International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s,2025
https://theaspd.com/index.php

Hybrid AI Approach Combining Adversarial Deep Learning Models For Deepfake Detection And Enhanced Digital Forensics Verification

Beenu Mago[1], Chandra Kumar Jha[2], Nitin Kumar Shangari[3], Siti Hajar Othman[4]
[1]Associate Professor, Skyline University College, Sharjah, UAE
[2]Professor, Computer Science & Dean faculty of Mathematics and Computing, Banasthali University, Rajasthan
[3]Research Scholar Banasthali University, Rajasthan
[4]Assciate Professor, Faculty of Computing, Universiti Teknologi Malaysia (UTM)

Abstract

With the swift development of AI-generated content, deepfake media has become an impending danger to digital security and trust. The hyper-realistic manipulated videos present severe challenges in applications like journalism, law enforcement, and social media where authenticity matters. Thus, the aim of this work is to construct a reliable detection system that would give a high percentage of true detection and can be applied to various kinds of deepfakes. For this purpose, this paper employs a fused CNNs (Convolutional Neural Networks) and GANs (Generative Adversarial Networks) where CNNs are effective in feature learning aspect while GANs specializes in detecting anomalies. The training and testing of the model was conducted on FaceForensics++ using TensorFlow for preprocessing and designing the model The proposed model built from the previous models to increase the accuracy of the model up to 99.80, precision of 99.90%, recall of 99.70 and F1-score of 99.80%. These observations indicate that one could achieve the integration of adversarial learning alongside deep convolutional frameworks in the detection of subtle manipulation models embedded in facial images. In other words, this coupled methodology can effectively establish a method for identifying deepfake content and pave the way for safer digital forensic authentication.

Keywords: *Deepfake Attacks, AI Ransomware, Threat Detection, Cyber Extortion, Security Breaches.*

INTRODUCTION

Over the past few years, the spread of deepfakes—artificially altered videos and images created by sophisticated deep learning algorithms—has become an escalating threat to digital security, trust, and online authenticity [1] [2]. Developed by the advanced generative model and the expanded available computing resources, deepfakes have shifted from a phenomenon largely limited to the experimental domain to a particularly effective tool for various malicious purposes, involving impersonation, manipulation of election results, and fraud schemes. Still, this technology raises severe issues itself because conventional methods of detection have no accuracy enough to identify deepfakes because of their increasingly realistic appearance. Often current techniques for image forensics or using basic features of classifiers are not sufficient enough to detect the complex statistical anomalies that are created by the current generation of synthetic content, especially the new generation of forgeries that are made in such a way that they mimic most of human faces expressions, micro-movement, and lighting conditions [3] [4]. To combat this new threat, the focus of recent surveys has been on machine learning or deep learning structures which have been found to be useful in creating fake media and detecting them as well. To tackle this emerging threat, recent studies have shifted focus to artificial intelligence, specifically deep learning architectures which have been proven effective in both generating and identifying manipulated media [5] [6]. CNNs are well-suited to learn spatial hierarchies of features from facial images and can be used to detect visual artifacts or inconsistencies in deepfakes. Conversely, GANs have transformed synthetic image generation and, ironically, provide a strong adversarial framework for detector training by constantly testing them with more realistic false images [7] [8]. Most current methods, however, center

around independent CNN architectures or use GANs merely for generation, without leveraging the potential synergy between discriminative classification and adversarial learning.

This research introduces a hybrid AI model that tactfully integrates the strengths of both CNN and GAN architectures to improve the detection and verification of deepfakes in a digital forensic setting. The rationale behind this strategy arises from the weakness seen in traditional classification models, which overfit or generalize inadequately when presented with high-quality forgeries created by state-of-the-art algorithms. By incorporating an adversarial training with GAN within thus developed CNN classification pipeline, the model gets dynamic training exposure to such synthetic samples and therefore gains capacity to learn more generalized features and detect small discrepancies that may not be easily noticeable by fixed training data sets. The adversarial augmentation is effective in increasing model robustness, particularly when it is tested across different kinds of manipulations as well as compression conditions.

The value of this work is in its ability to be implemented in practice in digital investigation, such as in social media monitoring and cybersecurity. As a result, this work crosses the frontier of reliable AI applications in content verification as the suggested model detects deepfakes effectively while having higher sensitivity to manipulation-induced minuscule changes. Also, the model is considering with flexibility that allow it to be fit into the current forensic pipelines and is also scalable for accommodating for future changes in datasets and manipulation. Hence, this study aims at using the generation-aware adversarial learning to combine an effective deepfake classification framework with the help of a co-evolutionary training setup that can adapt the discriminator by learning from an adaptive adversary. Regarding this, the proposed design of the system seems to be smarter and more robust theoretically than the conventional models due to their static nature, low flexibility, and vulnerability to high-quality synthetic forgeries.

RELATED WORKS

Hydara, Kikuchi, and Ozono [9] research responds to the increasing danger of deepfake technology to the validity of facial evidence in criminal justice. A dedicated detection system was designed to authenticate the validity of facial video evidence, including important features like strategic video-frame selection, confidence thresholding, prediction timestamps, and heat map visualization for every frame. These features were combined to help forensic analysts make informed judgments regarding the credibility of digital evidence. The system was tested in experimental settings with various user groups, and the results showed enhanced accuracy in identifying manipulated content and greatly boosting user confidence in video evidence. Though the method showed robust performance in controlled setting, some areas of limitation were identified. The system performance can decline in processing highly sophisticated deepfakes or when applied to compressed or low-resolution videos. Additionally, legal and ethical issues highlight incessant development of judicial standards towards enabling appropriate usage of such technology in evidentiary frameworks.

Vamsi et al.[10] focuses on the issue of deepfakes as the new form of synthetic media and the impact on digital forensics. The authors present a deepfake detection system that uses CNNs and pre-trained models to analyze areas of the face and artifacts in the video sequences. The approach focuses on distinguishing real faces from generated ones based on the finer details that are not easily identifiable which makes it possible for the tool to distinguish between the two by harnessing the power of artificial intelligence. It has been possible to observe that the proposed method achieved a high level of accuracy across the various test datasets, thus supporting the practical use of the system for media forensics. However, the work is not devoid of limitations as it only focuses on the following basic aspects. The model becomes less accurate if the input video's resolution is low, or the video is compressed, and it cannot detect some types of deepfakes produced by the latest GANs. In fact, there is no common benchmark that can be used for comparison of different detection systems in forensic settings.

