Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Artificial Intelligence Generated Deepfakes as Instruments of Disinformation: Examining Their Influence on Public Opinion, Digital Trust, and Governance

Muhammad Mudaber Jamshaid[1], Ahmed Hassaan[2]*, Zeeshan Akbar[3], Sikander Niaz[4], Muhammad Nouman Siddique[5], Salman Akbar[6]

[1]Harvard Graduate School of Education (HGSE), 13 Appian Way, Cambridge, MA 02138, USA

Email: mudabbir@alumni.harvard.edu

[2]Raymond A. Mason School of Business, William & Mary, 101 Ukrop Way, Williamsburg, VA 23186, USA

Email: ahassaan@wm.edu

[3]Raymond A. Mason School of Business, William & Mary, 101 Ukrop Way, Williamsburg, VA 23186, USA

Email: Zakbar@wm.edu

[4]College of Cybersecurity & Information Assurance, Virginia University of Science and Technology, 2070 Chain Bridge Rd STE 100, Vienna, VA 22182, USA

Email: Sniaz362@vust.edu

[5]School of Public Policy, The London School of Economics and Political Science, Houghton St, London WC2A 2AE, United Kingdom

Email: noumanlatki@alumni.lse.ac.uk

[6]School of Education, State University of New York at Albany, 1400 Washington Avenue, Albany, NY 12222, USA

Email: sakbar@albany.edu

* Corresponding Author Email: ahassaan@wm.edu

Submitted _Nov 7,2025 | Published _Dec 18,2025

ARTICLE INFO

Received: 07 Nov 2025

Revised: 10 Dec 2025

Accepted: 16 Dec 2025

ABSTRACT

The paper examines the use of artificial intelligence-generated deepfakes as a socio-technical threat, one which can be developed at three interlinked levels: technical detection, human perception, and governance. We test four image-only image-deepfake detectors, including GAN-based, diffusion-based, Vision Transformer (ViT-B/16), and CLIP ViT-B/32, using the FaceForensics image dataset, which has undergone a uniform preprocessing pipeline that includes facial-cropping, image-alignment, resolution normalization, and quality filtering. Transformer-based models are, by a significant margin, more successful than both GAN- and diffusion-based detectors, with CLIP ViT-B/32 being the most successful and obtaining the highest classification accuracy and an almost perfect ROC-AUC, which highlights the importance of large-scale pretraining and attention-based models on synthetic media forensics. To add to these technical experiments, there is a human-subject experiment indicating that participants are always more efficient in authentic image recognition than in deepfakes, with a general deepfake detection accuracy falling close to that of chance, and with a relative weakness in differentiating between age groups. Deepfakes not only result in a high rate of false categorization but also cause a significant decrease in the level of trust, despite the fact that the perceived credibility scores are not only displaced but also significantly lowered. Lastly, the policy and regulatory text topic modeling has shown an unequal panorama of emerging but inconsistent governance with a focus on identity protection and election protection issues, and minor reference to actual enforcement tools. Combined, the results can be interpreted to support the potential and shortcomings of modern AI-driven detectors, the susceptibility of human judgment, and the necessity of an improved and more enforceable regulation and specific media literacy to maintain online trust.

Keywords: Disinformation, Deepfake detection, Media trust, Governance, Policy analysis, FaceForensics.

Introduction

The visual media production and manipulation have been changed with the advent of artificial intelligence, which has been rapidly evolving [1]. One of the most influential ones is the emergence of deepfakes, which are a type of fake images and videos created through machine learning algorithms and can recreate a human face to a point where the simulation becomes extremely lifelike [2]. Although early deepfakes showed distinct artifacts and discrepancies, as of today, the high-level neighbors have been made very realistic through the use of various types of generative AI models, such as Generative Adversarial Networks (GANs) and diffusion-based models [3]. Simultaneously, large-scale vision foundation models have also increased the pace of fidelity and availability of synthetic media generation [4]. With this technological change, there have been emerging opportunities in the areas of entertainment, art, and accessibility, as well as threats of extensive threats of misinformation, identity exploitation, and loss of civic trust, of its severity.

The spread of deepfakes has questioned, outright, the effectiveness of visual information in online communication [5]. The difference between authentic and manipulated images can hardly be made by even highly educated people [6], and the exposed people have been found to lose trust not only in the media but also in the more general institutions [7]. These problems are aggravated by the propagation of automated manipulation tools and the growing accessibility of realistic deepfakes [8]. The spread of automated manipulation tools and the increased availability of lifelike deepfakes only exacerbate these issues [9]. As a result, the development of powerful, Artificial Intelligence-based systems that can locate and label manipulated images became a significant field of study [10].

Despite significant progress, existing approaches face several limitations. Most of the detection models have limitations to their ability to generalize to different manipulation strategies or data sets due to the small training distribution or use of surface artifact patterns that might not endure within generative systems [11][12]. In addition, although different studies assess deepfake detectors algorithmically in a purely technical viewpoint, a smaller number assess the system in terms of its human aspect, meaning how users respond to synthetic content, how their confidence is altered by exposure, and how human reactions connect with algorithm performance [13][14]. The gaps provide incentives to conduct more intensive, multi-layered analysis that will connect machine-level capabilities to detect any vulnerability of human perception to a governance implication.

This research is concerned with the increasing interest in effective image-only deepfake detector algorithms that are able to differentiate between authentic and manipulated facial images, in addition to focusing on the human and governance issues that have cropped up because of the spread of deepfakes. The essence of AI-based deepfake detection models and the analysis of the impact of deepfake exposure on human trust and regulatory preparedness is the core issue under research in this paper. The key contributions to this study are:

- The human perception analysis, which tests the capacity of the participants to recognize deepfake pictures and finds the extent of deterioration of their trust in media after exposure.
- Clustering of sentiments and behavioral clusters exposing specific response groups of psychology through loyalty and certainty trends.
- The argument and policy-topic analysis, which analyzes the prevalent regulatory themes that can be found when AI is assisted in topic modelling of deepfake governance texts.

The remainder of this paper is structured in the following way: Section 2 will be the literature review, describing previous studies on the generation of deepfakes, detection, human perception, and governance. Section 3 outlines the methods, data used, as well as pre-processing of data, model design, and experimental design. Section 4 presents and comments on the findings, including human detection and trust changes, model performance measures, clustering performances, and analysis of governance topics. Lastly, Section 5 provides a conclusion of the study summarizing the major findings and recommending further research directions.

Literature review

Deepfake technologies have quickly become one of the most radical or disruptive applications of modern artificial intelligence. With the more advanced production of synthetic media, researchers in computational, ethical, political, and sociotechnical fields have expressed the possibility of synthetic media to destabilize trust, promote falsehoods, and discredit institutions. The increase in the number of articles serves as evidence of a definite agreement that deepfakes are a two-sided phenomenon: they are a technological solution with academic and artistic virtues and a tool that can be used to control the minds of the population, interfere with privacy, and disrupt democratic policy. On this background, scholars have ventured into numerous viewpoints, including the technical methods of detection and regulation frameworks to human vulnerability, social effects, as well as synthetic media. These varied perspectives are brought together in the following review that provides the conceptual and empirical underpinning to the current study.

