation would highlight the restrictions of present forensic tools, the utmost applicable problems, and future challenges.

H. Digital Triage Approach

Digital triage has grown considerably in recent years, and it is currently rapidly growing. Digital triage aimed to remedy the delayed digital investigation methods, which are ineffective in time- criminal intelligence proceedings also as robbery or destruction to human beings. The term "triage" examines prioritizing electronic devices based on their significance in an inquiry into digital gadgets. *Digital* triage is a prioritizing procedure that aims to direct the search and seizure process in a criminal investigation and decrease the amount of data or objects that must be thoroughly reviewed at DFLs [40]. In investigations where time was significant, the CFFTPM model [43] was concerned about gaining actionable insights into the crime scene. It comprises integrity protection activities to ensure that any key evidence discovered may be examined and analyzed in a controlled environment back at the lab.

Despite the difficulties of gathering actionable insights and evidence collected on the ground, the digital triage forensics approach [37] facilitates the collection of helpful insight and additional evidence in terroristic attacks or post-blast inquiries. The growth of digital triage techniques emerged as laboratories failed to keep up with the growing inventory of digital devices [15]. The digital triage approaches are possible in time-critical circumstances, appropriate recognition of particular instance gadgets discovered at the crime scene, investigation of datasets and reduced DFLs backlog, and safeguards the criminal's confidentiality. This can be considering the search limitations from the entire device to only those areas of interest that are most likely to provide evidence.

- **Digital Triage in Mobile Forensics.** Mobile forensics has emerged as a viable solution to the need for forensically sound procedures with handheld phones. The NIST defined mobile forensics [2] as "the science of extracting digital evidence from a mobile phone using established procedures under forensically sound conditions." It is in charge of collecting, extracting, detecting, preserving, and preserving proof from mobile devices that can be used in court - for example, open-source mobile forensic systems [35]. Developing, testing, and releasing forensic tools is difficult, making selecting the best mobile forensics tool complicated [2].
- **Digital Triage in Computer Forensics.** Computer forensics employs deep investigation techniques to capture and organize information submitted as evidence from a particular computing device. It aims to perform a systematic analysis and preserve a documented chain of evidence to discover what happened on a computing device and who was responsible. Digital forensics [17] is a big data issue that many professionals are working on. Digital forensics is a combination of ML and statistical methodologies. ML systems have been suggested for analyzing and quarantining network traffic dumps or large forensic disc images.

ML-based digital triage[16] could be used to identify digital gadgets and construct an appropriate priority-based strategy using the gathered data. ML-based procedure [13] could be employed to assign files recovered from computers that many users may use to solve the multiuser carved data ascription challenge. Gomez [15] noted the necessity to train classification algorithms using offense-related attributes, comparable to searching devices using an offense-specific template. Hong [19] studied implementing a complete digital triage technique that complies with the Korean legal system's need to protect a victim's confidentiality throughout a criminal probe. The triage and quick consumption of digital media is made possible by a carving and extraction, and classification tool [14]. It scans the entire medium from beginning to end without seeking the disc head, identifying and extracting prominent elements from raw data useful in conventional digital evidence.

I. Machine Learning Approach

Machine Learning (ML) [12] is described as "a variety of strategies for automatically detecting patterns in data and then using those patterns to forecast future data or make other types of uncertain decisions." ML can be divided into descriptive, diagnostic, predictive, or prescriptive.

- Descriptive analytics uses statistics to work out what happened in history. It assists a company's understanding of its to provide clients with visibility into data. This can be accomplished via data visualizations such as graphs, charts, reports, and dashboards.
- Diagnostic analytics extends further qualitative information to provide a more in-depth examination of the goal. Root cause analysis is what it's called. This includes data discovery, mining, drilling down, and drilling beyond.
- Predictive analytics feeds past data into a machine learning model that considers prominent trends. Utilizing current information, the model is then utilized to predict what will happen next.
- Prescriptive analytics is a sort of analytics that goes beyond forecasting data. It considers various possibilities and weighs the pros and disadvantages of each.