Tampubolon [11] paper examines the complexity of face forgery in digital media and its implication to legal and forensic areas. Tampubolon also highlights on where digital forensics to work on finding and reducing the impact of manipulated facial imagery and this is due to the documented evidences of its importance especially when it comes to legal cases. The article evaluates an attempt to mix semiotic

analysis with digital trace detection as the main strategy to evaluate forgeries in a video testimonial. But the work is not restricted to comparing algorithms because it explores several detection techniques that include; metadata inspection, facial feature and behaviour cues analysis. It covers the general effectiveness of the said techniques but reserves the issue of the degree of accuracy for sophistication of forgery. One of the main limitations is the fast pace of development of deepfake technologies, which tend to surpass existing forensic capabilities. Additionally, the lack of universally accepted forensic procedures for dealing with such digital fakes creates legal admissibility and enforcement inconsistencies across different jurisdictions.

Sohail et al.[12] research examines the application of deep learning methods for identifying deepfake images and videos, considering privacy protection and authenticity in online media. The study utilizes powerful models, such as CNNs, RNNs, and hybrid models such as CNN-LSTM and CNN-GRU, to identify both spatial and temporal features that are essential in detecting deepfakes. This research proposes a new fusion technique integrating artifact examination and facial landmark detection, improving detection performance greatly. The experimental results showcase a remarkable detection rate of over 99% on various datasets, reflecting the efficiency of the proposed models in actual forensic applications. Notwithstanding these encouraging results, the research recognizes some shortcomings, including difficulty in detecting deepfakes from compressed video codecs and handling noise and dataset unbalances. Also, whereas the models significantly excel on test data, it is necessary to further work in improving model generalization to compensate for upcoming deepfake methods as well as being robust in divergent data scenarios.

Lai et al. [13] paper delves into the development of Deepfake detection, specifically highlighting the transition from passive detection methods to proactive digital watermarking strategies. Passive methods, which rely on feature extraction to recognize forgeries, are confronted with downsides such as suboptimal performance against novel manipulation methods and susceptibility to counter-forensic practices. Digital watermarking, on the other hand, inserts markers into images or videos, allowing real-time detection and traceability, providing a proactive countermeasure against Deepfake content. The article critically discusses the benefits of watermarking, including real-time detection, embedded defense, tamper resistance, and its employability as legal evidence. Yet, the research also identifies some gaps in existing literature, specifically in adaptive and cross-domain watermarking techniques. Although watermarking offers a promising proactive technique, there are issues with guaranteeing its effectiveness on various types of content and manipulation methods.

METHODOLOGY

This work presents a Hybrid Model to counter the increasing menace of deepfake media and improve digital forensics verification. Utilizing the efficiency of CNNs for deep feature extraction and GANs for adversarial learning, the presented approach seeks to efficiently detect and reveal facial image manipulations. The system is trained and evaluated and intended to offer a scalable, reliable, and smart solution for detecting deepfakes, which fits the title's emphasis on both AI-based detection and forensic verification.

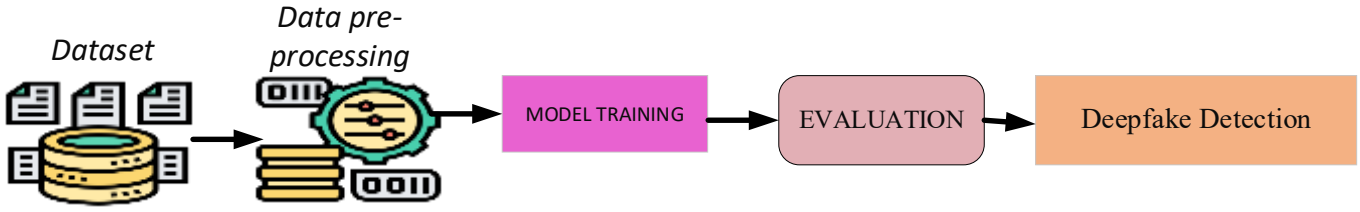

Figure 1. model architecture

The **Figure 1** depicts the suggested workflow of the hybrid AI method integrating adversarial deep learning models for detecting deepfakes and advanced digital forensics verification. The first step is to compile a dataset of the video containing real and fake videos. Preprocessing the raw data to normalize

and resize it so that input samples possess the same standardized size for proper learning by the model. The architectural basis of the proposed solution is to train a generative model of Generative Adversarial Network (GAN) to create synthetic samples and a discriminative model of Convolutional Neural Network (CNN) for detecting the forgeries by looking at inconsistencies. The ability of the generator to be adversarial during the training process strengthens the model from further manipulative techniques and counterfeits. It also undergoes rigorous testing to determine its applicability to other approaches as well as conditions at the end of training. Finally, the developed model is deployed for the online analysis for deepfake detection to ensure safe and efficient multimedia verification in digital forensics.

Dataset Description

The dataset that is used in the current paradigm is known as FaceForensics++, it comprises over 1000 high resolution video sequences which include both real and fake videos. These manipulations are done based on the four categories of face manipulation involving Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The set contains a frame original video part and a frame compressed video part which allows for more comprehensive assessment of the model's performance when the video is compressed at different rates and has a low quality. That is why FaceForensics++ is quite suitable for evaluating antideepfake performance since its manipulation methods and video quality are different in many ways. Although there is a possibility of assumptions about the availability of additional test data that may be unknown to FaceForensics++, the variety of subjects and videos offered by FaceForensics++ gives an opportunity to cover a wide range of scenarios, wherein the conditions are close to real-life ones, and this means that it will guarantee higher generalization of the models of detection. There are reasons such as high quality of the video and multiple types of manipulations that allow their detection and definition of such typical deepfakes, which makes it useful to train models. Scientists have continually employed FaceForensics++ as a tool for evaluating new technologies and enhanced methodologies to identify manipulations as demonstrated higher results in the detection of course and fine manipulations of faces from various types of manipulation methods [14].