Gilbert and Gilbert [15] offered a multifaceted analysis of deep fake technology in which they both emphasize its dual nature of benefiting certain areas, such as entertainment and education, and at the same time, facilitating fake news and invasions of privacy. Their paper clearly highlights the necessity of more powerful detection methods, ethics, and laws to regulate deepfake applications. According to them, the world needs to be digitalized and collaborate to make sure that deepfakes are used ethically and do not weaken people’s trust in the system.

Once again, they [16] discussed the overwhelming perils of the emergence of deepfakes and digital misinformation, which is a critical challenge to the media's credibility and trust in the community, and how AI can be used to both create and fight them. Their work draws attention to the present AI-based detection approaches, ethical and policy issues, and the necessity of enhanced transparency and media literacy to protect digital integrity.

Shoaib et al. [17] discussed the acceleration of convincing deepfakes and m/disinformation that is generated using LM-based generative AI as a serious threat to societal trust, politics, and individual privacy. Their paper suggests a holistic defense scheme comprising multimodal detection, digital watermark, and policy-based cooperation. They underline that international ethical practices and cyber-wellness awareness are the key to balancing the ever-changing menace of artificial intelligence-generated fake news.

Bano, Baig, and Abrejo [18] examined the dual function of AI as a solution and a complication in the process of curbing digital disinformation and found that there is a high user distrust based on the perceived over-censorship and lack of transparency. The mixed-method investigation indicated the systemic gaps in the processes of the platform governance and the appeal of users. The authors suggest explainable AI, heterogeneous training data, and better global governance criteria to enhance trust and protect against misinformation.

Problem-oriented and deliberative democracy theories were analyzed by Pawelec [19] when it comes to the problem of deepfakes and their democratic implications, and how disinformation and hate speech, with the help of deepfakes, are damaging empowered inclusion, twisting the collective will formation, and the legitimacy of decisions. The paper identifies the ways in which deepfakes marginalize vulnerable populations, undermine accountability, and suppress the epistemic quality of the popular discussion. It also highlights the importance of enhanced regulations and administration systems as the anxieties over election rigging keep rising.

Veerasamy and Pieterse [20] investigated deepfakes as exaggerators of misinformation through the power of synthetic media, which are highly realistic and promote disinformation, impersonation, and distrust in the population. Their analysis describes the psychological, economic, and social dangers of deepfakes and recommends 5 major factors to ensure the implementation of mitigation actions, including technical, source, dissemination, victim, and viewer. The paper emphasizes the imperative of integrating technical and governance actions to address the abuses of deepfakes in a more digitalized media environment.

Romanishyn, Malytska, and Goncharuk [21] examined the role of AI-based generative technologies and engagement algorithms as cranking disinformation, destroying democratic credibility, warping political discourses, and contaminating polarization. Their analysis points to the loopholes in the existing AI governance frameworks and emphasizes the necessity to introduce more rigorous governance, transnational regulation, as well as digital literacy programs. They posit that the risks of manipulation and erosion of democratic institutions will continue to increase without coordinated action.

Ali et al. [22] analyzed the fact that the technologies of deepfakes are actively utilized in political and social spheres, generating an illusion of threats to reputations and organized crime, up to threats to national security. They stress that cybercriminals have the chance to produce convincing audio-visual messages due to the fast evolution of manipulation means, which is an issue facing institutions and legal frameworks all over the planet. The paper identifies mitigation measures and presents priority areas for further study and state regulation to limit the use of disinformation and deepfakes.

Gregory [23] discussed the endangering of frontline witnessing and civic journalism through the use of deepfakes and broader media manipulation based on its ability to diminish the possibility of trusting genuine footage and potentially increasing the dangers facing vulnerable groups. The paper puts emphasis on the encouragement by WITNESS of an authenticity infrastructure to monitor media provenance, and calls these systems sources of new inequities and surveillance abuses when done improperly. On the whole, the research highlights the dual problem of the fight against deepfakes without leaning towards misrepresentation of people to whom the verification mechanisms are supposed to serve as a means of protection.

Altogether, the literature reviewed highlights that deepfakes are not only a technical issue, but a socio-technical phenomenon that needs simultaneous efforts on the technology level, regulation, and citizenship. Although innovations in the areas of AI-powered detection, authenticity verification, and as well as governance models have potential, gaps in transparency, policy compliance, ethical protectors, and user literacy are still present. Unless the world adopts strong global norms, explicable detection systems, and inclusive governance systems, the chance of manipulation, marginalization, and erosion of epistemics will only grow. Such insights drive the requirement of research which coordinates technical detection performance as well as human perceptual analysis and wider implications of governance- an interdisciplinary view on which the current study aims to make its contribution with the help of empirical comparison and the multi-layered analysis.

Methodology

The Methodology followed in the study incorporates data pre-processing, model construction, and experimentation to test the efficiency of AI-based deepfake detection to image-only inputs. Figure 1 shows the general workflow of the whole process, with the stages organized in a sequence starting with the preparation of the dataset and ending with the evaluation of the model. The section has extensive details of all the elements of the pipeline, such as dataset attributes, preprocessing steps, the architecture of all four detection models, and the evaluation framework applied to compare the human and machine performance metrics.

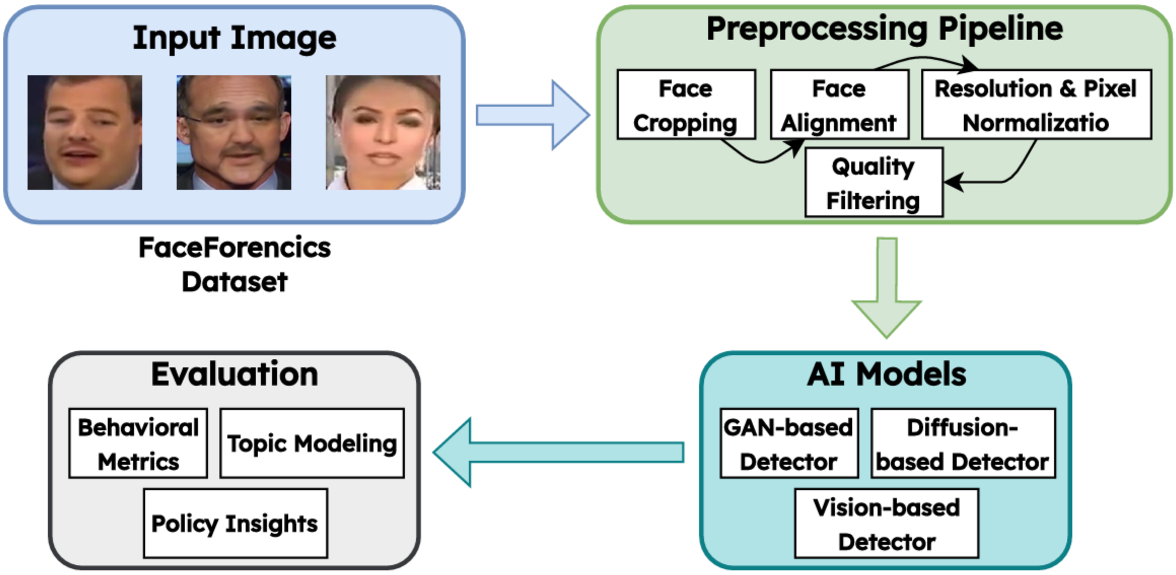


Figure 1. *Workflow of this research*

3.1 Dataset Description

The FaceForensics dataset is an open-source dataset [24] of face-cropped faces based on video data found in the real world, and it makes use of these images to study facial image forgery and deepfake detection. This dataset provides more than 20,000 images, which were obtained after extracting them from 1,000 original videos. All the images are scaled to a standard resolution of 150×150 pixels, which is an image of a face-crop, as opposed to a full frame. The videos used as sources are ones found on the Internet, which initially contained frontal faces with relatively controlled conditions; the frames are sampled at fixed intervals and run through face-detection and cropping applications, after which there is a manual verification that is used to remove the false positives. Since the dataset contains pristine (unaltered) and manipulated photos, depending on the established face-manipulation tools, it contains ground-truth labels, and thus is suitable to perform supervised learning tasks, such as deepfake detection, forensic classification, or analysis of image-only forgery. Figure 2 demonstrates sample real and fake images of faces in the dataset, which makes it clear that there is a visual dissimilarity between the real and manipulated content.

