

ISSN: 2582-7219



## **International Journal of Multidisciplinary** Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Issue 5, May 2025

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|



International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET) (A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

## Combating Adversarial Deepfakes using TAHN: Integrating Blockchain Consensus for Enhanced Trust and Robustness

Deepak Kumar B<sup>1</sup>, Franklin Jerold<sup>2</sup>, Pramoda N P<sup>3</sup>, Mohammed Mueen Pasha M<sup>4</sup> and Dr. Sandeep J<sup>5</sup>

Student, Department of Computer Science St Joseph's University, Bengaluru, India 1,2,3

Research Scholar, Department of Computer Science, CHRIST University, Bengaluru, India<sup>4</sup>

Associate Professor, Department of Computer Science, CHRIST University, Bengaluru, India 5

**ABSTRACT**: Over the last few decades, the use of technology has witnessed a manifold increase. Although the usage of technology such as machine learning, artificial intelligence and deep learning rapidly grows in genuine and legitimate means, but there are always unscrupulous and malicious users who try to exploit these technologies. Deep fake is one such domains that is used to generate fake images and videos that seems extremely authentic and challenges users from identifying and distinguishing the fake content in contrary to the original one. Deepfake is a synthetic media created using Generative Adversarial Networks (GANs). This paper explores the rise and consequences of deepfake technology, propelled by advancements in artificial intelligence and deep learning. In order to improve accuracy and reliability we discuss about the Textural Adversarial Hybrid Network (TAHN) as a novel approach to enhance model resilience against adversarial attacks. In detection of deepfake and adversarial defences, TAHN plays a vital role as an hybrid model that exhibits superior performance over the existing textual as well as image based models deployed for detection. To further enhance the security and transparency of the detection process, we propose the integration of blockchain consensus mechanisms for decentralized validation of deepfake detection results. This ensures tamper-proof logging and collaborative verification, reinforcing trust in TAHN's decision-making pipeline. In this paper we systematically assess the state-of-the-art deepfake detection method (TAHN) and feasibility of using this novel approach as a defence mechanism.

KEYWORDS: Deepfakes, GAN, Blockchain, Consensus mechanism, adversarial attacks, deepfake detection, TAHN.

## I. INTRODUCTION

The advancement in artificial intelligence (AI) and deep learning has led to the development of useful yet threaten tool that can generate highly realistic synthetic media, referred as Deepfakes. Deepfakes are synthetic media that is generated by Artificial Intelligence (AI) which includes images, videos and sound that represents the events that has not occurred or never existed in real. Deepfakes generates photos, videos or sounds that targets individuals using methods like Generative Adversarial Networks (GANs). This method makes it difficult to identify the difference between the real content and the faked information. The word Deepfake comes from a combination of deep, emerged from the deep learning in AI and fake, labelling something that is not real. The term came into use in 2017 where the subreddit post called "deepfakes" was launched by the reddit moderator which uses face switching technology to place someone's likenesses into the existing threaten videos on the internet. Some of the high profile deepfake instances that happened over the years includes, former US President Donald Trump involves in the clash with the policemen, Pope Francis in the puffer jacket, Mark Zuckerberg, CEO of Meta admitting corruption in his speech and Queen Elizabeth's video where she is dancing and presenting the speech about the technology. These events are never actually happened in the real world. Deepfake possess serious risks in the form of misuse, misrepresentations, fraud, malicious attacks and integrity invasion of an individuals, even it has the ability to be used in various innovative and education instructional purposes. Identifying deepfakes is crucial to maintain the integrity of individual's information, especially in the sensitive areas such as politics, media, and social media. In politics, deepfakes can be used to manipulate public by spreading negative information about the opposition parties. while in media, the fake contents destroy the trust among the people about journalism. On social media platforms, where information can spread quickly, identifying the real



content is crucial by distinguishing it from the fake contents. Hence strong deepfake detection methods are essential to preserve public trust, gain confidence and to ensure that the information passed among the people is accurate and true. Robust detection mechanisms are necessary to prevent the exploitation of deepfakes, as it continues to grow and threaten public. However, the same AI technologies are used to create adversarial attacks in deepfakes – The strategy to trick machine learning algorithms to generate false predictions or classifications. An adversarial attack is an input to machine learning models that are purposely designed to make a model to mistake and generate false outcomes despite giving a valid input. For example, In image classification, an adversarial attack may add small pixel level adjustments to an image that are invisible to humans but it leads to making an model to incorrectly classify the image.