Data set is used in descriptive and diagnostic analytics to describe what occurred and why. Predictive and prescriptive analytics use historical information to estimate future events as well as determine what measures might be taken to influence those results. Supervised and unsupervised

classification procedures [12] are the basic techniques for providing classed output.

- Learning from a training dataset of correctly identified drawings or training events based on past evaluations is known as supervised learning. This is achieved by examining behavior and actions and correlating it to historical data to identify patterns and structures that may be rationalized and also used to make choices. Data must be prepared to represent a measured' offense-related bit of knowledge kept on intelligent devices and structured to the characteristics of supervised learning.
- Unsupervised learning networks are built with unstructured samples and then free rein over the data. They extract concealed trends and correlations from the data provided. It's comparable to human brain training while constantly learning. Descriptive learning issues aim to classify information into separate categories, with objects in one class substantially more similar to others. Categorizing stats information on its fundamental resemblance is known as data clustering (or cluster analysis). Consequently, the clustering algorithm assembles dataset groups to detect cases with appropriate attributes.

ML helps users make predictions and develop algorithms that can automatically learn by using historical data. However, various ML algorithms such as Linear Regression, Logistic Regression, SVM, Decision Tree, Naïve Bayes, K-Means, random forest, Gradient Boosting, etc., require a massive amount of storage that becomes pretty challenging for a data scientist as well as ML professionals. On the criminal investigation or at DFLs, Fabio Marturana [11] described digital device pre-examination and categorization practice. In criminal proceedings, the goal has been to automate specific equipment collection. The criminal's anonymity is thought to be a severe problem. The innovative method defined herein offers numerous advantages over existing practices due to merging digital forensics practices and ml categorization.
Today, Cloud computing has become a game-changer for deploying ML models. The intelligent cloud that combines ML and cloud computing helps to enhance and expand ML applications.

J. Neural Network Approach

The neural network is an ML model (specifically, Deep Learning) utilized for unsupervised learning. It is a network of connected nodes that works the same way as neurons in the human brain. Data goes through numerous layers of interconnections in a neural network. While sending the findings on with other nodes in succeeding layers, each node determines the traits and data from the preceding layer. The neural network model is prepared up of numerous layers. The input layer is the first of three layers, accompanied by output and a hidden layer.

DL-powered technologies can generate bogus images and videos that are difficult to distinguish from real ones. To identify deepfake videos, Deep hybrid neural network architecture [47] reduced mistakes and decreased the discrepancy between fake and actual images in deepfakes. David Guera [7] presented a temporal-aware pipeline for detecting multiple deepfake videos using a convolutional neural network (CNN) and then used a recurrent neural network (RNN) to recognize whether a video has been manipulated or not. Samuel Henrique Silva [45] proposed a hierarchical explainable forensics algorithm that incorporates humans in the detection loop to share an explicable decision for forensic analyses.

K. Challenges

With the above review, it is proposed to design a model that resolves generalizability, interpretability, vulnerabilities, and restrictions issues.

- *Generalizability*. All recognized face modifications or particular datasets have been used to test present manipulation detection techniques. As a result, published diagnostic accuracy is too optimistic.
- *Interpretability.* Since deep learning is a black box, most detection methods are difficult to explain. Fake detectors currently offer a fakeness probabilistic model to a face sample, occasionally providing detection assurance.
- *Vulnerabilities.* Deep learning-based detection techniques can be significantly affected by disruptive assaults. Real-world fake detectors can deal with a variety of deprivations such as video/image loss, compaction, and so on, and they can be sensitive to malicious instances with undetectable incremental errors.
- *Restrictions*. Consistent frameworks and technologies for tampering with production and identification approaches may be generated, evaluated, configured, and compared with the assistance of such standardized frameworks.

III. METHODOLOGY

Because Deepfake is a contentious technique with numerous societal implications, significant effort has gone into creating detecting strategies to mitigate the possible negative impact of deepfakes. This study focused on content-aware device categorization in criminal proceeding studies using digital evidence. As a result, it is suggested that a model be developed to overcome problems that can be evaluated with replication. The NIST's forensic examination approach [34] was based on device collection, analysis, and disclosure, as depicted in Fig.2. The model includes extracting attributes (or features), feature processing, and device categorization, summarized as data gathering, processing, and displaying outcomes.

