



# A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact

Mohammad Wazid<sup>a,\*</sup>, Amit Kumar Mishra<sup>b</sup>, Noor Mohd<sup>a</sup>, Ashok Kumar Das<sup>c</sup>

<sup>a</sup> Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248 002, India

<sup>b</sup> Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun 248 002, India

<sup>c</sup> Center for Security, Theory and Algorithmic Research, International Institute of Information Technology, Hyderabad 500 032, India

## ARTICLE INFO

### Keywords:

Deepfake  
Artificial Intelligence (AI)  
Machine learning  
Cyber security  
Authentication

## ABSTRACT

Deepfake refers to synthetic media generated through artificial intelligence (AI) techniques. It involves creating or altering video, audio, or images to make them appear as though they depict something or someone else. Deepfake technology advances just like the mechanisms that are used to detect them. There's an ongoing cat-and-mouse game between creators of deepfakes and those developing detection methods. As the technology that underpins deepfakes continues to improve, we are obligated to confront the repercussions that it will have on society. The introduction of educational initiatives, regulatory frameworks, technical solutions, and ethical concerns are all potential avenues via which this matter can be addressed. Multiple approaches need to be combined to identify deepfakes effectively. Detecting deepfakes can be challenging due to their increasingly sophisticated nature, but several methods and techniques are being developed to identify them. Mitigating the negative impact of deepfakes involves a combination of technological advancements, awareness, and policy measures. In this paper, we propose a secure deepfake mitigation framework. We have also provided a security analysis of the proposed framework via the Scyther tool-based formal security verification. It proves that the proposed framework is secure against various cyber attacks. We also discuss the societal impact of deepfake events along with its detection process. Then some AI models, which are used for creating and detecting the deepfake events, are highlighted. Ultimately, we provide the practical implementation of the proposed framework to observe its functioning in a real-world scenario.

## 1. Introduction

Deepfake refers to synthetic media generated through artificial intelligence (AI) techniques, particularly deep learning [1]. It involves creating or altering video, audio, or images to make them appear as though they depict something or someone else. Deepfake technology uses mechanisms like neural networks and generative adversarial networks (GANs) [2]. It then manipulates the existing content. It has the capability to be used in different contexts, like the manipulation of an individual's voice or his/ her facial expressions. It can do replacement of faces in movies and create artificial surroundings that appear really genuine [3]. Though it is used for creative and entertaining activities, at the same time, it introduces some concerns due to its ability to be misused in different ways. It includes the production of false information and the spreading of fabricated information [4]. Further, there may be manipulation of content due to malicious intent, i.e., impersonation.

Due to these serious issues, some robust efforts should be made to detect deepfake events [5–7].

### 1.1. Issues, challenges, and threats with deepfake

Some strong efforts are required for the proper detection of deepfake events. There should be some increment in the awareness drives related to deepfake as it has a potential impact on various societal factors. Legislative measures and regulations should be implemented to overcome the ethical and legal problems introduced by deepfake. Some of the realistic deepfake cases are, celebrities (Bollywood and Hollywood) fake photos and video creations and spreading them all around the world. Some of them are related to sextortion cases. Even it has also impacted the famous politicians of the world. For example, the creation of fake videos of two politicians from different countries and showing them like

Peer review under responsibility of KeAi Communications Co., Ltd.

\* Corresponding author.

E-mail addresses: [wazidkec2005@gmail.com](mailto:wazidkec2005@gmail.com) (M. Wazid), [amitmishraddun@gmail.com](mailto:amitmishraddun@gmail.com) (A.K. Mishra), [noormohdcs@gmail.com](mailto:noormohdcs@gmail.com) (N. Mohd), [iitkgp.akdas@gmail.com](mailto:iitkgp.akdas@gmail.com), [ashok.das@iiit.ac.in](mailto:ashok.das@iiit.ac.in) (A.K. Das).

<https://doi.org/10.1016/j.csa.2024.100040>

Received 29 December 2023; Received in revised form 21 January 2024; Accepted 2 February 2024

Available online 5 February 2024

2772-9184/© 2024 The Authors. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

real-life couples. The details of issues and threats associated with deepfake are provided below [1,5–7].

- **Dissemination of incorrect information:** The application of deepfakes enables the creation of convincingly realistic but entirely fabricated information. Improper utilization of this can potentially engender confusion or erode trust in media outlets and sources of information, ultimately culminating in the dissemination of inaccurate information, the manipulation of public sentiment, and the proliferation of fabricated news stories. All of these objectives can be achieved through the improper utilization of this tool [1].
- **Privacy issues:** The utilization of deepfake technology enables the production of explicit content without the need for authorization, achieved by overlaying human faces onto explicit films. This objective is achieved by leveraging technology to generate the content. Unauthorized use of individuals' identities poses a serious risk of harm, thereby raising significant privacy concerns [5,6].
- **Impersonation issues:** When utilizing this technology, it is possible to create compelling imitations of well-known personalities or individuals from one's circle. The technology facilitates both of these options. As a result, there is a potential risk of identity theft, financial fraud, or harm to an individual's reputation [6].
- **Society and decline in faith:** With the advancing sophistication of deepfakes, discerning between authentic and manipulated content is becoming progressively challenging. It is conceivable that the erosion of trust in the media and online content may have significant societal implications [6,7].
- **Legal problems:** There are several significant legal and ethical difficulties that are associated with the use of deepfake technology. Some of these challenges include concerns with consent, intellectual property, and the responsibilities of platforms to control or delete problematic deepfake content [2].

### 1.2. Societal impact of deepfake

As the technology that underpins deepfakes continues to improve, we are obligated to confront the repercussions that it will have on society. The introduction of educational initiatives, regulatory frameworks, technical solutions, and ethical concerns are all potential avenues via which this matter can be addressed [6,7]. As a result of its ubiquitous application, deepfakes have a substantial impact on society in a variety of diverse areas. Examples of their effect that are particularly noteworthy can be found in the various key areas that are described below.