## **II. METHODOLOGY**

Deepfakes are not edited or photoshopped images or videos. They are designed with some specific mechanisms that combines existing and the new footage. A Generator and a discriminator (Neural Networks) are involved in the process of generating deepfakes. Generator achieves this by generating the synthetic data using the wanted output on the training dataset. Whereas the discriminator studies and determines the realism or artificialism of the fake data that is produced by the generator. This procedure goes into iterations where the generator later. These neural networks combine to create a Generative Adversarial Network (GAN). A GAN method finds patterns of images and creates the synthetic data by using those patterns. The deepfake picture is generated by a GAN system where it scans the image from the wide variety of sides to capture the subject's commonalities and perspectives. The GAN system views the video from multiple perspectives and analyzes the subject's characteristics to create a deepfake. These details are passed to the discriminator at multiple stages to make sure the output image or video looks real.

There are mainly two ways to create deepfake video, One is modifying the original video of target making the targeted subject stating or doing the activities that they had not done, While the second is implementation of the face interchanging mechanism in which they switch the face of the target of someone's real time video. When the audio deepfakes are generated, a GAN duplicates the voice of the targeted person, builds an AI model based on the subject's speech patterns and uses this model to generate any desired audio output. This technique is employed by the video game developers frequently. Another notable practice is lip syncing. This involves the alignment of recorded audio with the video by using this technology, where it appears as a targeted person in the video is uttering the recorded audio. If the passed audio itself a deepfake, then the video adds another layer of duplicity to the video. Recurrent neural networks facilitate this approach (The current layer's input is based on the output of the previous layer).

#### III. TECHNICAL BACKGROUND

The development of deepfakes is becoming more advanced due to the enhancement of some crucial technologies: **Generative Adversarial Network (GAN):** Deepfake contents are generated by GAN's using mechanisms namely generator and discriminator. The generator produces a fake digital content, whereas the discriminator tries to find the errors between the real and fake data in the generated content. This adversarial process paves the way to generate the fake realistic content in the today's world.

**Convolutional Neural Network (CNN):** CNN's are essential for identifying patterns in the visually represented data. These information are crucial for producing realistic deepfakes as they are widely used for facial recognition, identifying and tracking facial movements and understanding the structural hierarchies in the images.

Autoencoders: This type of neural networks is tasked to recognize and encode the relevant characteristics of a target such as body language and facial expressions. The autoencoder applies these characteristics onto the source video, enabling the mixing of various facial expressions of the target.

**Natural Language Processing (NLP):** NLP is used to create a deepfake audio. By analysing the speech patterns of a target, NLP algorithms can produce original speech that closely matches to the target's voice and speaking style. This is can be done often with a high degree of accuracy.



**High-Performance Computing (HPC):** Deepfakes require a significant computational power, particularly during the phases of training and generation. Faster and more complex deepfake models can be created due to the HPC's ability to handle the demanding processing tasks.

**Video Editing Software:** Modern video editing software integrates AI technologies into it. These technologies can enhance the overall realism of the deepfake outputs by tuning facial features, synchronizing audio and making other related editing adjustments. These technologies make deepfakes more accessible and realistic while also increasing their accuracy, ease of use and frequency.

**Blockchain and Consensus protocols:** Blockchain is a decentralized and tamper-resistant ledger technology that ensures data integrity across distributed nodes. Consensus algorithms, such as Proof of Work (PoW), Proof of Stake (PoS), or Practical Byzantine Fault Tolerance (PBFT), are used to achieve agreement among nodes on the validity of transactions or events. By integrating consensus mechanisms, blockchain can provide transparent, secure, and verifiable logging for deepfake detection results and model integrity validation. IV DEEPFAKE DETECTION TECHNIQUES

The deepfake attacks can be detected by several practices. The following are the signs that indicates possible deepfake content:

- Unusual or awkward face positions or expressions.
- Unnatural movement of the body or the face.
- Unnatural coloring in the content.
- Videos that appear strange when zoomed in or out.
- The audio is inconsistent.
- People who don't blink their eyes.
- The aging of the skin does not match with the aging of the hair and eyes.
- Despite the movements of the subject, the glasses glare angle remains constant, It can be no glare or have too much glare.

In textual deepfakes, these are some of the detection strategies:

- Misspelled words.
- Sentences that don't flow naturally.
- Suspicious email addresses as the source.
- Phrasing might be inconsistent with the sender's identity.
- Messages sent out of context are not relevant to any discussion, occasion or problems.

But AI is gradually overcoming some of these markers, such as the tools that support natural blinking and some biometric evidences.