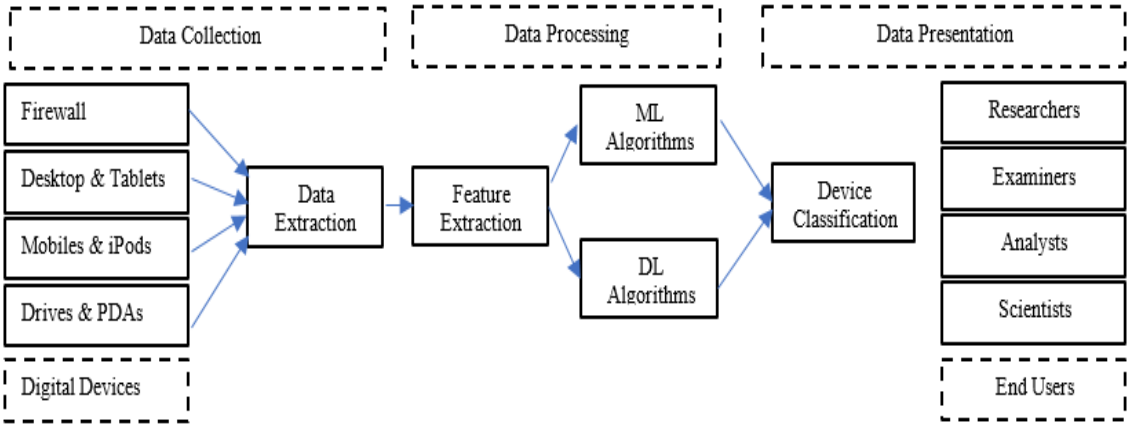

Figure 2: Digital Triage Model

Both neural networks (NN) and DL can be considered the deepfake detection technologies for *classification tasks* to detect deepfake videos automatically [34]. It is focused on the techniques of KNN and artificial neural networks (ANN)/ NN to handle classification challenges. NN resulted in greater classification accuracy when compared to KNN. The learning method used to train KNN and NN is binary exemplar categorization (i.e., pertinent or not pertinent) followed by model validation.

IV. RESULTS & DISCUSSIONS

This section describes the model for device identification on the crime scene and at DFLs. As the growth of connected gadgets provided for DFL inspection grows, congestion is forming, hindering probes, and severely influencing public protection and law enforcement agencies. Delays in obtaining data from mobile devices can have a severe influence on military and civilian safety and the mission statement of the military. Further robust on-scene triage techniques and technologies [31] are needed to provide forensic experts with relevant data and limit the number of gadgets reported to DFLs.

A. Data Corpus

Because there is not enough empirical research, the data corpus must be repeatable and drawn from real-world criminal proceedings for comparison purposes in digital forensics.

- Smartphones, storage devices, personal computers, portable music, media players, and similar devices are permissible.
- The end-users or stakeholders are scientists, investigative agencies, and digital forensics experts, and the stakeholders are researchers, crime investigators, and computer forensic examiners.
- The crime data shall be the past alteration date of essential files, investigation settings, and setup files, browse for hacking tools, investigate the password, lookup for keywords relevant to the event, browse for hidden secrets (cache), variations to documents (completely new files and file removals), retrieve a list of all e-mail, e-mail accounts and frequented and cached URLs, and so on data.
- It must highlight the issues of involvement, consumer and device identifiers, date/time, speech, and perhaps other environments such as address book, schedule, SMS origins, dialed, inbound, and images, voice, and video footage. It must also include dropped phone records and e-mail archives, surveillance video, online messaging Apps and Web surfing actions, e-documents, GPS data, and other personal computers and cell phone settings.

With the above data, a corresponding training set (i.e., number of observations) is generated and input to selected learners (i.e., training set classification). Several public datasets are available for deepfake detection: FFHQ, 100K-Faces, DFFD, CASIA-WebFace, VGGFace2, the eye-blinking dataset, and DeepfakeTIMIT.