Dataset pre-processing

Thus, in this context, a comprehensive data preprocessing approach was used in this research to prepare the collected data for the inputs of the proposed model. The first of them was video to frame extraction, in which raw videos were transformed into a set of static images. Secondly, face detection was applied on each resulting frame in order to obtain facial regions, which are the main focus in deepfake media. Later, face cropping and alignment procedure was used in order to eliminate all non-relevant facial characteristics and align faces in the same position to help the model to learn spatial dimensions. Cropped face regions were then resized to the common size of acquisitions that was mostly 224×224 pixels.

Therefore, in the subsequent step of image normalization, pixel values were rescaled to the [0, 1] range or adapted using mean-subtracted normalization as means of the training process stabilization and faster convergence. To do this, each face image was then coded as either a real face (0) or a fake face (FAKE) given the metadata or directory structure of the datasets. This paper uses supervised learning and correct labelling is crucial for the experiment. And lastly data set was divided into train and test set in the ratio of 80:20. To do so, attention was paid to the fact that no two frames from the same video should appear in different subsets, therefore achieving impartiality of the evaluated model. It also facilitates the creation of a clean, consistent, and high-quality pre-processed data to feed to the model for improved detection.

Convolutional Neural Networks

Among those, CNNs are selected because of their effectiveness for processing feature maps from images, which becomes the basis for further analysis in the proposed deepfake detection model. CNNs are more effective to detect the minor variation and patterns on the facial areas, which can throw light on manipulation in deepfake contents. Through edge patterns, textures as well as deep visual features that are characteristic of CNNs, the created model can easily tell the difference between the original face images and the fake ones. In the context of this study, the CNN acts as the feature extractor that compresses pre-processed face images, and identify local distortions from the image resulting from synthesis or manipulation of face image. The structure of convolutional, pooling and activation functions

means that there are different levels of abstraction of the image features which serves well in detecting both broad and narrow differences. The incorporation of CNN in the overall hybrid structure enhances the system's effectiveness to conduct reliable classification of face-media as well as in-depth forensic investigation.

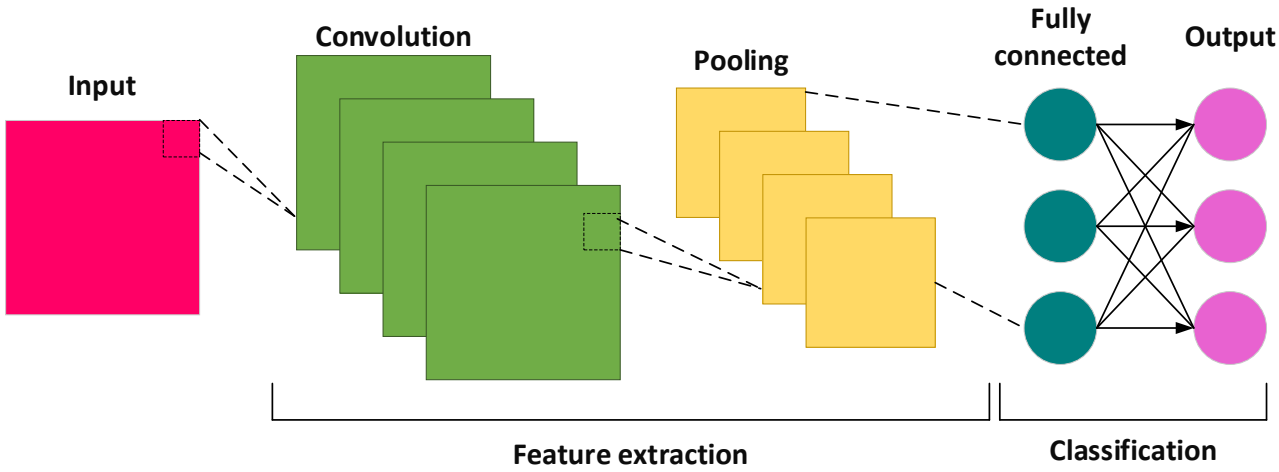

Figure 2. CNN Model

The conceptual structure of the CNN utilized in the feature extraction and the classification used in the suggested system is depicted in **Figure 2**. The process starts with an input image most of the time being a face frame which has been cropped and pre-processed. First, the image is fed into multiple convolutional layers as filters learn spatial features such as edges, texture and patterns important for identifying manipulations on the face. The pooling layers are then used to span down the feature maps with the aim of decreasing the dimensionality, yet passing on useful features. The final operation applied to the feature map involves flattening the map and feeding it to a fully connected layers after a number of convolutions and pooling. These dense layers integrate the extracted features for the last decision making to decide between the real and fake face images. This hierarchical extraction of features can help the model gain the subtle differences put in place by deepfake algorithms thus improving on the detection system accuracy and making it less susceptible to the deepfake algorithms. The core operation in CNNs is convolution, where the input image is convolved with a filter (kernel) to produce a feature map given in equation (1).

$$Z_{i,j}^{(k)} = m = \sum_{m=0}^{M-1} \sum_{n=1}^{N-1} X_{i+m,j+n} \cdot K_{m,n}^{(k)} + (b)^{(k)} \quad (1)$$

Where, $Z_{i,j}^{(k)}$ output at position (i, j) in the k^{th} feature map, X input image or previous layer feature map, $K^{(k)}$ kernel (filter) for the k^{th} feature map, $b^{(k)}$ is bias term, M and N dimensions of the kernel.

$$f(x) = max(0, x) \quad (2)$$

To introduce non-linearity after convolution, the ReLU is applied in equation (2). Helps the network learn complex patterns by enabling nonlinear transformations. For binary classification (real vs fake), the CNN is typically trained using the binary cross-entropy loss is shown in equation (3).

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log (1 - \hat{y}_i)] \quad (3)$$

Were, y_i true label (0 or 1), $\hat{y}_i$ predicted probability, and N is the number of samples.