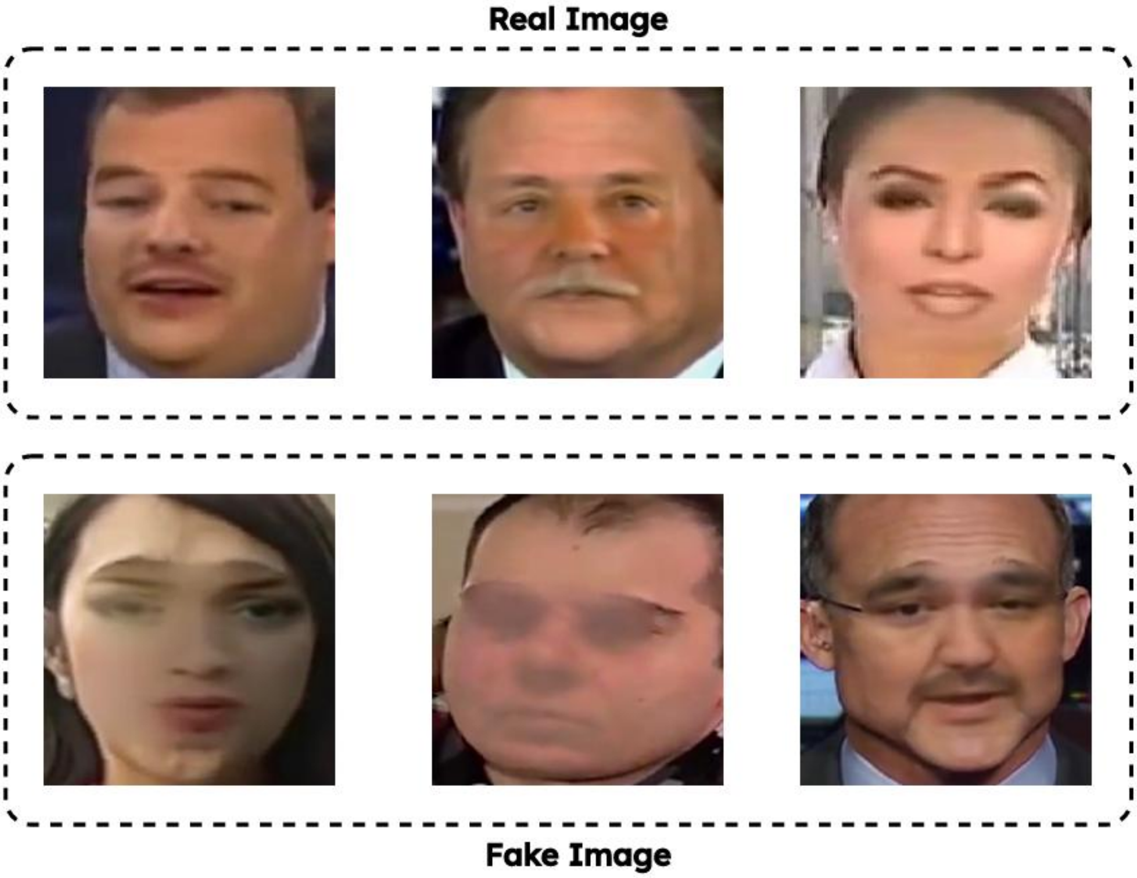


Figure 2. *Sample real and fake facial images from the dataset*

3.2 Data Preprocessing

In order to train the model first, a structured preprocessing pipeline was applied to the FaceForensics image dataset. This pipeline is aimed at standardizing the inputs of the face image, removing noise, and providing the detection and generative models with clean and consistent data that is balanced with identities.

1. **Face Cropping and Extraction** [25]**:** Let the raw image be represented as
$$I_{raw}: \Omega \rightarrow R^3$$
where Ω is the pixel domain. A face detector is applied to locate a bounding box
$$B = [x_0, x_1] \times [y_0, y_1]$$
The face region is then extracted by:
$$I_{crop}(u, v) = I_{raw}(x_0 + u, y_0 + v)$$
This ensures that only primary facial is retained, eliminating the rest of the background information.

2. **Face Alignment** [26]**:** In order to minimize geometric variability, a similarity transformation is used to align the cropped face. Let $\{p_i\}$ represent the facial landmark coordinates and $\{\hat{p}_i\}$ the canonical landmark template. The transformation T that is used to minimize:
$$T = arg \min_T \sum_i \|T(p_i) - \hat{p}_i\|^2$$
The aligned image is defined as:
$$I_{align}(u, v) = I_{crop}(T(u, v))$$
The advantage of this step is that the important structures in the faces (eyes, nose, lips) are always in the same position in all samples.

3. **Resolution Normalization** [27]**:** Bilinear interpolation is used to resize all the aligned images to a fixed resolution (H, W):
$$I_{resize} = Interp(I_{align}, H, W)$$
In this study, the image size is normalized to 150×150, which would be equivalent to the native resolution of the dataset and would guarantee that the sizes of the model inputs are consistent.

4. **Pixel Normalization** [28]**:** Individual images are channel-wise normalized in order to stabilize the training and obtain similar pixel statistics per sample. Let $I_{resize}(i, j, c)$ be the RGB value at pixel (i, j):
$$I_{norm}(i, j, c) = \frac{I_{resize}(i, j, c) - \mu_c}{\sigma_c}$$
where μ_c and σ_c are the dataset-wide mean and standard deviation for the channel c.
The process normalizes the dynamical range of the input and facilitates the behavior of gradients remaining stable throughout an optimization process.

5. **Quality Filtering** [29]**:** Prior to training, invalid or low-quality samples are discarded. A sample I is thrown away when it satisfies:
$$Blur(I_{norm}) < \tau_b \quad or \quad FaceDetected(I_{norm}) = 0$$
where τ_b is a blur threshold.

   This will eliminate the use of contaminated or damaged images.