B. The Digital Triage Model

When a bottleneck of uncategorized gadgets requires categorization as connected to a criminal investigation, the process described above is applicable from starting to DFLs inspection. Traditionally used data gathering and analysis methodologies can take hours to days to complete, except for the early training set gathering, categorization, and verification, which can be finished in a couple of moments. With the above training sets, the digital triage can identify previously undiscovered data and determine whether a digital gadget is appropriate for a criminal instance. Digital triage is practiced at DFLs with various digital gadgets assessed to the current criminal case. The proposed digital triage model is based on device retrieval, investigation, analysis, and presentation, and it can be used in criminal investigations, seizure of property, and DFL investigations.

- When a targeted gadget is discovered at a crime scene, it could be seized immediately.
- The gadget is then evaluated and sent to triage to narrow the focus of the smartphone investigation.
- The gadget could be forwarded to DFLs with further investigation, demonstration, and forensic inspection.

Working Principle. When the user uploads images and video from the front-end, the back-end calls equivalent detection methods for the submitted video. However, various detection algorithms are created using distinct programming paradigms and are dependent on varied environment parameters. As a result, a single architecture [25] may incorporate the most widely used deepfake detection techniques.

- *Step 1*. Offense-relevant information is recovered from internal storage (or forensic picture) by applying appropriate tools and following sound methods. The discovered devices (i.e., test items) are forwarded to DFLs and verified after a thorough analysis. The test item is captured in a forensic image and video.
- *Step 2*. The data retrieved from the specimen is transformed into functionalities, and a test case is generated.
- *Step 3*. The test sample is given to forensic experts, who are then categorized. The test sample is then classified as relevant or irrelevant to the study and treated appropriately.
- *Step 4*. The categorized data was sent to either the investigators on the incident site, who might proceed with the asset collection or the DFLs detective. They can determine whether to dig more into the confiscated device's analysis.

The workflow of the triage model illustrated in Fig. 3 encompassed data collection, processing, classification, and presentation.

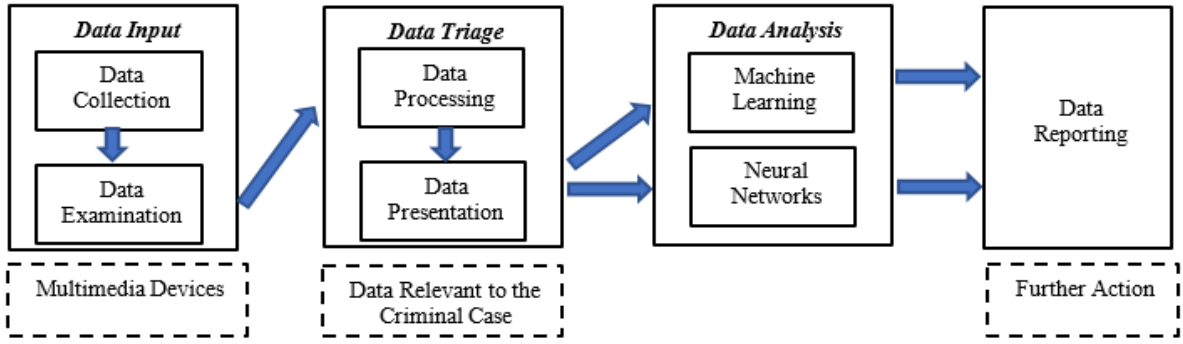

Figure 3: Digital Triage Workflow

Data Collection Plane. The data collecting plane entails locating, documenting, and gathering crucial digital data sources while adhering to data integrity norms and processes. Losing dynamic data can occur through battery-powered types of equipment (e.g., smartphones, laptops, tablets, and PDAs). According to the fragile nature of digital evidence, skilled employees involved in digital evidence recovery capture crucial data on the criminal investigation by applying specialized technology to protect evidence consistency and reliability (e.g., custody transmission) and ensure analysis uniformity.

Upon this criminal investigation or at DFLs, possibly offense intended data is collected from a desktop computer or smartphone. Once the data of particular interest, such as file information, call records or metadata, has been evaluated and retrieved from the examined devices, it is sent to the data processing stage to be turned into appropriate characteristics and instances. This was accomplished by questioning forensic investigation experts whose knowledge was precious to ensure a thorough data gathering. As a result, they could determine the information they thought was most essential in individual instances.