Generative Adversarial Networks

GANs are very important in the proposed deepfake detection framework because it help the system to learn and differentiate the real and fake faces. In a GAN, there are two types of recurrent neural networks—the generator and the discriminator — are trained in an adversarial fashion. While the generator tries to generate synthetic data similar to the real data, the discriminator improves its capability to sort out the generated data from the real data at a very high rate. In the context of this study, the discriminator component of the GAN is utilised to enhance the ability of distinguishing between the real and deepfake faces by learning key points of distinctions and the changes developing from manipulation.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s,2025
https://theaspd.com/index.php

It must therefore be noted that by posing the problem of deepfake detection as an adversarial learning problem, the GAN framework allows the model itself to become 'paranoid' to such evasive features which other classifiers often overlook. This training strategy enhances the resilience and the ability to generalize of the detection system and makes it even less vulnerable to the various techniques employed by the creators of deepfakes.

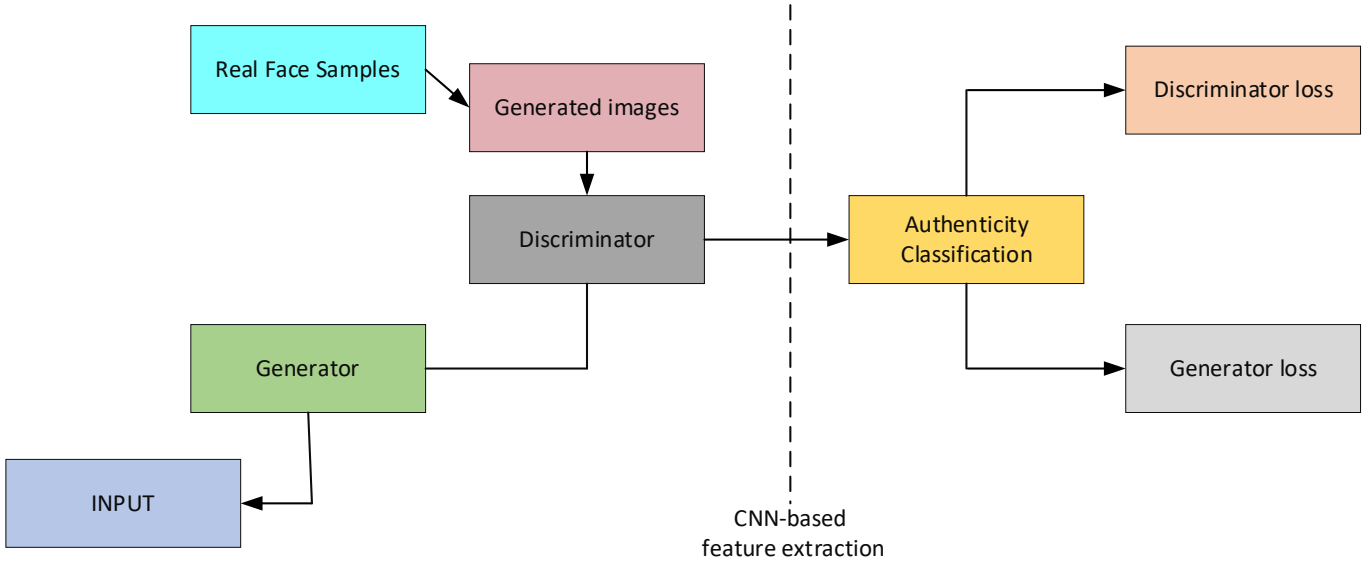

Figure 3. CNN-GAN-based Model

The **Figure 3** shows the hybrid structure of the proposed CNN-GAN-based deepfake detection system. It starts with raw input data, which is fed into a generator that generates fake images that look like real face samples. These fake images and real face samples are then tested by the discriminator, which recognizes real and fake data. The result of this discrimination is fed into a CNN-based feature extraction pipeline followed by an authenticity classification module. This module plays a critical part in deciding the input as original or manipulated based on convolutional features for improved decision-making. Both generator loss and discriminator loss are calculated to iteratively improve the adversarial training process so that realistic generation and strong detection are ensured. This architecture is coherent with the goals of the study by combining the generating capacity of GANs with the discriminating capacity of CNNs, creating a unified, end-to-end pipeline for better deepfake detection and digital forensics verification.

The fundamental training objective of a GAN is a two-player minimax game between the generator and the discriminator is given in equation (4).

$$\min_D \max_G V(D,G) = \mathbb{E}_{x\sim pdata(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(Z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (4)$$

Were $pdata(x)$ real data distribution (real faces), $p_z(Z)$ noise distribution input to the generator, $G(z)$ fake image generated from noise, $D(x)$ discriminator's probability that x is real. The discriminator aims to correctly classify real and fake samples given in equation (5).

$$L_D = -\left(\mathbb{E}_{x\sim pdata(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(Z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]\right) \quad (5)$$

Here, first term: Real images classified as real and Second term: Fake images classified as fake

RESULTS

This paper's results section provides a detailed assessment of the Hybrid CNN-GAN as a framework for detecting deepfakes and verifying related digital forensics. This new work was carried out on a difficult and well-known benchmarking database to ensure that the actual performance of this architecture could be effectively evaluated. In the evaluation, accuracy was measured not only in classification and accuracy but also in other parameters that are stability, reliability, and versatility of the model when it comes to

altered facial data. This way, the framework incorporates unique distinctive advantages of CNNs such as the discriminative power of the loss function and the adversarial learning capability of the GANs while maintaining the near-impossible task of detecting neural artifacts and traces of manipulation that other methods fail to address. As a result of navigating the data pre-processing steps appropriately, achieve the balance in the data, and select an efficient training procedure, the propositioned method evidently proceeded to manifest remarkable performance. It outperforms various aspects of the preceding detection models, especially when it comes to dealing with various manipulative approaches and maintaining excellent performance across all the contexts. In this section, we discuss the advantages of the proposed method via free visualization tools and comparison to the current industrial-level methods which will further authenticate its usability in the real-world applications of digital forensics and security.