- **False, misleading, and misleading news:** The utilization of deepfakes to create deceptive narratives poses a challenge in distinguishing between authentic information and manipulated content, leading to a blurred line between truth and falsehood. The utilization of deep fakes enables the creation of fabricated narratives. The use of deepfake technology helps to build narratives, which may be misleading. The possibility of the creation of inaccurate information or fabricated news going to increase day by day. Further, it leads to a decline in the trust factor associated with the big media houses and the public figures [1].
- **Politics and an influenced voting procedure:** The introduction of deepfakes technology causes a significant problem. As it can modify comments, speeches, or events involving influential politicians. Due to this, it introduces a potential threat to the political system of a democracy. The distribution of those fabricated materials produces a great influence on public sentiments. It may cause social disturbances and can affect election results [5,6].
- **Privacy and the theft of identities problems:** The use of deepfake technology can do victimization of an individual. It results in their identity theft, infringements their privacy, and also causes detrimental effects on their reputations, i.e., misuse of their fabricated photographs and videos [6,7].

- **Encouraging financial fraud:** Deepfake technology can also be used to perform financial fraud. One method, which can be used for the making of fake/ fabricated videos/ audio files. It has the purpose of deceiving individual users into making money transactions, or there may be the disclosing of sensitive data [5,6].
- **Problems in law enforcement:** There are chances that advancement of deepfake technology may present strong challenges for law enforcement agencies. This is due to the potential for individuals to create counterfeit evidence or deceptive material. In addition, this technology has the potential to be utilized to impersonate individuals, thereby posing significant security risks for numerous businesses. This situation has the potential to be hazardous [2].
- **Impact on one's mental state:** Continuous exposure to deepfakes can erode trust, leading individuals to question the legitimacy of visual and aural content, which in turn leads to an increase in skepticism and confusion. There is a correlation between deepfakes and the deterioration of trust [5–7].

## 2. Research motivation

As the technology underpinning deepfakes continues to improve, we are obligated to confront the repercussions it will have on society. The introduction of educational initiatives, regulatory frameworks, technical solutions, and ethical concerns are all potential avenues for addressing this matter. As a result of its ubiquitous application, deepfakes have a substantial impact on society in a variety of diverse areas. There should be some increment in the awareness drives related to deepfake as it has a potential impact on various societal factors. Legislative measures and regulations should be implemented to overcome the ethical and legal problems introduced by deepfake. From this discussion, it is clear that a secure deepfake mitigation framework is necessarily required. Therefore, we work on a secure deepfake mitigation framework in the paper.

### 2.1. Research contributions

The research contributions of the paper are given below.

- We propose a secure deepfake mitigation framework using Elliptic Curve Cryptography (ECC)-based signature and encryption schemes.
- The architecture of deepfake mitigation framework through general guidelines is provided. Further, the cryptographic design of deepfake mitigation framework is given.
- We have also provided a security analysis of the proposed framework via the Scyther tool-based formal security verification. It proves that the proposed framework is secure against various cyber attacks.
- We also discuss the societal impact of deepfake events along with its detection process.
- Further, some of the AI models used for creating and detecting deepfake events are highlighted.
- We further provide the practical implementation of the proposed framework to observe its functioning in a real-world scenario.

### 2.2. Article organization

The remainder of the paper is organized as follows. In Section 3, we provide the details of the existing similar schemes. The detection methodologies of deepfake events' detection are given in Section 4. The details of various artificial intelligence (AI) models used in creating and detecting deepfakes are given in Section 5. The details of the proposed deepfake mitigation framework are given in Section 6. Further, the formal security verification of the proposed framework is given in Section 7. Furthermore, the practical implementation of the proposed framework is provided in Section 8. Finally, the work is concluded in Section 9 with some concluding remarks and future research works.

**Table 1**  
A critical analysis of the related existing frameworks and the proposed framework

| Scheme                     | Summary                                                                                                                                                                                                                                                                                                                                     | Limitations/ advantages                                                                                                |
|----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| Qureshi <i>et al.</i> [6]  | They provided a proof-of-concept deepfake detection system. Their system could identify fake news video clips made using voice impersonation. It involved the incorporation of digital watermarks into the audio track of a video by a hybrid speech watermarking technique.                                                                | Security analysis of the scheme was not provided.                                                                      |
| Li <i>et al.</i> [7]       | They provided an unsupervised learning-based detection method for recognizing deepfake films. The primary principle underpinning their solution was the facial region in the genuine video.                                                                                                                                                 | Security analysis of the scheme was not provided.                                                                      |
| Fehring and Bonaci [9]     | They discussed the societal implications of deepfakes. The technology for deepfakes, such as autoencoders and generative adversarial networks (GANs), continues progressing and becoming more accessible. Due to that, deepfakes are becoming easier to construct and more convincing.                                                      | Proper design of the deepfake mitigation mechanism was not provided. Security analysis of the scheme was not provided. |
| Mangaokar and Prakash [10] | They explored five common fallacies regarding deepfakes. They adopted a security perspective to propose three recommendations to minimize backward-facing research efforts.                                                                                                                                                                 | Security analysis of the scheme was not provided.                                                                      |
| Nagothu <i>et al.</i> [11] | They introduced “DeFakePro.” It was based on a decentralized consensus mechanism for online videos. Utilizing “electrical network frequency (ENF),” an environmental fingerprint incorporated in digital media recording. That made it possible to create a consensus process that was referred to as the “proof-of-ENF (PoENF) algorithm.” | Security analysis of the scheme was not provided.                                                                      |
| Proposed framework         | A secure deepfake mitigation framework using Elliptic Curve Cryptography (ECC)-based signature and encryption schemes                                                                                                                                                                                                                       | Proper security analysis, cryptographic design, and practical implementation have been provided.                       |

### 3. Related Work

In this section, we summarize some of the works related to deepfake. The details are given below.

To completely capture the semantic content of the text, Guo *et al.* [3] proposed a strategy for identifying fake news that would apply to the setting of mixed languages and would be implemented using a multiscale transformer. It was anticipated that by extracting more fruitful feature levels from the initial textual contents, it would be possible to create more robust feature spaces for the semantic properties of mixed languages.

Sudhakar and Shanthi [4] provided a comprehensive literature review of deepfake detection schemes. Their study highlighted the numerous methods that were utilized in the development and detection of deepfakes. Rana *et al.* [5] have undertaken a literature review to provide an updated overview of the research efforts in deepfake detection. They summarized some related works which were published between 2018 and 2020. Patil and Chouragade [8] examined recent research publications to identify deep phony videos shared on social media.