Data Processing Plane. Extraction of features, dataset construction, and instance processing are all included. Consequently, when specific instance information is retrieved from intelligent devices related to a criminal process is identified, it is then standardized into features (i.e., extraction of features from metadata) to build a training occurrence or training set (e.g., one per gadget). After gathering all training examples, the appropriate dataset (i.e., dataset creation) is established, from which categorization can be performed.

Features are discrete variables with values derived from countable groups that can be finite or infinite depending on whether a test or a variety of alternatives is used. A test on a specific characteristic, for example. One example is several possibilities. Counting the number of picture files and video footage taken from a device is an example of assessing the frequency of a given feature. In the approach discussed in this article, numerical features take values from infinite countable sets, whereas notional aspects get quantities from finite, countable sets.

Training exemplars are categorized, including a regression model termed class, which shows a measure of digital device importance in a criminal proceeding, once the dataset is created (i.e., instance processing). Consequently, the user estimates the result with learning accuracy as a minimal binary value, i.e., pertinent or not pertinent. One frequent solution to the issue of imbalanced data is sampling, which involves pre-processing exemplars to reduce class disparity.

Data Classification Plane. In the stage of processing, ML and DL were employed to solve the particular tasks. The steps in data management are to create a training set, evaluate the learning accuracy, create test data, categorize the test, and analyze the efficiency.

- *Create Training Set*. It creates a visual representation of how categorization is the initial stage in the supervised learning process. Examples are selected based on their relevance to the crime being investigated. The training input set is generated by aggregating all relevant exemplars and feeding them to the classifiers of preference.
- *Classify the Training Set*. Supervised learning techniques are used to analyze the training data set. Multiclass classification and binary classification are two possible classification strategies. In multiclass categorization, each occurrence is identified as the more frequent offending group in the training set. Scenarios include child pornography distribution, intellectual property theft, phishing, killing, and terrorist acts in the data corpus. In binary classification, classifiers are trained for each offense category. Whether the digital technology corresponds to the goal violation, the class is provided a

binary value (i.e., pertinent or not pertinent) in binary classification. The binary classification was used as the classification method to prevent conflicts if the training data was incomplete.

- *Evaluate the Learning Accuracy*. The model is validated based on the performance indicators such as *confusion matrix*, *precision*, *recall*, *f-measure*, and *learning accuracy*.
- *Create The Test Set*. Once the training step is over, and learning accuracy has been computed, a test instance is created. The *test input matrix* is populated as a collection of all the available test instances.
- *Classify the Test Set and Evaluate the Performance*. Selected learners are given the test input matrix, and test examples are categorized correspondingly. The performance above measures is used to calculate testing accuracy.

Data Presentation Plane. It may include summarising the activities taken, detailing how procedures and technologies were chosen, identifying what subsequent steps must be taken, and making suggestions for better rules, regulations, approaches, technologies, and other areas of the forensic examination. Based on the circumstances, the results that will be presented may differ substantially. New data sets will be examined, detected weaknesses will be secured, and improved current security mechanisms will be.

The overall architecture of the digital triage model is shown in Fig. 4 and encompasses the data collection (extraction) plane, data processing plane, data classification plane, and data presentation plane.

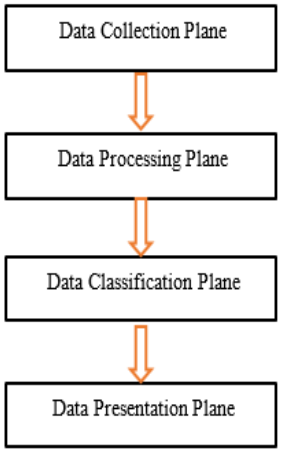

Figure 4: The Proposed Digital Triage Architecture Model

As a result, a proposed digital triage model considers the existing workforce, device accessibility, technical limits, and time limitations when conducting an on-site investigation and seizures. Bulk data analysis is excellent for triage since it retrieves information that otherwise would have been missing by current technologies. It can instantly identify whether hard drives or cell phones have relevant e-mail contacts and generate new leads.