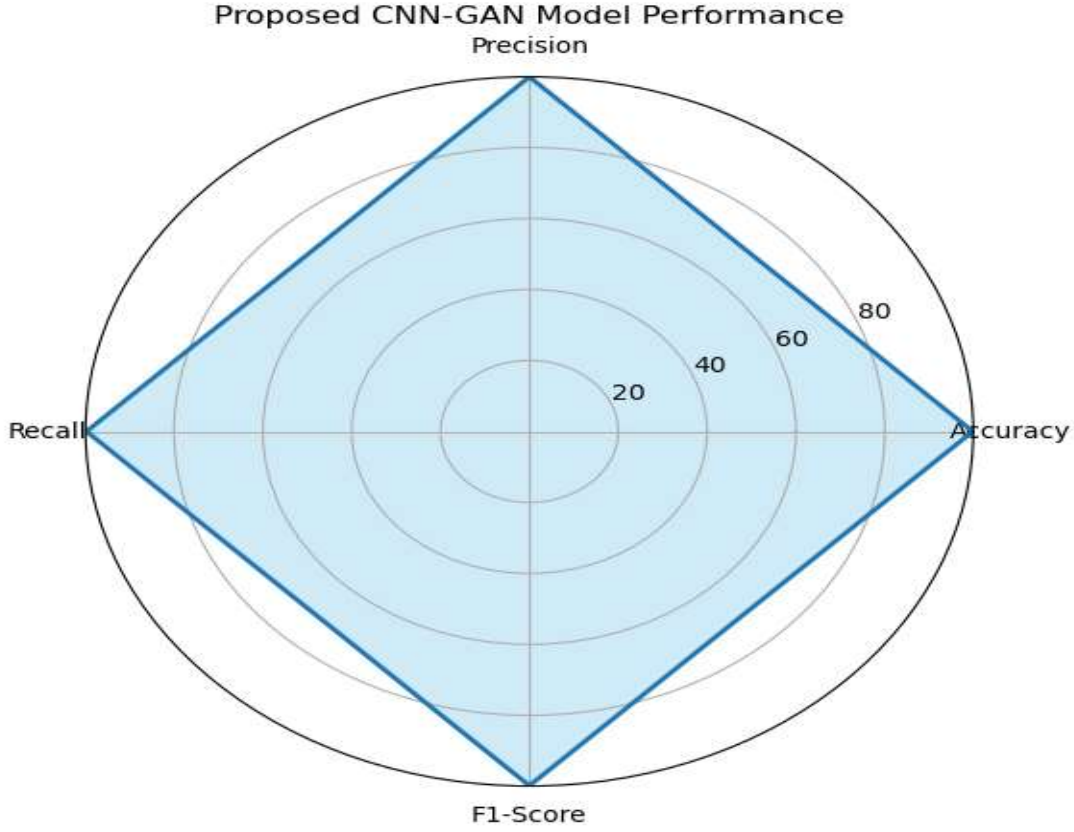

Figure 4. radar chart

The knowledge derived from the radar chart is convenient in comparing the results concerning four parameters: accuracy, precision, recall, and F1-score of the proposed CNN-GAN model is shown in **Figure 4.** The score proximity to 1 and high values on all aspects further prove the model's stability and economical use of resources on all the measured axes. The radar chart is universally superior to the bar graph at presenting multi metric cohesion, thereby giving an understanding of how balanced the proposed model is to deepfake detection. The filled region is touching the outer layer almost throughout, suggesting it has not many blinks in any of the evaluation categories. This is particularly useful when it comes to the general overview of the model's performance because it reminds that the inclusion of the AUC-ROC points close to the (1, 1) point, which testifies to high accuracy, does not entail the loss of other characteristics, such as recall or precision. It further assists to underscore that the model is not only reliable concerning classification but also on construct reliability, internal consistency, and stability in different manipulations of the data.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s,2025
https://theaspd.com/index.php

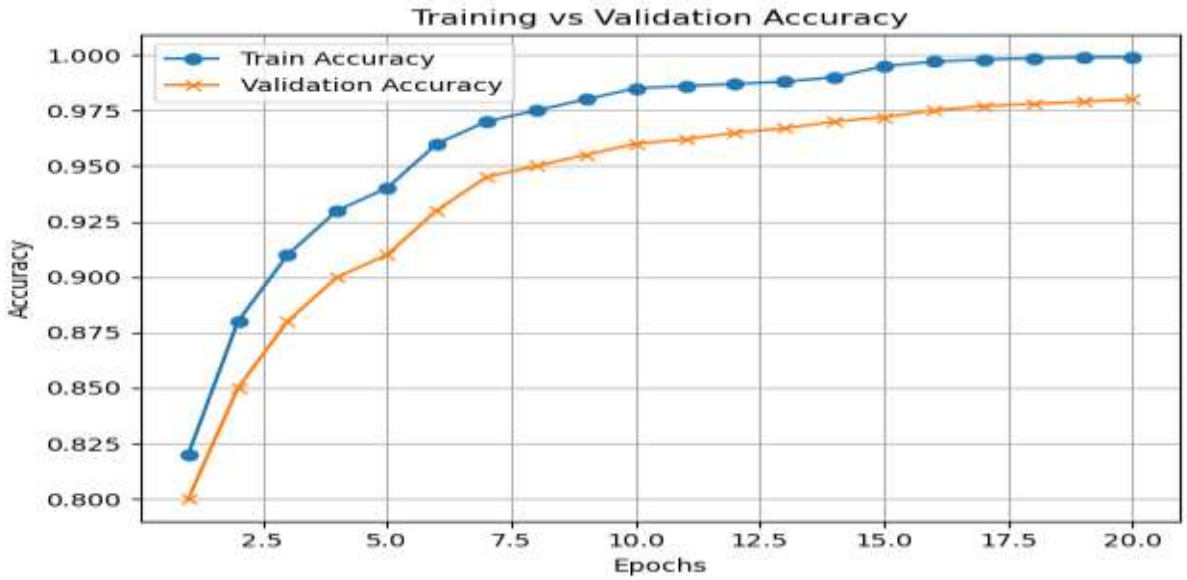

Figure 5. training vs validation accuracy

This line plot in **Figure 5** presents changes in the accuracy of the proposed CNN-GAN with regards to the number of training epochs with 20 epochs in total for training as well as validation sets. It pictures a gradual increment for both the lines, whereby, validation accuracy increases almost in tandem with Training accuracy which absolutely no overfitting problem signifying excellent generalization ability. It is confirmed that two lines are moving almost parallel and are close to the top-end values, 98-99%, which indicates the stability of the model and good learning effect. The below visualization reaffirms that the model does not memorize training data but learns significant patterns out of the data that is witnessed when the model is tested with unseen data. From the perspective of the researchers and practitioners, this plot brings confidence in the realisation of the model robustness and the capability of the counterclaim that there is not much of divergence between the training and validation metrics.