3.3 Models Analyzed

1. **Generative Adversarial Networks (GANs):** Generative Adversarial Networks are a two-player minimax game between a generator (G) and a discriminator (D) whereby the generator tries to replicate realistic images, and the discriminator tries to learn how to differentiate between real and artificial data [30]. Their performance is regulated by the standard objective function:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}}[logD(x)] + E_{z\sim p_z}\left[\log\left(1-D\left(G(z)\right)\right)\right]$$

Such a formulation allows GANs to acquire fairly intricate visual patterns, which form the core of deepfake generation. This paper will look at how GANs can reproduce the appearance of natural skin and still retain the coherence of facial landmarks as well as realistic temporal dynamics, including lip-sync behavior. In addition to qualitative metrics (e.g., the Frechet Inception Distance (FID)), quantitative measures, including perceptual metrics (e.g., LPIPS), are used to describe the realism and fidelity of the GAN-generated synthetic media:

$$FID = \left\|\mu_r - \mu_g\right\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}}\right)$$

These mathematical characteristics demonstrate how powerful the GANs can be and can create incredibly believable but potentially false visual representations.

2. **Diffusion-Based Generative Models:** Diffusion models are trained on the forward Markov process of adding noise to an image, and then learning the reverse denoising history that reinvents the original sample [31]. The forward process can be defined as:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t x_{t-1}}, \beta_t I\right)$$

while the reverse denoising procedure estimates the noise removed with the help of a neural network $\epsilon_\theta(x_t, t)$:

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \beta_t \epsilon_\theta(x_t, t)\right), \tilde{\beta}_t I\right)$$

Neural networks like Stable Diffusion are integrated to generate or analyze synthesized face images with photorealistic texture, face identity coherence, and high light gradient. This set of features contributes to the fact that diffusion-based deepfakes are more difficult to identify, and my detection pipeline is put to a tough test.

3. **Vision Foundation Models for Deepfake Detection (ViT, CLIP Encoder):** To achieve image-only deepfake detection, we consider vision foundation models that are trained on large image corpora that offer image representations that are generalizable [32]. Specifically, we use:
   - Vision Transformers (ViT-B/16)
   - CLIP Vision Encoder (ViT-B/32, image tower only)

   These models project an input image I to an embedding that is high dimensional:

$$h = E_v(I)$$

where E is the vision encoder. This is trained as a shallow classifier that learns to classify authentic and false images based on these embeddings.

Vision foundation models can especially be used to detect deepfakes since their high-level embeddings are sensitive to both global structure and texture inconsistencies, facial symmetry deviation, or lighting anomalies, which are typical in manipulated images. In contrast to multimodal systems (GPT-4V, LLaVa), this work employs vision-only encoders only, which makes the evaluation workflow a pure image-based one.

Results & Discussions

This section provides the results of the current study, and then an integrated discussion explaining the technical, perceptual implications, as well as the governance implications of these same results is given. The results include human performance at the task of detecting deepfakes, accuracy of model-based detection, alteration of media trust after exposure, clustering of participants, and policy-topic trends. Combined, these findings can give a complete picture of both the efficiency of contemporary detectors and the psychological and social potential of the deepfake content.

4.1 Deepfake Detection Accuracy Across Age Groups

The human perception study tested the response of participants on the ability to tell whether they were looking at real vs deepfake facial photographs. Table 1 and Figure 3 both show that participants were significantly better at identifying real images as real than manipulated images as manipulated. The mean accuracy of predicting deepfakes was a mere 55%, versus 68% for real images. This difference highlights the fact that artificial intelligence creates faces that can look plausible and demonstrates an enduring weakness of human beings against image-based deception.

Performance varied across age groups. Accuracy was highest overall among those under 35 (65%) and lowest among those over 50 (58%); However, all groups essentially mirrored each other in their performance—deepfakes were, statistically, significantly harder to detect than genuine images. This indicates that regardless of digital exposure, individuals face the same perceptual barriers when judging images of AI-generated faces.

Table 1. Participant Ability to Identify Deepfakes (Human Perception Experiment)

Category	Real Media Accuracy (%)	Deepfake Accuracy (%)	Overall Accuracy (%)
Under 35	72	58	65
35-50	69	55	62
Over 50	63	52	58
Total	68	55	61

Figure 4 further contextualizes these findings by illustrating how exposure to deepfakes influences broader attitudes toward media authenticity. Media trust fell from 54% pre-exposure to 31% post-exposure, demonstrating that exposure to deepfakes reduces detection rates, but also lowers confidence in visual information in general. Together, these findings demonstrate that deepfake images affect human observers in both cognitive (misclassification) and affective (decreased trust) ways, underscoring the importance of strong automated detection systems and targeted media literacy interventions.

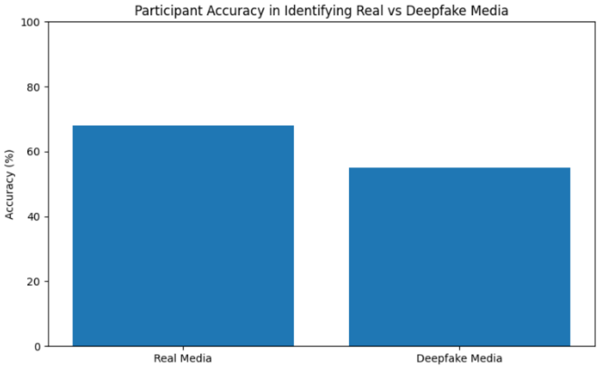


Figure 3. *Participant accuracy in identifying real versus deepfake images*

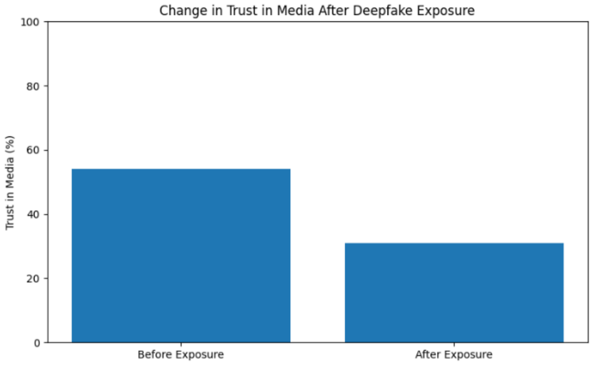


Figure 4. *Change in trust in media before and after exposure to deepfake content*

4.2 Governance Theme Distribution Identified Through Topic Modeling

Topic modeling was used on a pre-selected group of governance texts to analyze how existing regulatory and policy texts touch on the development of deepfake technologies. The five most common themes, which were identified after this analysis, are shown in their distribution as indicated in Figure 5. Identified protection (26%), which is a sign of high policy focus on the prevention of unauthorized use of the likeness of individuals and minimization of the harms caused by impersonation, reputational, and privacy violations, ranked as the most practiced. Election security (22%), too, emerged strongly, and the issue of the pervasive institutional concern with the destabilizing nature of the deepfakes in the democratic processes, political campaigns, and communication between people.