Model Validation. Apart from facts, assessment is an essential feature of any research method that verifies its efficacy and functionality. Assessment and verification are required for the suggested technique to sufficiently identify new data once learned. This should help increase legal confidence in the system by assigning a monetary value to its correctness. The optimal strategy to analyze a classifier is to gather a lot of data to base the recommendation. Furthermore, this is unrealistic because acquiring data can be complex [12]. As previously stated, Digital Triage researchers confront challenges in obtaining real-world inquiry data (e.g., legal limitations, secrecy protection, etc.).

Limitations. Digital triage research is still in its initial stages, with much more studies ongoing. Challenges such as a shortage of court information and empirical studies to confirm the methodological approach should be considered, leading to research limitations that must be reviewed in a case-by-case scenario. Furthermore, without initially examining the data of a single file, it is difficult to deduce a device's relation to the case for specific offenses, such as fraud. The fundamental investigation methods used to retrieve critical data from the criminal's gadget and transform it into the model's attributes must encompass a variety of devices, operating systems, file types, and other aspects throughout time, which is a significant difficulty in digital forensics.

C. Case Study

A case study on mobile phone classification in court cases of child pornography exchange as an application of the model with some experimental results [27].

V. CONCLUSION AND FUTURE WORKS

This article discusses the new concept of Digital Triage in Digital Forensics, which involves categorizing digital device studies relevant to a criminal proceeding. The machine learning-based approach implies the importance of gadgets to a criminal proceeding consists of a set of offense-related exemplars and attributes, either at the criminal investigation or at DFLs. Part of a digital triage practice is raw data, extraction and classification, handling, and displaying outcomes. Appropriate training patterns are gathered to develop ML and NN strategies. After being trained, the classifiers were offered additional data examples collected from digital devices whose significance to a criminal case has yet to be determined.

The proposed model seeks to give the categorization result to either crime scene detectives or forensic examiners at DFLs to assist them in deciding which following actions to perform. The development of a multi-cloud approach for Digital Triage is considered for future study. As forensic procedures are currently unable to follow up with extensive data, this can be intended to increase horizontally to cope with rising storage density.

People's belief in media messages has been weakened by deepfakes and seeing them no longer correlates to trusting them. They have the potential to cause pain and injury to those targeted, raise misinformation and hate speech, and intensify public opinion, terrorism, or conflict. This is crucial because deepfake methods have become more widely available, and social media platforms may quickly spread fraudulent material. This review examines the issues, possible developments, and future deepfake production and identification possibilities. As a result, the artificial intelligence research field will benefit from this study in creating successful countermeasures to deepfakes.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Yisroel Mirsky, Wenke Lee (2021). The Creation and Detection of Deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1): 1–41, 2021.

[2] Ayers R, Jansen W, et al. (2007). Cellphone Forensic Tools: An overview and Analysis Update. Technical Report, National Institute for Standard and Technology (NIST).

[3] Cantrell G, Dampier D A, et al. (2012). Research toward a partially-automated and crime-specific digital triage process model. Computer and Information Science, 5(2), 29–38.

[4] Ying Zhang, Lili Zheng, Vrizlynn LL Thing (2017). Automated face swapping and its detection. The 2nd Int. Conf. on Signal and Image Processing, 15–19. IEEE, 2017.

[5] Xinsheng Xuan, Bo Peng, Wei Wang, Jing Dong (2019). On the generalization of GAN image forensics. In Chinese Conference on Biometric Recognition, 134–141. Springer, 2019.

[6] Darius Afchar, Vincent Nozick, et al. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. Int. Workshop on Information Forensics and Security, 1–7. IEEE.

[7] David Guera Edward J. Delp (2018). Deepfake Video Detection Using Recurrent Neural Networks, Video and Image Processing Laboratory (VIPER), Purdue University.

[8] David W Stewart (2021). Forensic Engineering Analysis of a Commercial Dry Storage Marina Reinforced Concrete Runway Slab, National Academy of Forensic Engineers, 38(1), 141-151.