## V. ADVERSARIAL TRAINING PROCESS

Algorithm for the training loop:

**Step 1:** The Generator produces a fake sample data using some noisy data (Usually drawn from Gaussian Distribution) as an input.

Step 2: The discriminator is fed with both the real data and fake sample data from generator as input. It identifies any data as fake or real data.

**Step 3:** The discriminator is trained as it classifies real and fake data accurately. This is accomplished by optimizing the loss function that punishes for incorrect classifications.

**Step 4:** A Generator is trained to pull discriminator towards the probability of making a mistake and calling it as a real data instead of fake. This is achieved by minimizing the loss function, where a generator gets penalized if its outputs could clearly sound as fake or fraud.



**Step 5:** The above steps are repeated for several iterations. The generator becomes better at generating realistic data with each iteration. As the discriminator gets better at identifying fakes.

## VI. EXISITING SYSTEM

## 6.1 Textural Adversarial Hybrid Network (TAHN)

Over recent years, adversarial attacks have obtained remarkable recognition in the field of machine learning, especially in deep learning models. These attacks necessitate manipulation of an input data in a way that is often unnoticeable to humans but causes the model to process incorrect results. Despite the fact that adversarial attacks are frequently connected with image data, with the expeditious progress in Natural Language Processing, there has been a growing emphasis on text-based adversarial attacks. Depending on how the network is built and put into effect, the idea of a Textural Adversarial Hybrid Network represents an engrossing way to merge textural analysis and adversarial techniques that could be applied to both text and image inputs.

#### 6.2 Definition and Components

In order to enhance the robustness of deep learning models, TAHN integrates adversarial learning technique with texture-based feature extraction. The network's dual ability to supervise adversarial perturbations, which can impact text and picture data, as well as the texture information which is essential for image processing is what sets it apart.

**6.3 Adversarial Component:** The objective of this network's element is to generate adversarial instances, that involve textural or visual modifications made sufficient to trick the model into predicting the incorrect outcome. Adversarial instances in a text environment might involve in terms that are misspelled or replaced using synonyms to mislead the model without altering the meaning altogether. This could involve tiny pixel alterations for images that are imperceptible to the human eye yet has significant effects on the model's outcome.

**6.4 Textural Component:** When dealing with images, the elements of a network's textural analysis is critical. The physical arrangements of pixels in an image provides with essential characteristics regarding the patterns, edges, and several other visual elements. In the context of text, texture could be recognized metaphorically such as the structure of text, together with syntax, grammar, and sentence complexity. The textural component focuses into capturing and making the best use of these elements in order to enhance the network's comprehension and robustness.

**6.5 Hybridization:** The hybrid point of the network is recognized with incorporating adversarial and textural characteristics into a single model. This hybrid method can be immensely productive in handling multi-model data, which includes both text and images. The network can be constructed to handle text and image inputs individually through their discrete adversarial and textural tracks. It can then integrate these insights in to make predictions that are more dependable and well-informed.

**6.6 Consensus Mechanism:** Consensus mechanisms are protocols used in blockchain to ensure that all participating nodes agree on a single version of truth without relying on a central authority. They provide fault tolerance and prevent tampering by validating transactions or data entries through collective agreement. In the context of deepfake detection, consensus can be used to securely validate and log detection outcomes, ensuring integrity and trust in the model's decisions.

#### 6.7 Compatibility with Text and Image Input

Because if its inherent versatility, the Textural Adversarial Hybrid Network can be adapted to process both text and image inputs. Although the architecture might need some modifications based on the specific use case of the application.

**6.8 Text Input:** The network's adversarial element would concentrate on producing adversarial text sample for text data which involves techniques as synonym replacement, word insertion, or paraphrasing, which can slightly alter the text while preserving its original meaning. In the context of text, the textural component could involve in examining syntactical structures, linguistic patterns, sentence complexity, all of which contributes completely towards the efficiency of the model.



**6.9 Image Input:** The textural component takes on a more conventional role when dealing with image data by involving pixel pattern analysis, edge analysis, and other visual characteristics which defines the image texture. The adversarial part would focus on generating adversarial images that include precise changes in order to fool the model. Later in the hybridization phase, these two streams of data would be integrated to enhance the capacity of a model in defending against adversarial attacks.