Qureshi *et al.* [6] provided a proof-of-concept deepfake detection system. Their system could identify fake news video clips made using voice impersonation. Their proposed method involved the incorporation of digital watermarks into the audio track of a video by utilizing a “hybrid speech watermarking methodology.” Li *et al.* [7] provided an unsupervised learning-based detection method for recognizing deepfake films. This method was initially developed in the publication. The primary principle underpinning their solution was that the facial region in the genuine video was captured by the camera, whilst its counterpart in the deepfake video was often generated by the computer. Fehring and Bonaci [9] discussed about the societal implications of deepfakes. They discussed that technology for deepfakes, such as autoencoders and generative adversarial networks (GANs), continues progressing and becoming more accessible. Due to that, deepfakes are becoming easier to construct and more convincing. Further, Mangaokar and Prakash [10] explored five common fallacies regarding deepfakes. Then, they adopted a security perspective to propose three recommendations designed to minimize research efforts that were backward-facing. These efforts were useful for the detection of deepfakes.

Nagothu *et al.* [11] introduce “DeFakePro,” a deepfake detection technology based on a decentralized consensus mechanism. It was utilized in online video conferencing tools. Utilizing “electrical network frequency (ENF),” an environmental fingerprint that was incorporated in digital media recording, made it possible to create a

consensus process that was referred to as the “proof-of-ENF (PoENF) algorithm.”

A critical analysis of the related existing schemes is given in Table 1.

### 4. Detection of deepfake

In this section, we provide the details of detecting deepfake problems. The deepfake technology advances just like, the mechanisms, which are used to detect them [1]. There’s an ongoing cat-and-mouse game between creators of deepfakes and those developing detection methods. Hence, multiple approaches need to be combined to identify deepfakes effectively [5,6]. Detecting deepfakes can be challenging due to their increasingly sophisticated nature, but several methods and techniques are being developed to identify them.

- **Identification of inconsistencies:** Deepfakes may exhibit small anomalies, such as lip motions that do not match, facial expressions that are not genuine, or weird artifacts around the face. Similar to the last example, audio deepfakes may contain glitches, pauses that are not natural, or inconsistencies in the voice, all of which can be indications that the voice has been manipulated. A further point to consider is that the metadata associated with the media file could occasionally reveal abnormalities or evidence of modification [2].
- **Genuineness of source:** It may be helpful to cross-reference with a number of different sources or to confirm the legitimacy of the content with sources that are well-known and trusted when one is looking to discover inconsistencies [7].
- **Forensic analysis of content:** The content can be analyzed by sophisticated forensic techniques, which can discover any anomalies, modifications, or evidence of tampering that may be present in the pixels, frames, or audio waveforms [5,6].
- **Utilization of AI-based schemes:** Technologies that can combat deepfakes are now in the process of being developed. These technologies make use of machine learning algorithms in order to recognize patterns that are specific to content that is believed to be deepfake. These systems are trained on massive datasets of both real and fraudulent forms of media in order to acquire the knowledge necessary to differentiate between the two types of media (i.e., original or fabricated) [6,7].
- **Watermarking with blockchain:** Several sites are investigating the possibility of employing blockchain technology and digital watermarking to validate the genuineness of media assets and learn where they originated [8].

- **Application of human expertise:** When identifying irregularities that automated technologies would potentially overlook, human knowledge in assessing small visual or aural cues can be extremely beneficial [1].

## 5. Different artificial intelligence (AI) models used in creating and detecting deepfakes

In the context of the deepfake ecosystem, AI models perform a variety of functions, including the generation of synthetic media that is realistic and the detection and identification of content manipulations. The field of deepfakes is always evolving, which means that new and improved models and strategies are constantly being developed to construct deepfakes and to counteract them. Multiple AI models are utilized in the process of developing and detecting deepfakes. Detailed information is available down below [5–7].

- **Generative adversarial networks (GANs):** The use of GANs is the primary model that is applied for the generation of deepfakes. Both a generator that creates fake content (photos, movies, and so on) and a discriminator that makes an effort to differentiate between real and fake content are examples of neural networks that are utilized in the components that comprise these systems. As a result of the competitive process in which both networks learn from one another, the quality of the information that is generated by both networks gradually improves over the course of time [1].
- **Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs):** Examples of sequential data that can be stored in these are audio and video files. As a result of their ability to gain knowledge of temporal correlations and patterns, they are able to generate or modify sequential content, such as facial expressions or speech [5,6].
- **Deep convolutional neural networks (CNNs):** Considering that deep CNNs can interpret visual data, they are frequently utilized in deepfake identification. They can recognize patterns and irregularities in films and photographs that may indicate tampering [8].
- **Transformers:** Text-based deepfakes are produced by employing models such as Generative Pre-trained Transformer (GPT) and its variants. These deepfakes are developed using deep learning techniques. Not only are these models capable of producing text material that is realistic, but they also do very well in tasks that include natural language understanding [2].
- **Autoencoders:** The utilization of these models is necessary for the encoding and decoding of data. In the context of deepfakes, a variational autoencoder, also referred to as a VAE for short, is an ad-

ditional type of autoencoder applied. Using this particular kind of autoencoder, it is possible to generate new data by gaining an understanding of the representation that lies behind a dataset [1,2].

## 6. The proposed deepfake mitigation framework

Mitigating the negative impact of deepfakes involves a combination of technological advancements, awareness, and policy measures. Combating deepfakes requires a multi-faceted approach that combines technological innovation, education, policy, and responsible practices within the technology and media industries. As deepfake technology evolves, ongoing efforts to mitigate its negative impact must also adapt and advance. In this section, we provide the details of the proposed deepfake mitigation framework. It is also depicted in Fig. 1. It has been designed as follows.

### 6.1. General guidelines

The general guidelines of the deepfake mitigation framework are given below [5–8].