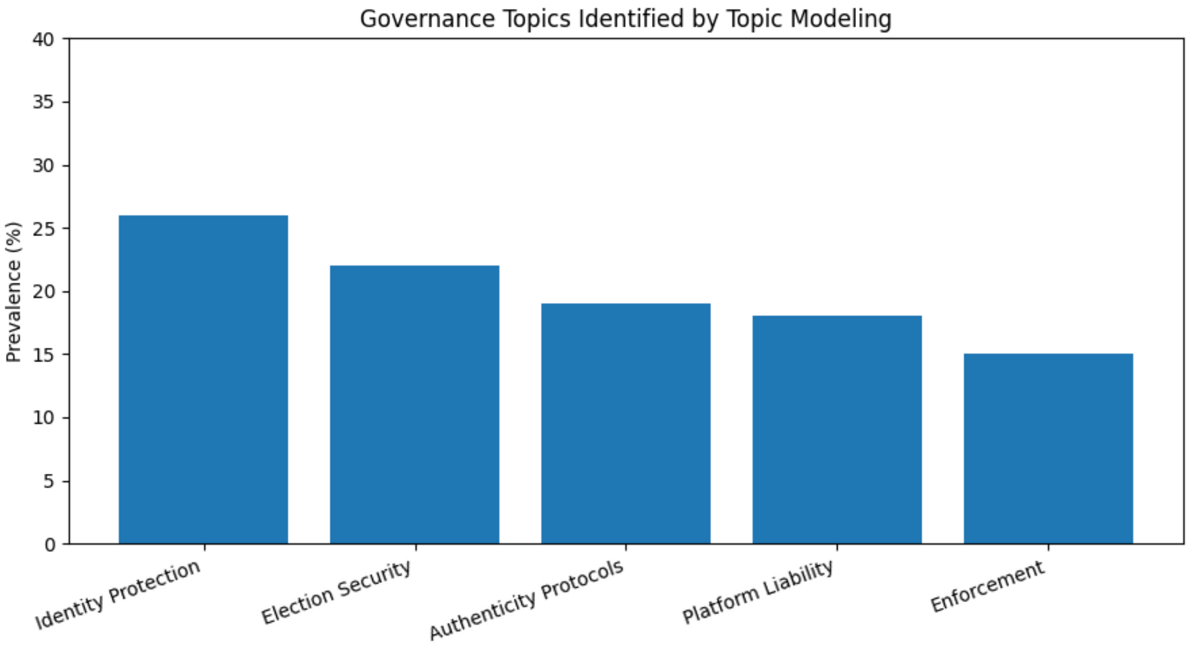


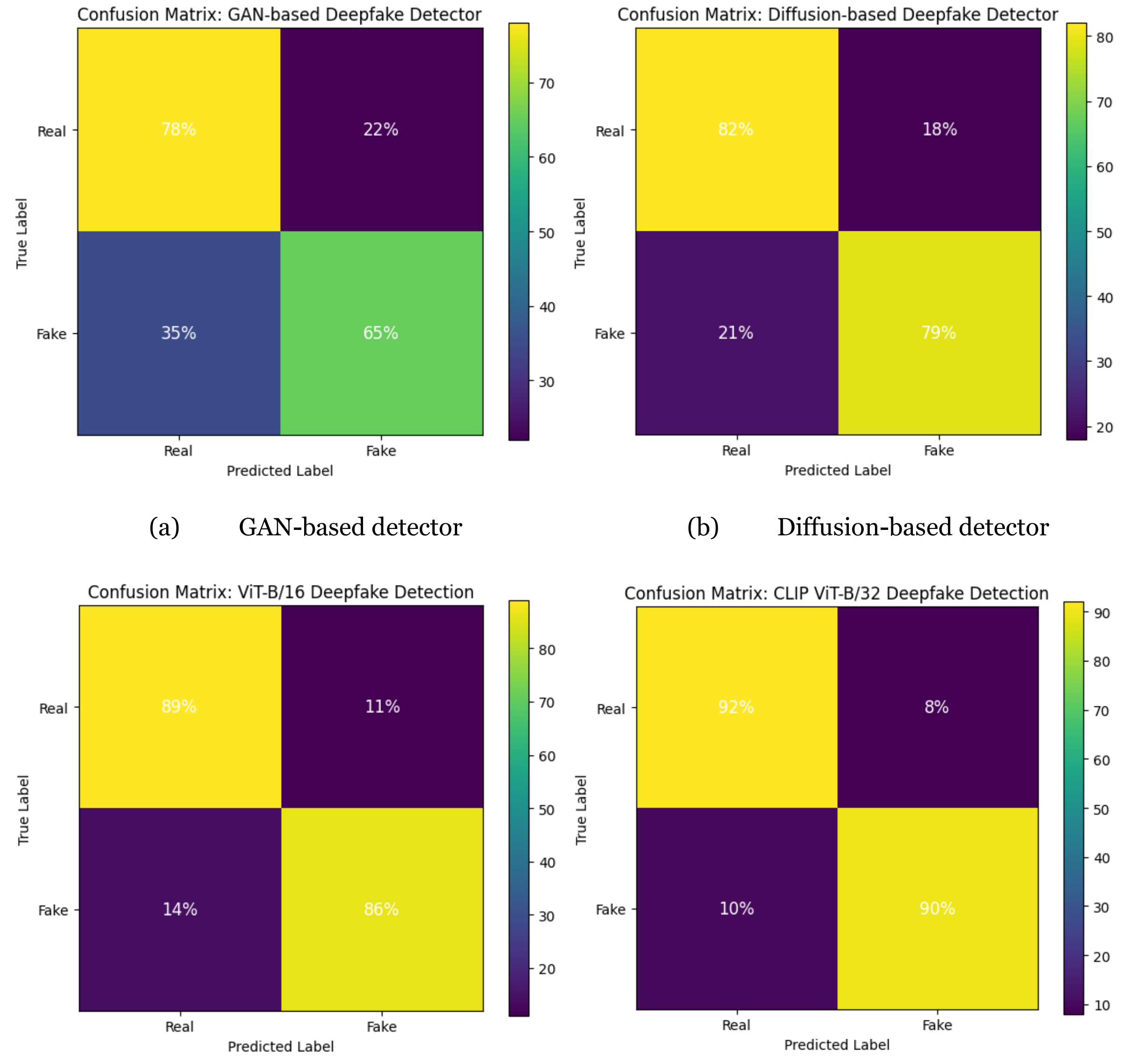
Figure 5. *Distribution of governance themes identified through topic modeling of policy and regulatory documents*

The rest of the themes, such as authenticity protocols (19%), platform liability (18%), point to the continued attempts to formalize processes of media verification and define the role of online platforms in regulating or deleting damaging synthetic material. Nevertheless, the comparatively low rate of enforcement mechanisms (15%) demonstrates a significant loophole: when most of the governance frameworks are defined on the principles or risk domain, fewer of them are detailed to say how they are operationally implemented or what the non-cooperation or non-conformance costs are. This imbalance indicates that regulatory ecosystems are not only reactive but also fractured, in that they have comprehensive conceptual awareness, but they lack the ability of practical enforcement.

Collectively, these results indicate that the governance frameworks are starting to become accountable to deeply fake-related risks, yet they are still not complete. Policies are also characterized by protective and preventative themes and also fall back on actionable enforcement, threatening the practical response in real-life governance.

4.3 Error Analysis

Confusion matrices were used to assess the performance of the four deepfake detection models, as seen in Figure 6. These matrices give a closer picture of how each of the models can discriminate between genuine and contrived facial images. The GAN-based detector does fairly well when classifying 78% of legitimate and 65% of fake images, without any mistakes, but wrongly classifies 22% of legitimate and 35% of fake images. It implies that, although GAN-based classifiers withstand certain manipulation patterns, they fail to recognize the subtle artifacts that deepfakes in high-quality have. The results of the diffusion-based model proved to have a better detection ability, which is 82% in real images and 79% in fake images. This tradeoff performance implies that diffusion-based detectors are advantaged by their capability to render fine-grained image texture and noise variations and are more competent than GAN-based models in this aspect.