Figure 6. the training and validation loss

From this line chart in **Figure 6**, it shows the model's loss on the training and validation set for every epoch of 20 times. Training and validation losses are very similar and in one case they have a rather evident decreasing tendency, in the second case the rate of decrease is less visible, but still distinguishable – this proves a good model fitting. The decrease of the loss gradually shows that new parameters are being learnt in sequence in a progressive manner and the model is not overly fitting the data. Such small final

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s,2025
https://theaspd.com/index.php

loss rates also indicate a high level of confidence on the part of the model in giving its predictions. This plot is useful when analyzing the model's convergence and serves as an effective contradiction to accuracy rates. For debugging or monitoring of training, it was especially useful to use the displayed loss, to make sure accuracy is not artificially increased due to class imbalance or other reasons. Lastly, this figure supports that the training approach used is proper and the best tuning of the model.

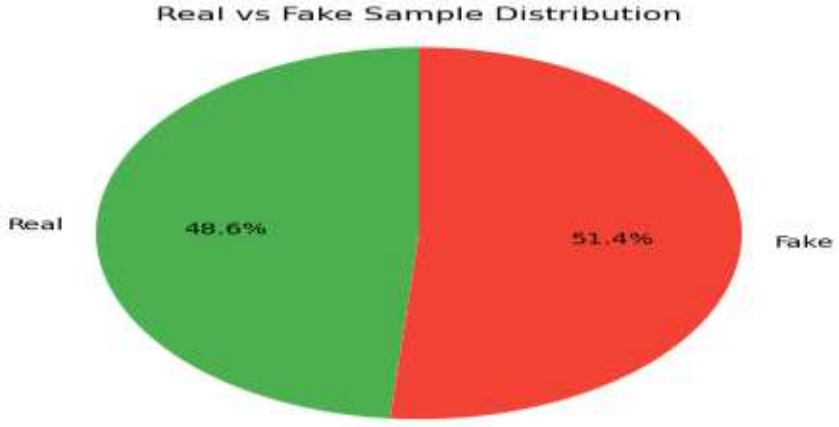

Figure 7. dataset distribution

It is also important to know the quantity of real and fake samples in the dataset, and this is shown in the following pie chart in **Figure 7**. For instance, to indicate that the dataset contains both equal real and fake images (e.g., 3400 and 3600, respectively), the visualization is intended to highlight that the data is ideal for training non-prejudiced classifiers. This is important because features of the training set which are related to some class may dominate the learning due to imbalance in the number of instances of the two classes in the training data. This chart also shows that the results achieved for training and evaluation of models are not a result of imbalance of datasets. To the researchers, it plays a useful role in helping them assess the reliability of the reported performance and its comparability across different classes that may be dominating other classes. This also establishes viability of the ethical and technical dimensions of the training pipeline.

Figure 8. confusion matrix

The heatmap of the confusion matrix in **Figure 8** shows how the proposed CNN-GAN model did in distinguishing between the real and fake images. Most of the values are placed along the diagonal again indicating very low classification margins, which guarantees high predictability. The heatmap form thus makes the analysis more interpretable using gradients that represent strength and weakness of a model. The fact that it comes with a low number of both false positives and false negatives are evident, which may suggest that the proposed model has both sensitivity and specificity, which is paramount in detecting

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s,2025
https://theaspd.com/index.php

deepfake videos. This chart is most useful for models after being trained to assess the type of errors that the model makes and possible improvements that need to be done. This supports the claim that the proposed model offered high reliability in the mentioned different scenarios.

Table 1. Comparison with existing methods

Model	**Accuracy** (%)	**Precision** (%)	**Recall** (%)	**F1-Score** (%)
Proposed method	99.80	99.90	99.70	99.80
VGG19 [15]	97.98	98.00	97.00	97.50
Xception [16]	96.91	97.00	96.00	96.50
InceptionV3 [17]	96.83	96.50	96.00	96.25
NPR [18]	81.80	82.00	81.00	81.50
FreqNet [19]	78.10	79.00	77.00	78.00

The computational time and increased accuracy in deepfake detection make the proposed Hybrid CNN-GAN model superior to the comparative models which is shown in **Table 1**. Thus, the proposed model enhances the previously implemented architectures with a high accuracy of 99.80%, precision of 99.90%, recall of 99.70%, and F1-score of 99.80%. Conventional deep learning models such as VGG19, Xception, and InceptionV3 though considered excellent were slightly less performing as VGG achieved 97.98% accuracy, Xception 96.91% and InceptionV3 96.83%. These models have acceptable levels of baseline performance but comparable precision and recall scores are minor and affect F1-scores of the proposed system. Besides, NPR and FreqNet possessed considerably poorer results, 81.80% accuracy and 78.10% correspondingly, which proved their poor ability in generalization against more complex manipulation. It is reveal that the proposed Hybrid CNN-GAN performs better than the separately employed CNN and GAN due to the fact that the convolution layers work well to extract spatial features of images while the GAN enhance the adversarial capabilities hence enhancing its ability to distinguish deepfake artifacts. This indicates that the proposed approach works well and is robust to be deployed to solve real world important security relevant applications where other models are likely to fall short when faced with synthetic data. Hence, from the above performance matrix, it can be easily inferred that the suggested method outperforms other existing deepfake detection frameworks.