1. **Development of detection schemes:** Investing in research and developing more sophisticated detection tools and algorithms to identify deepfakes is crucial. This includes using AI, machine learning, deep learning, and generative AI schemes to create counter-technology that detect inconsistencies in audio, video, or image content [7].
2. **Societal awareness programs:** Educating the public about the existence and potential dangers of deepfakes is important. Teaching media literacy skills can help individuals critically evaluate information and be cautious of manipulated content. The government authorities can initiate these programs and involve some non-government organizations for better involvement of people. Under those circumstances, the results and outcomes will be good [1].
3. **Encourage verification of source:** Encouraging and promoting the verification of the source of information can help prevent the spread of deepfakes. Platforms can integrate features that indicate the authenticity or origin of content. Schemes like entity authentication, content originator authentication, and content authentication should be applied [7].
4. **Implementation of regulations and policies:** Governments and tech companies can collaborate to establish guidelines, regulations, or policies specifically addressing creating and distributing deepfake content. This could include standards for labeling or removing po-

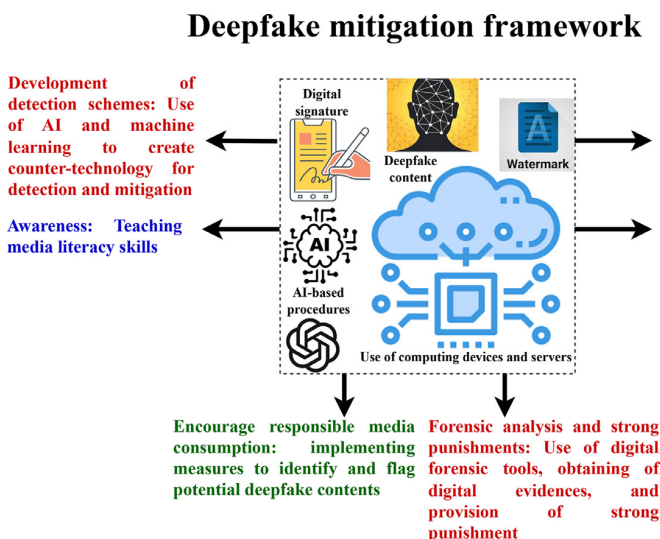


Fig. 1. Architecture of deepfake mitigation framework-general guidelines

tentially harmful content. We should also try to include the potential people of the society in creating these policies [6].

5. **Implementing digital watermarking and digital signature:** Implementing digital watermarking or digital signatures on content can help track the source and verify authenticity. The watermarks mechanisms such as visible watermarks, invisible watermarks, public watermarks, and fragile watermarks are available. Digital signature schemes, like the RSA digital signature algorithm, digital signature algorithm (DSA), Elliptic Curve Digital Signature Algorithm (ECDSA), etc., are available. Emerging potential technologies such as blockchain can also be deployed for better results and outcomes [8].
6. **Encourage responsible media consumption:** Platforms and media outlets can take responsibility for the content they host or share by implementing measures to identify and flag potential deepfakes and providing context or fact-checking information alongside suspicious content. Encouraging collaboration among researchers, tech companies, policymakers, and law enforcement agencies can lead to a more coordinated effort in tackling the issues surrounding deepfakes [1,2].
7. **Forensic analysis and strong punishments:** Advanced forensic tools can analyze the content for anomalies, alterations, or traces of manipulation in the pixels, frames, or audio waveforms. Digital forensic tools, like Wireshark, FTK imager, Volatility, EnCase, Nmap, FTK forensic toolkit, etc., are available. Effective digital evidence can be obtained through these tools for their use in courtroom proceedings. Based on that, strong punishments can be given to the malicious people who indulge in those activities [6].

## 6.2. Threat model

In the proposed framework, we have considered the guidelines of the Dolev-Yao (DY) model [12]. It states that the communicating parties communicate through the general open channel (i.e., the Internet).

This channel is insecure. Therefore, the potential adversary  $\mathcal{A}$  gets some chances to delete, read, or modify the exchanged messages. Apart from that, we have followed the guidelines of the CK-adversary model [13]. This model says that  $\mathcal{A}$  has all capabilities like the DY model. In addition to that,  $\mathcal{A}$  can steal the session states and the associated information, i.e., session keys can be revealed to  $\mathcal{A}$  if it is not handled properly [14,15].

## 6.3. Cryptographic design of deepfake mitigation framework

In this section, we provide the details of the design of the deepfake mitigation framework. We use various cryptographic parameters and algorithms in the design of this framework [16,17]. The details are given below.

1. **Registration phase:** In this phase, there is an assigned trusted registration authority  $TA$ , which does the registration of the various entities of the network. For example, the content provider  $CP_i$  and content requester  $CP_j$  are registered. Here, it is important to mention that  $CP_i$  can be deployed through the cloud servers.  $TA$  chooses a “non-singular elliptic curve over a finite field. There are two constant numbers “ $u \in Z_q$  and  $v \in Z_q$ , where  $Z_q = \{0, 1, \dots, q-1\}$  and  $q > 3$  should be a prime”. Then  $TA$  selects “a non-singular elliptic curve  $E_q(u, v): y^2 = x^3 + ux + v$  over the finite field  $GF(q)$ , such that  $4u^3 + 27v^2 \neq 0 \pmod{q}$  with a point at infinity or zero point  $\mathcal{O}$ ”. The parameter  $G$  is a base point in  $E_q(u, v)$  with a similar big order like  $q$ .  $TA$  also chooses “collision-resistant one-way hash function  $h(\cdot)$  (for example, SHA-256 hash algorithm)” to accomplish different tasks.
2. **Mutual authentication and secure session establishment:** In this phase, there will be the execution of the standard mutual authentication and session key establishment among the registered entities of the network [18,19]. To execute this task, the procedure steps available in these articles can be followed [16,17]. After the completion of the procedure, there is a session key establishment between  $CP_i$  and  $CP_j$  as  $SK_{CP_i, CP_j} (= SK_{CP_j, CP_i})$ .

```
hashfunction h; const xor:Function;
const cat:Function; const add:Function;
const mul:Function; const mod:Function;
protocol deepfake(CP, CR)
{
  role CP
  {
    const TKREQCONK, RIDCPj, REQCONK, SKCPiCRj, SKCRjCpi, kCpi, TS1, TS2,
    srcpi, q;
    macro SGREQCONK = add(kCpi, mul(h(cat(RIDCPj, TS2, h(REQCONK))),
    mod(srcpi, q)));
    macro SKCPiCRj = SKCRjCpi;
    recv_1(CR, CP, cat({TKREQCONK}SKCRjCpi, TS1));
    send_2(CP, CR, cat({REQCONK}SKCPiCRj, {SGREQCONK}SKCPiCRj, TS2));
    claim_CP1(CP, Secret, (kCpi));
    claim_CP2(CP, Secret, (srcpi));
    claim_CP3(CP, Niagree); claim_CP4(CP, Nisynch);
    claim_CP5(CP, Secret, (SKCPiCRj));
    claim_CP6(CP, Weakagree); claim_CP7(CP, Alive);
  }
  role CR
  {
    const TKREQCONK, RIDCPj, REQCONK, SKCPiCRj, SKCRjCpi, kCpi, TS1, TS2,
    srcpi, q;
    macro SKCRjCpi = SKCPiCRj;
    send_1(CR, CP, cat({TKREQCONK}SKCRjCpi, TS1));
    recv_2(CP, CR, cat({REQCONK}SKCRjCpi, {add(kCpi, mul(h(cat(RIDCPj, TS2,
    h(REQCONK))), mod(srcpi, q)))}SKCRjCpi, TS2));
    claim_CR1(CR, Niagree); claim_CR2(CR, Nisynch);
    claim_CR3(CR, Secret, (SKCRjCpi));
    claim_CR4(CR, Weakagree); claim_CR5(CR, Alive);
  }
}
```