## **VII. APPLICATIONS AND IMPLEMENTATION**

TAHN as a wide range of potential applications spanning across various domains such as natural language processing, medical imaging, security, and in any field where adversarial robustness is a concern. For instance, TAHN could be used to improve the detection model's ability to hold up against adversarial attacks that aim to evade detection mechanism in the context of deepfake detection — a relevant area for this research. In order to enhance the robustness of machine learning models, the Textural Adversarial Hybrid Network (THAN) combines adversarial learning methods with texture-based feature extraction. Two parallel pathways of the input data are processed through the network: one that analyses textural elements like pixel patterns in image or syntactical structures in text, while the other generates and defends against adversarial examples. Later by integrating these insights, THAN is able to make more reliable and accurate predictions, effectively providing defense against adversarial attacks on various data sources. In order to improve the robustness of the model against adversarial attacks, TAHN is designed to process both image and textual data. The network uses a sequential design that combines textural and adversarial features for accurate predictions.

## 7.1 Input Data Layer

**Textual Data:** This pathway aims to capture the text's syntactic and semantical structure. The way it handles the textual input is by preprocessing the data and then converting it into a format that is suitable for embedding (e.g., contextual embeddings like BERT or word embeddings such as Word2Vec, Glove).

Adversarial Data: This pathway deals with adversarial examples, such as minor adjustments made to the text or image data. Adversarial examples may involve synonyms or misspellings for text data. On the other hand, for images, there could be a subtle pixel alteration intended to trick the model.

1.0 Layered Architecture of the proposed model





### 7.2 Concatenation Layer

The textual and adversarial pathways results are combined and added into single representation. When adversaries are used the same feature cannot be passed twice to the rest of your network which is why we introduced a concatenation layer that brings into account both normal and adversarial features so they entire interaction between them. In order to improve the network to make more informed predictions, the concatenated vector would be a rich representation by having both adversarial and semantic information.

$$C = [T_{text}, A_{adv}]$$

 $T_{text}$  = Textual feature vector,  $A_{adv}$  = Adversarial feature vector, C = Concatenated vector passed to the next layers.

#### 7.3 Dense Layer1

Now the concatenated input is passed through Dense Layer1; which is a fully connected neural network layer. The model does so by adding non-linearity using fixed set of weights and biases per input, as well in addition the activation function such as ReLU or Tanh. The main purpose of this layer is to grasp the intricate patterns and relations among adversarial elements, related textual part.

#### 7.4 Dense Layer2

The output which is received from Dense Layer1 is now passed to Dense Layer2, which proceeds refining from the features learned in the previous layer. This additional layer helps the network to learn higher-level abstractions and dependencies, making it better at handling complex scenarios, specifically in adversarial settings, while it is crucial for making robust predictions. The network's depth helps in capturing minute information which single-layer model might miss out.

$$D_i = f(W_i \cdot C + b_i)$$

 $W_i$  = Weight matrix,  $b_i$  = Bias vector, f(.)= Activation Function (ReLU, Tanh)

#### 7.5 Output Layer (Binary Output)

Finally, the Output Layer receives the processed data at this point, which is typically a binary classifier (like a sigmoid activation function). A binary output indicating the model determined if an adversarial perturbation is present. The predictions are made by the network on analysing the textual or image data is real or altered by any adversarial perturbations.

$$0 = \sigma(W_o \cdot D_i + b_o)$$

 $\sigma(\cdot)$  = Sigmoid Function, W<sub>i</sub> = Weight Matrix of the Output Layer,  $b_o$  = Bias Vector,  $\boldsymbol{0}$  = Value between 0 and 1 which represents the probability of the input being adversarial. In summary, TAHN is an efficient methodology which effectively uses both textural and adversarial features for improved robust predictions. By blending these features in the Concatenation Layer and then refining them through multiple Dense Layers, the network can better defend and sustain effectively against adversarial attacks while still able to generate accurate binary classifications on whether the input data is real and its originality is maintained or the integrity of the data which was sent as input has been altered. Thus, making TAHN a highly secured and powerful in cases such as Deepfake detection, Adversarial defense in Machine learning which is prone to the malicious attack.

### VIII. IMPORTANCE OF TAHN

In the battle against adversarial threats, TAHN stands out as a critical advancement because of its combination in two differentiable parts – an adversarial learning and texture-based feature extraction. Unlike the other traditional models which generally includes defensive strategies in adversarial robustness or texture analysis independently, TAHN can learn both of the modules simultaneously and achieve a comprehensive defense mechanism. The need for this hybrid approach is crucial as it allows the network to not only detect subtle adversarial manipulations but also analyses texture and the inherent structure of the data, which leads in predicting more accurate results. Incorporating a blockchain-based consensus mechanism into the TAHN model enhances the overall trust, security, and resilience of the deepfake detection pipeline. By using consensus protocols, such as Practical Byzantine Fault Tolerance (PBFT) or Proof of Stake (PoS), the outputs generated by TAHN (i.e., whether a media file is authentic or manipulated) can be validated across