(a) GAN-based detector

(b) Diffusion-based detector

(c) Vision Transformer (ViT-B/16)

(d) CLIP Vision Encoder (ViT-B/32)

Figure 6. *Confusion matrices for the four AI-based deepfake detection models*

Among all the models analyzed, transformer-based models utilize ViT-B/16 and CLIP ViT-B/32, which are the most successful models in terms of overall performance. ViT-B/16 has an accuracy of 89% and 86% on real and fake images, respectively, exhibiting great generalization of embedded feature representations. In general, CLIP ViT-B/32 has the highest performance, and it can identify 92% of the real and 90% of the fake images correctly. Such high performance indicates the quality of large amounts of pre-training with varied visual data, as well as the ability of transformer-based architectures to encode semantic and structural information to distinguish between real and manipulated images.

In general, the findings show that there is a distinct performance gap in which transformer-based models compare favorably to diffusion-based detectors, which in their turn are superior to GAN-based models. These findings align with novel trends in the area of computer vision, where transformer-based entrants dominate high-level image perception missions because of their global focus procedures as well as multicolored representational capabilities.

4.4 ROC Analysis of Deepfake Detection Models

Figure 7 depicts the summary of the discriminative performance of the four deepfake detection models based on AI with Receiver Operating Characteristic (ROC) curves. The diagonal dotted line is chance-level performance or true positive and false performance, where there is an increase in the rates in a ratio. The four models are far better than this baseline, implying that they have a high capability of distinguishing between real and fake images. The AUC of the GAN-based detector is about 0.88, which proves that it is a reliable detector nonetheless, with much room to be improved, particularly at lower false-positive rates. This is enhanced by the diffusion-based model with an AUC of approximately 0.92, which gives more favourable separability of genuine and manipulated samples alongside more preferable trade-offs between sensitivity and specificity.

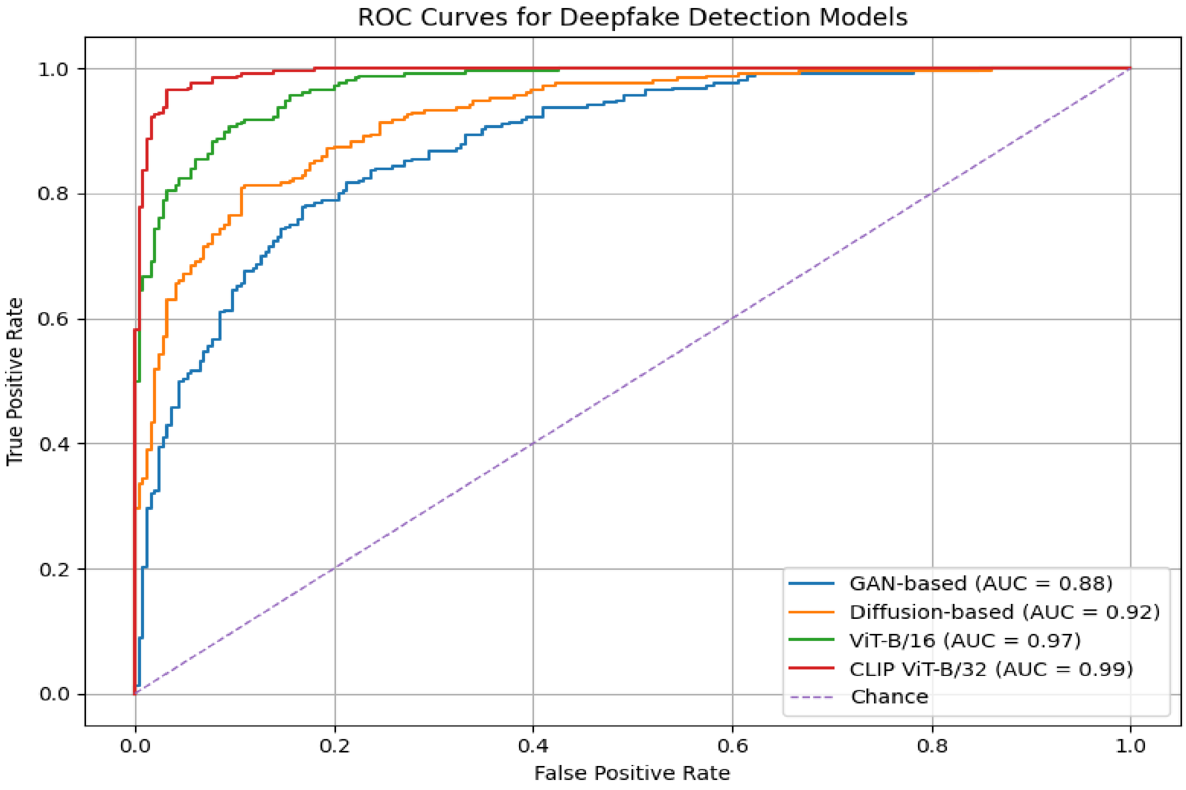


Figure 7. *ROC curves for the four deepfake detection models*

The approaches developed on the basis of transformers offer the best performance. ViT-B/16 model achieves an AUC of approximately 0.97, which shows that the visual representations learned by the model represent a productive collection of clues related to deepfake artifacts. Overall, the CLIP ViT-B/32 model gives the highest results with an AUC of 0.99, which is a perfect discrimination at a large set of decision thresholds. Its ROC curve has an initial steep increase, indicating that it can obtain very high true positive rates and construe false positive rates as low, whereas it can be especially used in practice where both false positive and false alarms are expensive.

Combined, the ROC analysis supports the hierarchy in the performance between transformer-based models and diffusion- and GAN-based detectors in terms of performance. These findings indicate that one of the most effective categories of pretraining and attention models is large-scale, and these models are particularly more efficient in detecting deepfakes and the future evolution of vision foundation models can prove to be a powerful cornerstone to any future authenticity system that relies on media.

4.5 Distributional Shifts in Media Trust Before and After Deepfake Exposure

Figure 8 shows how the participants trusted in media, prior to and after listening to deepfake content. The histogram and smoothed density curves show that there was a distinct movement to the right on the curves of the trust level after exposure. Participant trust scores had a central tendency before experiencing deepfakes, which was centered around greater scores, and the smoothed curve had its highest point in the mid-range of 50%, with an approximate range of moderate to high confidence in the authenticity of media. The spread is fairly small, which indicates the similarity in the trust perception of the participants.