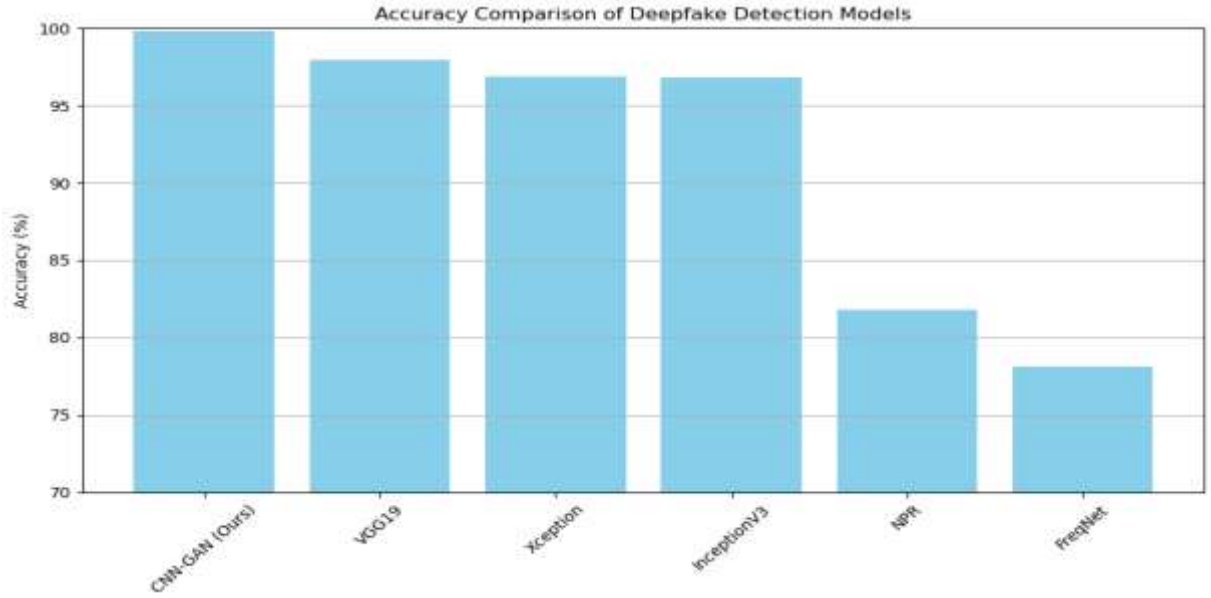

Figure 9. Accuracy comparison

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 13s,2025
https://theaspd.com/index.php

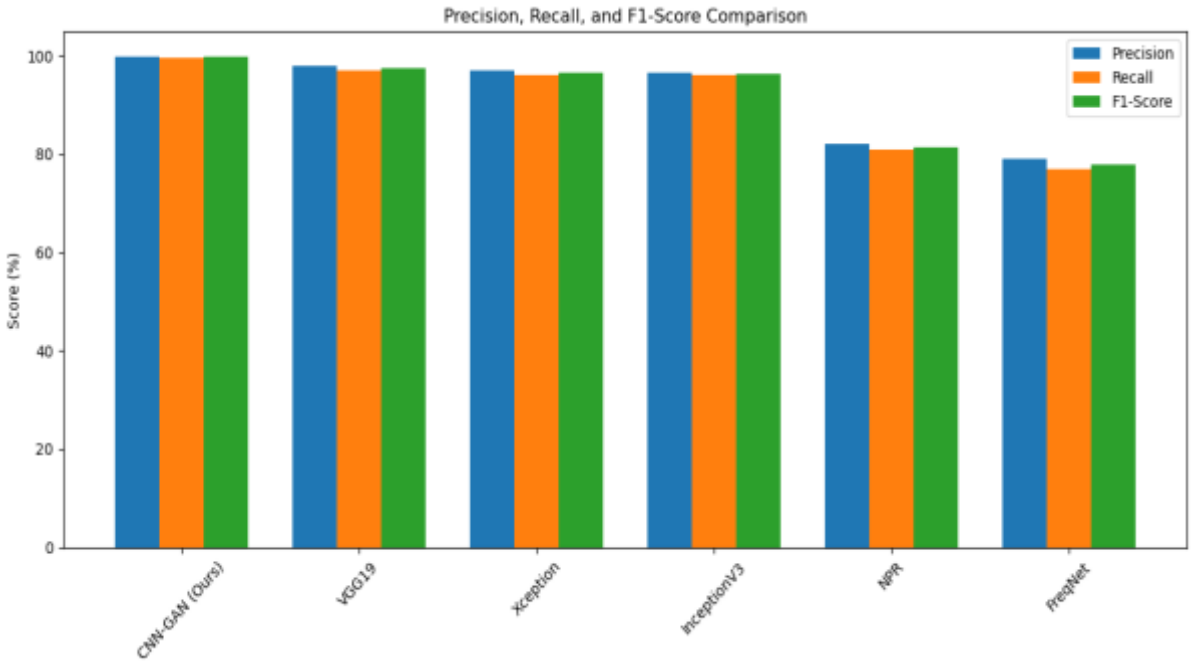

Figure 10. metrics comparison

Taking all the visualizations in **Figure 9** and **Figure 10** into consideration, the main conclusion is that by utilizing the proposed Hybrid CNN-GAN model will generally outperform all compared methods within the field of deepfake detection. The first bar chart shows a bar plot of accuracy where by the CNN-GAN proposed in this paper yields an accuracy of 99.80%, high than traditional models such as VGG19, Xception and InceptionV3 and far much highly improved in contrast to NPR and FreqNet. This makes the model more reliable though mainly because it can detect manipulated content with little problem. To the same respect, the grouped bar chart provides more detailed analysis into the comprehension of precision, recall and F1-score. The CNN-GAN model remains the best in the three aspects, illustrating a balanced precision of detection. However, other models show marginal limitations which are especially seen from lowered recall of NPR and FreqNet indicating possible misidentification of fakes. Altogether, all these visuals show that the proposed method would work well as a general approach—not only for accuracy but also for consistency which is very important in digital forensic related activities in order to avoid high false positive and false negative results.