[9] Drezewski R, Sepielak J, Filipkowski W. (2012). System Supporting Money Laundering Detection. Digital Investigation, 9, 8–21.

[10] Ekraam Sabir, Jiaxin Cheng, et al. (2019). Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 3(1):80–87, 2019.

[11] Fabio Marturana (2014). Device Classification in Digital Forensics Triage, Ph.D. Thesis, Universita Degli` Studi Di Roma Tor Vergata Dipartimento Di Ingegneria Civile E Ingegneria Informatica.

[12] Witten I H., Frank E, and M. A. (2011). Data Mining Practical Machine Learning Tools and Techniques. Elsevier, 3rd Hall Ed.

[13] Garfinkel S L, Farrell P, Roussev V, Dinolt G. (2009). Bringing science to digital forensics with standardized forensic corpora. Digital Investigation, 6(1), S2–S11.

[14] Garfinkel S L (2013). Digital Media Triage with Bulk Analysis and Bulk Extractor, Computers & Security, 32, 56–72.

[15] Gomez L S M. (2012). Triage in-lab: Case Backlog Reduction with Forensic Digital Profiling. In Simposio Argentino de Informtica y Derecho.

[16] Grillo A, Lentini A, Me G, et al. (2009). Fast user classifying to establish forensic analysis priorities. In 5th IEEE Int., Conf., on IT Security Incident Management & IT Forensics.

[17] Guarino A (2013). "Digital forensics as a Big Data Challenge," in ISSE 2013 securing electronic business processes: Springer, 197-203.

[18] Haya R Hasan and Khaled Salah (2019). I am combating Deepfake Videos using Blockchain and Smart Contracts. IEEE Access, 7:41596– 41606, 2019.

[19] Hong I, Yu H, Lee S. and Lee K. (2013). A new triage model conforms to the need for selective search and seizure of electronic evidence. Digital Investigation, 10(2), 175–192.

[20] Irene Amerini and Roberto Caldelli (2020). Exploiting Prediction Error Inconsistencies through LSTM-based Classifiers to Detect Deepfake Videos. Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security, 97– 102, 2020.

[21] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis (2017). Fast face-swap using Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision, 3677–3685, 2017.

[22] Jan Chorowski, Ron J Weiss, Samy Bengio, Aaron Van Den¨ Oord (2019). Unsupervised Speech Representation Learning using WaveNet Autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12):2041–2053.

[23] Jansohn C, Ulges A, Breuel T M. (2009). Detecting Pornographic Video Content by Combining image features with motion information. In the 17th ACM Int., Conf., on Multimedia.

[24] Kent K, Chevalier S, Grance T, Dang H. (2006). Guide to Integrating Forensic Techniques into Incident Response. Technical Report, National Institute for Standard & Technology (NIST).

[25] Khormali A, Yuan J S (2022). DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer. Appl. Sci. 2022, 12, 2953. https://doi.org/10.3390/app12062953

[26] Luisa Verdoliva (2020). Media Forensics and Deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing, 14(5): 910–932, 2020.

[27] Marturana F, Bert, R., Me G, Tacconi S. (2011b). A Quantitative Approach to Triaging in Mobile Forensics. In IEEE International Joint Conference of TrustCom-11/ICESS-11/FCST-11.

[28] Marturana F. Tacconi S. (2013). A Machine Learning-Based Triage Methodology for Automated Categorization of Digital Media. Digital Investigation, 10(2), 193–204.

[29] Matthew Groha, Ziv Epsteina, Chaz Firestoneb, Rosalind Picard (2021). Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds, PNAS 2022 119(1), e2110013119, 1-11. DOI:10.1073/pnas.2110013119

[30] Ming-Yu Liu, Xun Huang, et al. (2021). Generative Adversarial Networks for Image and Video Synthesis: Algorithms and applications. Proceedings of the IEEE, 109(5):839–862, 2021.

[31] Mislan R P, Casey E, and Kessler G C. (2010). The growing need for on-scene triage of mobile devices. Digital Investigation, 6(3-4), 112–124.