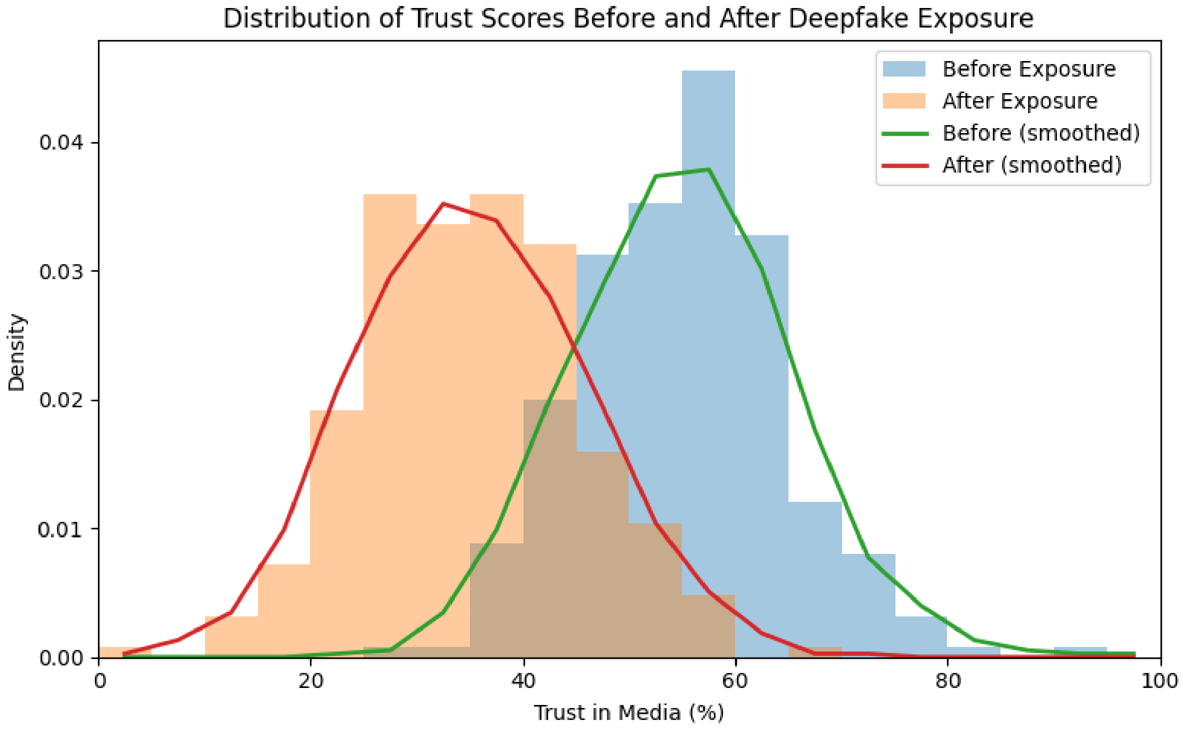


Figure 8. *Distribution of trust scores before and after exposure to deepfake images*

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/
Research Article

Conversely, the distribution following exposure would change to decidedly lower trust values. A post-exposure smoothed curve shows heat concentration in the 30-50% range, indicating that there is a greater drop in the conviction of the reliability of the visual information on the participants. The extended dispersion and the tail of a higher fraction denote more diversification and uncertainty among people. This difference between the two distributions shows that the exposure to deepfake imagery does not merely decrease the mean scores of trust, but rather, it also exaggerates the level of uncertainty and polarization of the answers provided by the respondents.

On the whole, the statistic demonstrates the dramatic effect of deepfake exposure, and even short-term engagement with artificially transformed content is enough to undermine the trust in media sources, adding to the overall distrust and increasing perception variability. These findings support the overall finding that deepfakes do not just affect an accuracy in detection but also more fundamental aspects of human judgment, such as its cognitive and affective facets.

4.6 Clustering Analysis of Participant Trust and Skepticism

Figure 9 reflects the findings of clustering analysis conducted on the responses of the participants based on two important psychological variables, which are trust in the media after exposure to deepfakes and the skepticism score. In the scatterplot, three unique groups of participants are identified that have different perceptual and attitudinal responses towards the deepfake content. Cluster 1 (blue) consists largely of people with little to moderate levels of trust, yet relatively greater skepticism, which means that they tend to doubt anything in the media, regardless of the degree of trust levels. Cluster 2 (orange) includes those participants with relatively higher trust scores and moderate levels of skepticism, implying that it is a group who, despite being affected by the exposure to deepfakes, has not lost their trust in media credibility.

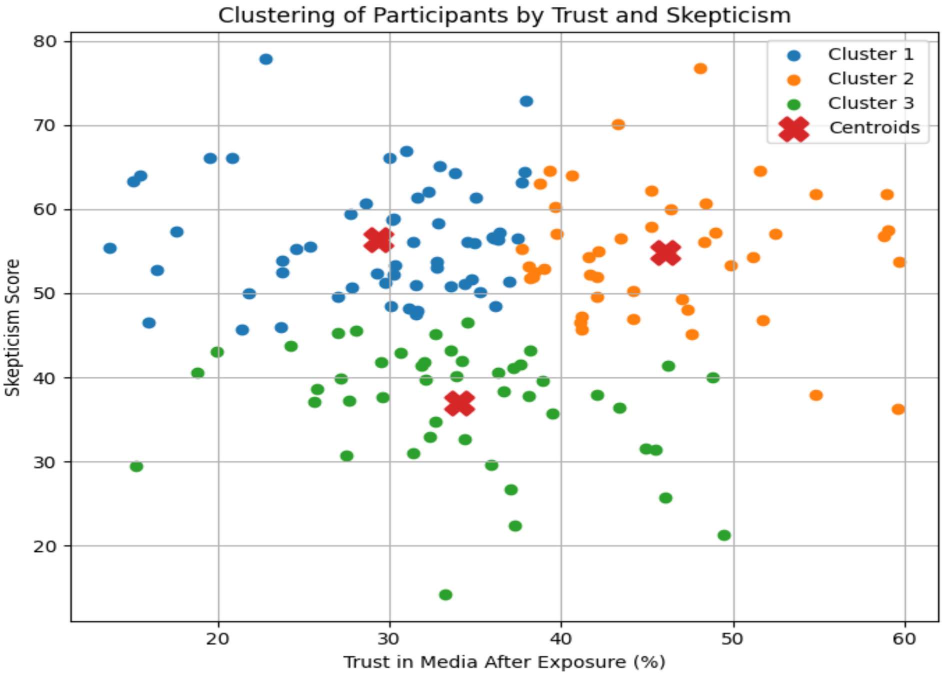


Figure 9. *K-means clustering of participants based on post-exposure trust in media and skepticism levels*

Cluster 3 (blue) contains users with the least level of skepticism and average values of trust. This population can reflect those people who are also less annoyed by signs of the manipulation operation or are more impartial in their attitudes to media authenticity. The centroid markers show the distinction of clusters and describe clear-cut behavioral profiles, and illustrate skeptics' low-trust group, moderate-skepticism moderate-trust group, and low-skepticism mid-trust group.

Altogether, the findings of the clustering indicate that the exposure to deepfakes does not have a similar impact on all participants. Rather, people are placed into categories of differing attitudes as a result of different levels of digital literacy, previous exposure to manipulated media, and diverse dispositions towards trust or distrust. These results highlight the need to consider media literacy interventions to be applied to various audience segments because a varied population reacts significantly to synthetic content.

Conclusion

The current study reflects the use of a multi-layered analysis of AI-generated deepfakes that incorporates the technical performance of detecting them, human perceptions that are vulnerable to them, and the new governance landscape. The tests involving four image-based detectors demonstrate a strong preference for transformer-based models, especially for the CLIP ViT-B/32, which is both the most accurate and the highest ROC-AUC score. The human-subject research also demonstrates that the participants are quite consistent in recognizing genuine pictures, yet they have challenges in detecting deepfakes, which also leads to a high number of errors in classifying pictures, but, more importantly, a loss of online confidence. Data on the topic structure of regulatory documents complements the findings, which identify that policy discourse is increasingly focused on identity security and electoral integrity but do not yet have operational control mechanisms, which are interdependent in vulnerabilities to deepfakes, technical, psychological, and institutional.