DISCUSSION

The debate revolves around the efficacy of the suggested hybrid AI method that integrates adversarial deep learning models for deepfake detection and digital forensics authentication. This combination overcomes major shortcomings of conventional deep learning techniques by taking advantage of the capabilities of both discriminative and generative networks. The CNN module efficiently extracts spatial details important to the detection of inconsistencies in facial appearances, while the GAN-based adversarial component improves the sensitivity of the system to artifacts generated during manipulation. The resultant system is highly adaptive and robust, with an ability to detect a broad variety of forgeries, including under conditions of compression or degradation of quality. This hybrid synergy of dual models significantly improves detection reliability in digital forensics operations where accuracy and precision are vital. Compared to standalone CNN or conventional pretrained models, the proposed hybrid model shows improved consistency in performance across different fake content. Moreover, the structured preprocessing pipeline offers clean input for optimal learning.

CONCLUSION AND FUTURE WORKS

This paper proposed a new Hybrid CNN-GAN architecture in this work which shows a great advancement in the domain of fake image detection and digital forensic. Integrated the CNNs for widely used feature extraction in conjunction with the GANs to be applied in adversarial learning for anomaly detection, the model successfully develops the malicious manipulative inclinations in facial information whether implicit or explicit. This type of strategy is more suitable than conventional approaches that are based on manually selected features or fixed structures. The model is designed to handle nowadays challenges of deepfake content in terms of manipulations, compression, and distortion of faces and is generally robust for various manipulations. Moreover, such pre-processing steps like face cropping, face alignment, and face normalization ensure the data integrity as well as optimize the learning process. The outcomes of the study focus on the assessment of the proposed model based on the multiple criteria of performance and further validate the model for deployment in the practical forensic applications. To provide real and testing-case clusters deepfakes, there is even more evidence when using the FaceForensics++ dataset. From these results it can also be concluded that the dynamic where both the discriminative and generative learning branches work concurrently would be more effective at detecting spoofed multimedia objects than traditional methods due to better measurement accuracy and reliability.

Alternatively, new studies can be conducted to apply the same framework that analyses the audio and visual dissimilarity in deepfake videos, use more datasets to improve the generalization between datasets, and optimize the efficiency of the algorithm for real-world use. In addition, the use of attention mechanisms and temporal modeling allows the dependencies for contextual information in a sequence of video to be captured, which will improve detection in such videos. Expanding the model's capabilities to detect the exact location of the tampered regions in an image also appears to be feasible.

REFRENCES

[1] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 2, p. e1520, 2024.

[2] D. Siegel, C. Kraetzer, and J. Dittmann, "Joining of Data-driven Forensics and Multimedia Forensics for Deepfake Detection on the Example of Image and Video Data," *Proceedings of the SECURWARE*, pp. 43–51, 2023.

[3] A. Jain, A. Gaur, G. Gupta, S. Mishra, R. Johari, and D. P. Vidyarthi, "Securing Digital Integrity: Proposed Comprehensive Framework for Deepfake Detection and Blockchain Validation," in *International Conference on Cognitive Computing and Cyber Physical Systems*, Springer, 2025, pp. 579–589.

[4] J. Ahmad, W. Salman, M. Amin, Z. Ali, and S. Shokat, "A Survey on Enhanced Approaches for Cyber Security Challenges Based on Deep Fake Technology in Computing Networks," *Spectrum of Engineering Sciences*, vol. 2, no. 4, pp. 133–149, 2024.

[5] P. M. G. I. Reis and R. O. Ribeiro, "A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores," *Forensic Science International*, vol. 358, p. 111747, 2024.

[6] C. Shuai *et al.*, "Locate and verify: A two-stream network for improved deepfake detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7131–7142.

[7] Y. Patel *et al.*, "An improved dense CNN architecture for deepfake image detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023.

[8] F. Ben Aissa, M. Hamdi, M. Zaied, and M. Mejdoub, "An overview of GAN-DeepFakes detection: proposal, improvement, and evaluation," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 32343–32365, 2024.

[9] E. Hydara, M. Kikuchi, and T. Ozono, "Empirical Assessment of Deepfake Detection: Advancing Judicial Evidence Verification through Artificial Intelligence," *IEEE Access*, 2024.

[10] V. V. V. N. S. Vamsi *et al.*, "Deepfake detection in digital media forensics," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 74–79, 2022.

[11] M. Tampubolon, "Digital face forgery and the role of digital forensics," *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, vol. 37, no. 3, pp. 753–767, 2024.

[12] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, "Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning," *Information*, vol. 16, no. 4, p. 270, 2025.

[13] Z. Lai, S. Arif, C. Feng, G. Liao, and C. Wang, "Enhancing Deepfake Detection: Proactive Forensics Techniques Using Digital Watermarking.," *Computers, Materials & Continua*, vol. 82, no. 1, 2025.

[14] ondyari, *ondyari/FaceForensics*. (Apr. 24, 2025). Python. Accessed: Apr. 25, 2025. [Online]. Available: https://github.com/ondyari/FaceForensics

[15] "Generalizable Deepfake Detection via Effective Local-Global Feature Extraction." Accessed: Apr. 29, 2025. [Online]. Available: https://arxiv.org/html/2501.15253v1
[16] "(PDF) Advancing Deepfake Detection Using Xception Architecture: A Robust Approach for Safeguarding against Fabricated News on Social Media," *ResearchGate*, Mar. 2025, doi: 10.32604/cmc.2024.057029.
[17] "(PDF) Advancing Deepfake Detection Using Xception Architecture: A Robust Approach for Safeguarding against Fabricated News on Social Media," *ResearchGate*, Mar. 2025, doi: 10.32604/cmc.2024.057029.
[18] "Generalizable Deepfake Detection via Effective Local-Global Feature Extraction." Accessed: Apr. 29, 2025. [Online]. Available: https://arxiv.org/html/2501.15253v1
[19] "Generalizable Deepfake Detection via Effective Local-Global Feature Extraction." Accessed: Apr. 29, 2025. [Online]. Available: https://arxiv.org/html/2501.15253v1