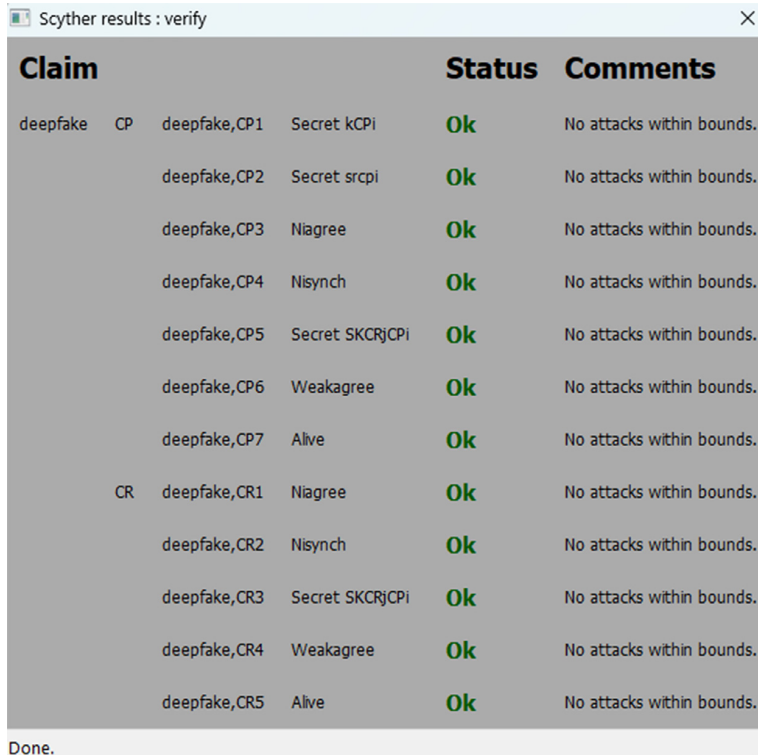
Fig. 2. SPDL snippet for the role of  $CP_i$  and  $CP_j$

**3. Secure content delivery to the requester:** In this phase, we provide the details of the secure content delivery process between  $CP_i$  and  $CP_j$ . The  $CP_i$  provides the requested content  $REQ_{conk}$  to  $CP_j$  in a secure way through the established session key  $SK_{CP_i,CP_j}$  ( $= SK_{CP_j,CP_i}$ ). It can be executed as follows. First of all,  $CR_j$  sends a content request as  $MSG_1 = \{E_{SK_{CR_j,CP_i}}(TK_{REQ_{conk}}), TS_1\}$ , where  $TK_{REQ_{conk}}$  is a identifier for requested content  $REQ_{conk}$  and  $TS_1$  is a freshly generated timestamp value. Here  $E_{x,y}()$  represents a general encryption procedure, which can be done through algorithms like Advanced Encryption Standard (AES). At the arrival of  $MSG_1$  of  $CP_i$  does the verification of  $TS_1$  through the condition  $|TS_1 - TS_1^*| \leq \Delta T$ , where  $\Delta T$  stands for the maximum transmission delay and  $TS_1^*$  for the message  $MSG_1$ 's receiving time to verify timestamp value  $TS_1$ . If the condition holds,  $CP_i$  decrypts  $D_{SK_{CR_j,CP_i}}(E_{SK_{CR_j,CP_i}}(TK_{REQ_{conk}}))$  and obtains  $TK_{REQ_{conk}}$ . Here  $D_{x,y}()$  represents a general encryption procedure, which can be done through algorithms like Advanced Encryption Standard (AES). Through  $TK_{REQ_{conk}}$ ,  $CP_i$  tries to determine which content has been requested by  $CR_j$ . Then  $CP_i$  generates another fresh timestamp value  $TS_2$  and creates a message as  $MSG_2 = \{E_{SK_{CP_i,CP_j}}(REQ_{conk}), E_{SK_{CP_i,CP_j}}(SG_{REQ_{conk}}), TS_2\}$ , where  $E_{x,y}()$  represents a general encryption procedure, which can be done through algorithms like Advanced Encryption Standard (AES).  $CP_i$  then sends the message  $MSG_2$  to  $CP_j$ . At the arrival of  $MSG_2$ ,  $CP_j$  performs the verification of timeliness of  $MSG_2$  using the condition  $|TS_2 - TS_2^*| \leq \Delta T$ , where  $\Delta T$  stands for the maximum transmission delay and  $TS_2^*$  for the message  $MSG_2$ 's receiving time to verify timestamp value  $TS_2$ . If the condition holds, then  $CP_j$  decrypts the requested digital content as  $REQ_{conk} = D_{SK_{CP_i,CP_j}}(E_{SK_{CP_i,CP_j}}(REQ_{conk}))$ , where  $D_{x,y}()$  represents a general decryption procedure, which can be done through algorithms like Advanced Encryption Standard (AES). After that,  $CP_j$  computes  $h(REQ_{conk})$  and stores it for its future use. It further decrypts the signature part and obtains  $SG_{REQ_{conk}} = D_{SK_{CP_i,CP_j}}(E_{SK_{CP_i,CP_j}}(SG_{REQ_{conk}}))$ .