multiple decentralized nodes before being accepted. This ensures that no single entity can alter the results, making the detection process tamper-proof and verifiable. Additionally, consensus enables distributed logging of adversarial attacks and detection outcomes, allowing for transparent auditing and collaborative defense strategies in real time. This strengthens TAHN's reliability in adversarial environments by introducing an immutable and trustless verification layer. TAHN is efficient compared to any other model in the fields like deepfake detection because of its ability to simultaneously examine the authenticity of both the textual and visual patterns while being resilient to adversarial perturbations making it more superior from other models. Because of its ability to handle various type of data, including images, text and multi-modal inputs – TAHN ensures its wide range of applicability, while at the same time establishes it as a powerful tool in enabling security of machine learning system under adversarial settings.

#### VIII.CONCLUSION

The rapid advancements in deepfake technology present both remarkable opportunities and significant threats to the digital world. While deepfakes are utilized for educational and creative purposes, they have also become powerful tools for misinformation, fraud, and security breaches, posing a serious risk to the authenticity of the media consumed across various platforms. Numerous research efforts have been undertaken to mitigate the risks associated with deepfakes, yet as AI continues to evolve, so does the quality and sophistication of deepfake technology, making detection increasingly challenging. To address this growing concern, we have proposed a novel approach called TAHN (Textural Adversarial Hybrid Network). Deepfake technology poses a significant threat due to its ability to generate highly convincing synthetic media, making robust detection methods essential. In this study, we examined the effectiveness of the Textural Adversarial Hybrid Network (TAHN) in resisting adversarial attacks and improving deepfake detection accuracy. To further enhance the trust, transparency, and security of the detection process, we proposed the integration of blockchain consensus mechanisms.

By leveraging decentralized consensus, the detection results can be securely validated and immutably recorded, preventing tampering and enabling collaborative verification. This integration not only strengthens the resilience of TAHN in adversarial environments but also lays the groundwork for a transparent and trustworthy deepfake detection ecosystem. This innovative framework combines two distinct yet complementary technologies—adversarial learning and texture-based analysis—to enhance the robustness and accuracy of deepfake detection. Unlike traditional deepfake detection models that rely on a single detection strategy, TAHN integrates both adversarial and textural analysis, significantly reducing inconsistencies and error rates while improving prediction accuracy.

Future research should focus on further strengthening TAHN's capabilities by enhancing its scalability, adaptability, and real-time efficiency. As deepfake technology continues to evolve, advancements in AI-driven defense mechanisms like TAHN will be crucial in safeguarding digital authenticity, preventing misinformation, and ensuring the integrity of media across various domains. By continuously improving detection frameworks, we can stay ahead in the battle against sophisticated deepfake manipulations and protect the reliability of information in an AI-driven era.

#### REFERENCES

- 1. Mirsky, Y., & Lee, W.: The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys (CSUR), 54(1), 1-41 (2021).
- 2. Neekhara, P.: Adversarial Attacks on Artificial Intelligence Based Video and Audio Deepfake Detectors. rXiv preprint arXiv, 2002.12749 (2020).
- 3. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box Adversarial Attacks with Limited Queries and Information. In: International Conference on Machine Learning (ICML), pp. 2137–2146 (2018).
- 4. Dong, Y., Liao, F., Pang, T., Hu, X., Zhu, J., Su, H., Li, J., Wei, Z.: Boosting Adversarial Attacks with Momentum. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9185–9193 (2018).
- Akhtar, N., Mian, A.: Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access 6, 14410–14430 (2018).
- 6. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. In: IEEE Symposium on Security and Privacy (SP), pp. 582–597 (2016).
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing Properties of Neural Networks. In: International Conference on Learning Representations (ICLR), pp. 1–10 (2014).





- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y.: Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, 2672-2680 (2014).
- 9. Zhu, X., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, pp. 1–14 (2002).
- 10. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics 3(6), 610–621 (1973).
- 11. Muniraju Hullurappa, Sudheer Panyaram, "Quantum Computing for Equitable Green Innovation Unlocking Sustainable Solutions," in Advancing Social Equity Through Accessible Green Innovation, IGI Global, USA, pp. 387-402, 2025.
- 12. Tech Target Webpage, <u>https://www.techtarget.com/whatis/definition/deepfake</u>, last accessed 2024/09/10.
- 13. Britannica Webpage, https://www.britannica.com/technology/deepfake, last accessed 2024/09/09.





# INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com