[32] Oscar de Lima, Sean Franklin, et al. (2020). Deepfake Detection using Spatiotemporal Convolutional Networks. arXiv preprint arXiv:2006.14749, 2020.

[33] Wang W, Dong J, and Tan T. (2009). A survey of passive image tampering detection. In 8th International Workshop on Digital Watermarking.

[34] Pan D, Sun L, Wang R, Zhang X, and Sinnott R. O. (2020). "Deepfake Detection through Deep Learning," 2020 IEEE/ACM Intel., Conf., on Big Data Computing, Applications and Technologies (BDCAT), 134-143, DOI: 10.1109/BDCAT50828.2020.00001.

[35] Walls R J, Learned-Miller E, and Levine B N. (2011). Forensic triage for mobile phones with dec0de. In the 20th USENIX Conference on Security.

[36] Pavel Korshunov, Sebastien Marcel (2019). Vulnerability Assessment and Detection of Deepfake Videos. In 2019 International Conference on Biometrics (ICB), 1–6. IEEE.

[37] Pearson S. and Watson R. (2010). Digital Triage Forensics-Processing the Digital Crime Scene. Syngress.

[38] Tianchen Zhao, Xiang Xu, et al. (2021). Learning self-consistency for deepfake detection. In Proc. of IEEE/CVF Int., Con., on Computer Vision, 15023–15033, 2021.

[39] Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, et al., (2022). Deep Learning for Deepfakes Creation and Detection: A Survey, arXiv:1909.11573v4.

[40] Pollitt M. M. (2013). Triage: A Practical Solution or Admission of Failure. Digital Investigation, 10(2), 87–88.

[41] Siwei Lyu (2020). Deepfake Detection: Current Challenges and Next Steps. In IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 1–6. IEEE, 2020.

[42] Robert Chesney, Danielle Keats Citron (2018). Deep fakes: A looming challenge for privacy, democracy, and national security. Democracy, and National Security, 107, 2018.
[43] Rogers M. K, Goldman J, Mislan R, and Wedge T. (2006). Computer forensics field triage process model. In Conference on Digital Forensics, Security and Law.
[44] Sakshi Agarwal and Lav R Varshney (2019). Limits of deepfake detection: A robust estimation viewpoint. arXiv preprint arXiv:1905.03493, 2019.
[45] Samuel Henrique Silva, Mazal Bethany, et al. (2021). Deepfake Forensics Analysis: An Explainable Hierarchical Ensemble of Weakly Supervised Models, Forensic Science International: Synergy 4 (2022) 100217.
[46] Samuel S (2019). A Guy Made a Deepfake App to Turn Photos of Women into Nudes. it didn't Go Well. https://www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nudewomen-porn
[47] Santosh Kolagati, Thenuga Priyadharshini V, Mary Anita Rajam (2022). Exposing Deepfakes using a Deep Multilayer Perceptron – Convolutional Neural Network Model, I Information Management Data Insights, 2(1), 100054. DOI: 10.1016/j.jjimei.2021.100054.
[48] Shruti Agarwal, Hany Farid, et al. (2019). Protecting World Leaders Against Deep Fakes. Computer Vision and Pattern Recognition Workshops, 1, 38–45, 2019.

ABOUT THE AUTHORS

L Mohmed Afridi received his Bachelor of Technology in Computer Science & Engineering from JNTU Anantapur and pursuing a Master of Technology in Network and Information Security from Pondicherry University, Puducherry, India. Presently he is working as a Cyber Security Analyst. His field of interest is Cyber Forensics, Cyber Security, and computer networks. He has more than two years of work experience as a Python Developer (Data Science) and research in Cyber Forensics at Pondicherry University.

Palanivel Kuppusamy received his Bachelor of Engineering in Computer Science & Engineering and Master of Technology in Computer Science & Engineering from Bharathiar University, Coimbatore, India, and Pondicherry University, Puducherry, India, in 1994 and1998 respectively. Presently he is working as a Systems Analyst in the Computer Centre of Pondicherry University, India. His field of interest is Software architecture, emerging technology, computer networks, and Design Patterns, and he has published more than 50+ research publications. He has more than 20 years of experience in computer applications, teaching, and research at Pondicherry University.