**4. Digital content's genuineness checking and source verification:** In this phase, we provide the details of the digital content's genuineness checking and source verification procedure. It is executed through the procedure of Elliptic Curve Cryptography (ECC). To execute this process,  $CP_i$  generates the signature of the digital content  $REQ_{conk}$  as  $SG_{REQ_{conk}} = k_{CP_i} + h(RID_{CP_j} || TS_1 || h(REQ_{conk})) * sr_{CP_i} \pmod{q}$ , where  $Q_{CP_i} = k_{CP_i}.G$  are the public key ( $Q_{CP_i}$ ) and private key ( $k_{CP_i}$ ) of  $CP_i$  and  $SR_{CP_i} = sr_{CP_i}.G$  are the public parameter ( $SR_{CP_i}$ ) and secret number ( $sr_{CP_i}$ ) of  $CP_i$ . Here, it is important to mention that  $RID_{CP_j} = h(ID_{CP_j} || k_{TA})$  is the pseudo-identity of  $CP_j$  and  $k_{TA}$  is a private key of  $TA$ . These digital signatures are also provided to  $CP_j$  through message  $MSG_1$  in a secure way. To perform the verification of signatures,  $CP_j$  executes the procedure as  $SG_{REQ_{conk}}.G = Q_{CP_i} + h(RID_{CP_j} || TS_1 || h(REQ_{conk})).SR_{CP_i}$ . If both sides of this equation match, then  $CP_j$  assumes that the obtained digital content  $REQ_{conk}$  is original, and it also verifies that source  $CP_i$ 's genuineness to  $CP_j$ . Using this mechanism  $CP_i$  and  $CP_j$  exchange the digital content securely without any malicious middle entity's forgery.

## 7. Formal security verification of the proposed framework

In this section, we provide the formal security verification of the proposed framework. For this task, the Scyther tool is used. [20–22]. This tool is very effective compared to the other tools, like, ProVerif and AVISPA. Scyther tool has been implemented with secure cryptographic principles and guidelines. It tries to discover as many security vulnerabilities as possible in a newly designed security protocol. It uses the "Security Protocol Description Language (SPDL)" language for the coding and implementation. In the SPDL specification, every participant in a communication assumes the position of a role. Roles can carry out a variety of tasks, including the sending and receiving of messages, the management of events, and the submission of any necessary security claims [23]. One of the ten various adversarial models that Scyther conforms to is the Dolev-Yao (DY) model. Other adversarial models that it



| Claim                                | Status | Comments                  |
|--------------------------------------|--------|---------------------------|
| deepfake CP deepfake,CP1 Secret kCPI | Ok     | No attacks within bounds. |
| deepfake,CP2 Secret srcpi            | Ok     | No attacks within bounds. |
| deepfake,CP3 Niagree                 | Ok     | No attacks within bounds. |
| deepfake,CP4 Nisynch                 | Ok     | No attacks within bounds. |
| deepfake,CP5 Secret SKCRjCPI         | Ok     | No attacks within bounds. |
| deepfake,CP6 Weakagree               | Ok     | No attacks within bounds. |
| deepfake,CP7 Alive                   | Ok     | No attacks within bounds. |
| CR deepfake,CR1 Niagree              | Ok     | No attacks within bounds. |
| deepfake,CR2 Nisynch                 | Ok     | No attacks within bounds. |
| deepfake,CR3 Secret SKCRjCPI         | Ok     | No attacks within bounds. |
| deepfake,CR4 Weakagree               | Ok     | No attacks within bounds. |
| deepfake,CR5 Alive                   | Ok     | No attacks within bounds. |

Done.

Fig. 3. The scyther security analysis results

adheres to include CK and eCK. With its suite of tests, Scyther asserts that it can validate security properties such as “privacy, authentication, synchronization, aliveness, weak agreement, and agreement.” It is important to note that the digital content requester (CR) and the digital content provider (CP) both play crucial roles in the phases required by the proposed framework. Scyther tool performs execution of the SPDL file and runs multiple analyses on the proposed framework. Fig. 2 depicts the SPDL code for the roles of  $CP_i$  and  $CR_j$ , respectively. Fig. 3 depicts the output of the Scyther tool’s analysis. The analysis outcomes indicate that the proposed framework is protected under the claims and would be protected against various attacks.

## 8. Practical implementation of the proposed framework

In this section, we do the practical implementation of the proposed framework. The details are given below.

### 8.1. Simulation environment and settings

The details of the simulation parameters are given in Table 2. We used Intel(R) Core(TM) i5-8250U CPU@1.60GHz processor with Windows 11 Home platform. We have taken the Google Colab platform for

**Table 2**

Simulation parameters and their values

| Parameter                       | Value                                  |
|---------------------------------|----------------------------------------|
| Processing unit                 | Intel(R) Core(TM) i5-8250U CPU@1.60GHz |
| Platform used                   | Windows 11 Home                        |
| Environment                     | Google Colab                           |
| Programming environment         | Python 3.7                             |
| Random access memory (RAM) size | 8 GB                                   |
| Number of scenarios             | 3 (scenario-1, scenario-2, scenario-3) |
| Scenario-1                      | One $CP_i$ and ten $CP_j$              |
| Scenario-2                      | Three $CP_i$ and thirty $CP_j$         |
| Scenario-3                      | Five $CP_i$ and fifty $CP_j$           |

the implementation along with Python 3.7 programming language. The size of random access memory (RAM) was 8 GB. In the implementation, we have taken three different scenarios, i.e., scenario-1 having one  $CP_i$  and ten  $CP_j$ . Further, in scenario-2, we have taken three  $CP_i$  and thirty  $CP_j$ . Apart from that, in scenario-3, we have taken five  $CP_i$  and fifty  $CP_j$ .

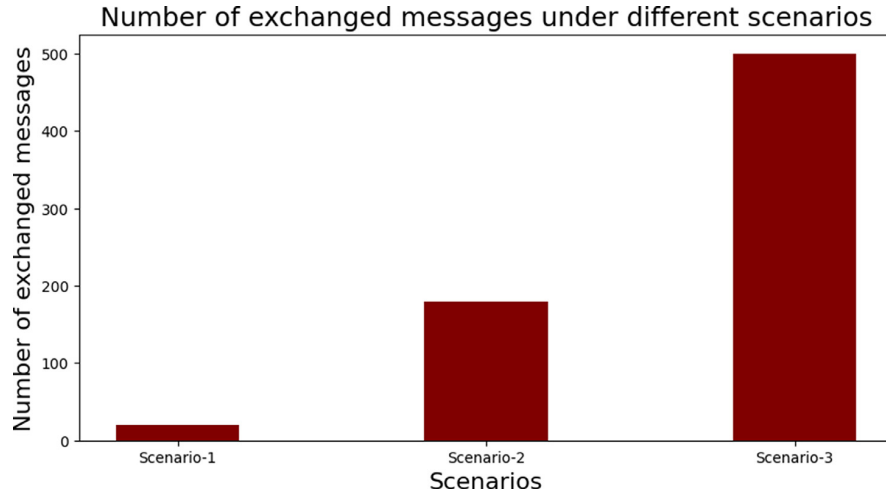


Fig. 4. Number of exchanged messages under different scenarios

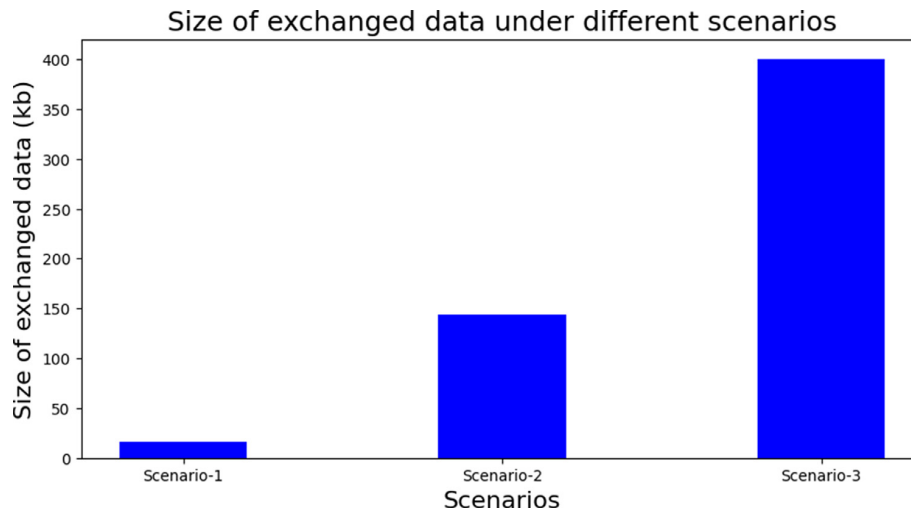


Fig. 5. Size of exchanged data (kb) under different scenarios

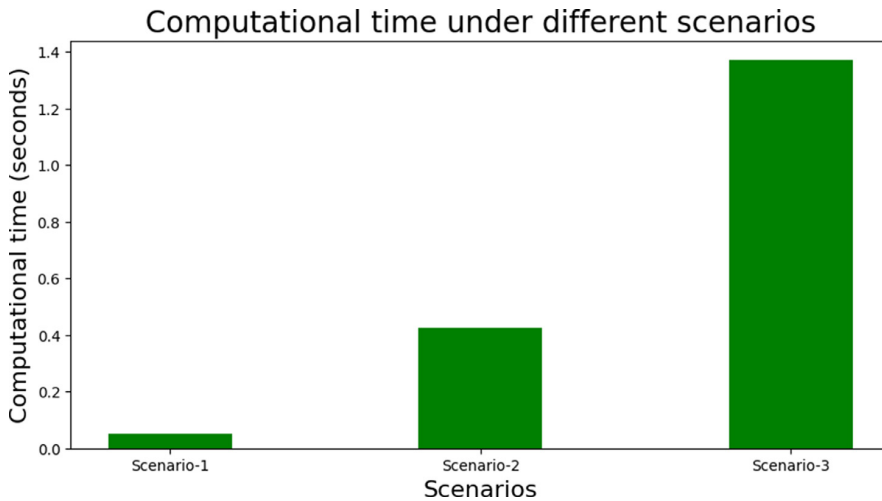


Fig. 6. Computational time (seconds) under different scenarios

## 8.2. Obtained results

In the practical implementation and analysis, we have analyzed and estimated important results, like the “number of exchanged messages,” the “size of exchanged data (kb), and the “computational time (seconds)”. Their details are provided below.

- **Number of exchanged messages:** In the considered scenarios, we have estimated 20, 180, and 500 messages in scenario-1, scenario-2 and scenario-3, respectively. Here, it is important to mention that the number of exchanged messages increases from scenario-1 to scenario-2 and scenario-2 to scenario-3. As in scenario-2 and scenario-3, we have more number of  $CP_i$ s and  $CP_j$ s that incurs the exchange of more number of messages. The results are also depicted in Fig. 4.
- **Size of exchanged data (kb):** In the considered scenarios, we have estimated 16 kb, 144 kb, and 400 kb of data in scenario-1, scenario-2 and scenario-3, respectively. Here, it is important to mention that the value of exchanged data (kb) increases from scenario-1 to scenario-2 and scenario-2 to scenario-3. As in scenario-2 and scenario-3, we have more number of  $CP_i$ s and  $CP_j$ s that incurs the exchange of more number of messages. It causes an increment in the exchanged data. The results are also depicted in Fig. 5.
- **Computational time (seconds):** In the considered scenarios, we have estimated 0.05070 seconds, 0.42586 seconds, and 1.37018 seconds of computational time's values in scenario-1, scenario-2 and scenario-3, respectively. Here, it is important to mention that the value of computational time (seconds) increases from scenario-1 to scenario-2 and scenario-2 to scenario-3. As in scenario-2 and scenario-3, we have more number of  $CP_i$ s and  $CP_j$ s that incurs the exchange of more number of messages. It requires some time. Hence, the computational time increases accordingly. The results are also depicted in Fig. 6.

## 9. Conclusion and future work

Mitigation of deepfakes becomes essential. We proposed a secure deepfake mitigation framework. We provided a security analysis of the proposed framework via the Scyhter tool-based formal security verification. It proved that the proposed framework was secure against various attacks. We also discussed the societal impact of deepfake events along with its detection process. Finally, we provided the practical implementation of the proposed framework to observe its functioning in a real-world scenario.

In the future, we would like to add more functionality features to the proposed framework. Furthermore, some machine learning-based

algorithms can also be used for the detection and mitigation of deepfake cases.

## Declaration of competing interest

Ashok Kumar Das is an editorial board member for *Cyber Security and Applications* and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

## CRediT authorship contribution statement

**Mohammad Wazid:** Writing – review & editing, Writing – original draft, Validation, Visualization, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Amit Kumar Mishra:** Resources. **Noor Mohd:** Data curation. **Ashok Kumar Das:** Validation.

## References

- [1] H. Lamb, Can you believe your eyes? 'deepfake' technology is racing towards realism. But what can we do when fakes are accepted as true and reality is dismissed as false? *Eng. Technol.* 15 (2020) 18–21.
- [2] P. Fraga-Lamas, T.M. Fernandez-Carames, Fake news, disinformation, and deepfakes: leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality, *IT Profession.* 22 (2020) 53–59.
- [3] Z. Guo, Q. Zhang, F. Ding, X. Zhu, K. Yu, A novel fake news detection model for context of mixed languages through multiscale transformer, *IEEE Trans. Comput. Soc. Syst.* (2023) 1–11, doi:10.1109/TCSS.2023.3298480.
- [4] K.N. Sudhakar, M. Shanthi, Deepfake: an endanger to cyber security, in: *IEEE International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 2023, pp. 1542–1548, doi:10.1109/ICSCSS57650.2023.10169246. Coimbatore, India
- [5] M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, Deepfake detection: a systematic literature review, *IEEE Access* 10 (2022) 25494–25513.
- [6] A. Qureshi, D. Megias, M. Kuribayashi, Detecting deepfake videos using digital watermarking, in: *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1786–1793. Tokyo, Japan
- [7] L. Zhang, T. Qiao, M. Xu, N. Zheng, S. Xie, Unsupervised learning-based framework for deepfake video detection, *IEEE Trans. Multimedia* 25 (2023) 4785–4799, doi:10.1109/TMM.2022.3182509.
- [8] U. Patil, P. Chouragade, Deepfake video authentication based on blockchain, in: *IEEE Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1110–1113, doi:10.1109/ICESC51422.2021.9532725. Coimbatore, India
- [9] J.A. Fehring, T. Bonaci, It looks like me, but it isn't me: on the societal implications of deepfakes, *IEEE Potential.* 42 (2023) 33–38.
- [10] N. Mangaokar, A. Prakash, Dispelling misconceptions and characterizing the failings of deepfake detection, *IEEE Secur. Privacy* 20 (2022) 61–67.
- [11] D. Nagothu, R. Xu, Y. Chen, E. Blasch, A. Aved, Defakepro: decentralized deepfake attacks detection using ENF authentication, *IT Profession.* 24 (2022) 46–52.
- [12] D. Dolev, A. Yao, On the security of public key protocols, *IEEE Trans. Inf. Theory* 29 (1983) 198–208.

- [13] R. Canetti, H. Krawczyk, *Universally Composable Notions of Key Exchange and Secure Channels*, Springer Berlin Heidelberg, Amsterdam, The Netherlands, 2002. L.R. Knudsen, *Advances in Cryptology — EUROCRYPT*, 337–351.
- [14] S. Ahmad, S. Mehfuz, Efficient time-oriented latency-based secure data encryption for cloud storage, *Cyber Secur. Appl.* 2 (2024) 100027.
- [15] W.S. Admass, Y.Y. Munaye, A.A. Diro, Cyber security: state of the art, challenges and future directions, *Cyber Secur. Appl.* 2 (2024) 100031.
- [16] M. Wazid, A.K. Das, V. Bhat K, A.V. Vasilakos, LAM-Clot: lightweight authentication mechanism in cloud-based iot environment, *J. Netw. Comput. Appl.* 150 (2020) 102496.
- [17] M. Wazid, A.K. Das, V. Odelu, N. Kumar, M. Conti, M. Jo, Design of secure user authenticated key management protocol for generic iot networks, *IEEE IoT J.* 5 (2018) 269–282.
- [18] Y. Yang, D. He, P. Vijayakumar, B.B. Gupta, Q. Xie, An efficient identity-based aggregate signcryption scheme with blockchain for iot-enabled maritime transportation system, *IEEE Trans. Green Commun. Netw.* 6 (2022) 1520–1531.
- [19] J. Shen, H. Yang, P. Vijayakumar, N. Kumar, A privacy-preserving and untraceable group data sharing scheme in cloud computing, *IEEE Trans. Depend. Secure Comput.* 19 (2022) 2198–2210.
- [20] B. Khadem, A.M. Suteh, M. Ahmad, A. Alkhayyat, M.S. Farash, H.S. Khalifa, An improved WBSN key-agreement protocol based on static parameters and hash functions, *IEEE Access* 9 (2021) 78463–78473.
- [21] C.J.F. Cremers, *Scyther : semantics and verification of security protocols*, <https://pure.tue.nl/ws/files/2425555/200612074.pdf>. Accessed on December 2023.
- [22] M. Tanveer, A.H. Zahid, M. Ahmad, A. Baz, H. Alhakami, Lake-iod: lightweight authenticated key exchange protocol for the internet of drone environment, *IEEE Access* 8 (2020) 155645–155659.
- [23] M. Adeli, N. Bagheri, H.R. Meimani, On the designing a secure biometric-based remote patient authentication scheme for mobile healthcare environments, *J. Ambient Intell. Humanized Comput.* 12 (2021) 3075–3089.