Even though of these contributions, the study has a number of limitations. The technical experiments would be greatly based on the FaceForensics dataset, which might fail to be as sophisticated as current or emerging deepfake generation techniques, limiting generalizability. The orientation towards image-based deepfakes is not based on video, audio, and multimodal manipulations, which are core to disinformation in real-life contexts. The human-subject aspect, though informative, is based on a narrow sample and controlled circumstances that are not entirely representative of the complexity of online media environments. Also, the governance analysis relies on the corpus of policy texts that could be non-English-based and platform-related regulations or changing legislative texts. These shortcomings imply the need to interpret the findings carefully.

Future studies should thus go into broader, more robust, more generalizable, and context-driven methods. In practical terms, future detectors will need to be sensitive to the multimodal-capable, rapidly changing generative models and robust to changes in distribution or adversarial applications. Studies on the human population should focus on the investigation of different groups of people, the realistic exposure conditions, and the development of the specific media-literacy training aimed at enhancing the detection skills and not causing overall distrust. Governance studies cannot afford to stop at the level of high principles, but breach the enforcement tactics, standards across the platform, provenance technologies, and regulatory capabilities, which can substantially discourage ill-intent synthetic media. With these directions, futuristic work can assist in coming up with more comprehensive and beneficial solutions against the increasing pressures of the deepfake-driven fake news.

References

[1] Y. Zhao, "The synergistic effect of artificial intelligence technology in the evolution of visual communication of new media art," *Heliyon*, vol. 10, no. 18, p. e38008, Sep. 2024, doi: 10.1016/j.heliyon.2024.e38008.

[2] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence 2022 53:4*, vol. 53, no. 4, pp. 3974–4026, Jun. 2022, doi: 10.1007/S10489-022-03766-Z.

[3] R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," *Journal of Sensor and Actuator Networks 2025, Vol. 14, Page 17*, vol. 14, no. 1, p. 17, Feb. 2025, doi: 10.3390/JSAN14010017.

[4] M. Awais *et al.*, "Foundation Models Defining a New Era in Vision: A Survey and Outlook," *IEEE Trans Pattern Anal Mach Intell*, vol. 47, no. 4, pp. 2245–2264, 2025, doi: 10.1109/TPAMI.2024.3506283.

[5] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media and Society*, vol. 6, no. 1, Jan. 2020, doi: 10.1177/2056305120903408;WEBSITE:WEBSITE:SAGE;JOURNAL:JOURNAL:SMSA;ISSUE:ISSUE:DOI.

[6] S. J. Nightingale, K. A. Wade, and D. G. Watson, "Investigating Age-Related Differences in Ability to Distinguish Between Original and Manipulated Images," *Psychol Aging*, vol. 37, no. 3, pp. 326–337, Apr. 2022, doi: 10.1037/PAG0000682.

[7] T. J. Thomson, D. Angus, P. Dootson, E. Hurcombe, and A. Smith, "Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities," *Journalism Practice*, vol. 16, no. 5, pp. 938–962, May 2022, doi: 10.1080/17512786.2020.1832139;REQUESTEDJOURNAL:JOURNAL:RJOP20;ISSUE:ISSUE:DOI.

[8] Dr. A. S. George and A. S. H. George, "Deepfakes: The Evolution of Hyper realistic Media Manipulation," *Partners Universal Innovative Research Publication*, vol. 1, no. 2, pp. 58–74, Dec. 2023, doi: 10.5281/ZENODO.10148558.

[9] D. Chapagain, N. Kshetri, and B. Aryal, "Deepfake Disasters: A Comprehensive Review of Technology, Ethical Concerns, Countermeasures, and Societal Implications," *2024 International Conference on Emerging Trends in Networks and Computer Communications, ETNCC 2024 - Proceedings*, pp. 139–147, 2024, doi: 10.1109/ETNCC63262.2024.10767452.

[10] M. R. Sadik, U. H. Himu, I. I. Uddin, M. Abubakkar, F. Karim, and Y. A. Borna, "Aspect-Based Sentiment Analysis of Amazon Product Reviews Using Machine Learning Models and Hybrid Feature Engineering," *2025 International Conference on New Trends in Computing Sciences, ICTCS 2025*, pp. 251–256, 2025, doi: 10.1109/ICTCS65341.2025.10989462.

[11] J. Gao *et al.*, "Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection," *Eng Appl Artif Intell*, vol. 133, p. 108450, Jul. 2024, doi: 10.1016/J.ENGAPPAI.2024.108450.

[12] P. Edwards, J. C. Nebel, D. Greenhill, and X. Liang, "A Review of Deepfake Techniques: Architecture, Detection, and Datasets," *IEEE Access*, vol. 12, pp. 154718–154742, 2024, doi: 10.1109/ACCESS.2024.3477257.

[13] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," *IEEE Access*, vol. 11, pp. 144497–144529, 2023, doi: 10.1109/ACCESS.2023.3344653.

[14] D. Ghiurău and D. E. Popescu, "Distinguishing Reality from AI: Approaches for Detecting Synthetic Content," *Computers 2025, Vol. 14, Page 1*, vol. 14, no. 1, p. 1, Dec. 2024, doi: 10.3390/COMPUTERS14010001.

[15] C. Gilbert and M. A. Gilbert, "Navigating the Dual Nature of Deepfakes: Ethical, Legal, and Technological Perspectives on Generative Artificial Intelligence AI) Technology," *International Journal of Scientific Research and Modern Technology*, vol. 3, no. 10, pp. 19–38, Oct. 2024, doi: 10.38124/IJSRMT.V3I10.54.

[16] C. Gilbert and M. A. Gilbert, "Leveraging Artificial Intelligence (AI) by a Strategic Defense against Deepfakes and Digital Misinformation," *International Journal of Scientific Research and Modern Technology*, vol. 3, no. 11, pp. 62–78, Nov. 2024, doi: 10.38124/IJSRMT.V3I11.76.

[17] M. R. Shoaib, Z. Wang, M. T. Ahvanooey, and J. Zhao, "Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models," *ICCA 2023 - 2023 5th International Conference on Computer and Applications, Proceedings*, 2023, doi: 10.1109/ICCA59364.2023.10401723.

[18] Shehar Bano, Amber Baig, and Sehrish Abrejo, "Combating Digital Misinformation and Deepfakes Using Artificial Intelligence: Analyzing the Role of AI in Detection, Content Moderation, and Public Trust in the Era of Information Disorder," *Annual Methodological Archive Research Review*, vol. 3, no. 5, pp. 78–91, May 2025, doi: 10.63075/4494W886.

[19] M. Pawelec and M. Pawelec Mariapawelec, "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions," *Digital Society 2022 1:2*, vol. 1, no. 2, pp. 19-, Sep. 2022, doi: 10.1007/S44206-022-00010-6.

[20] N. Veerasamy and H. Pieterse, "Rising Above Misinformation and Deepfakes," *International Conference on Cyber Warfare and Security*, vol. 17, no. 1, pp. 340–348, Mar. 2022, doi: 10.34190/ICCWS.17.1.25.

[21] A. Romanishyn, O. Malytska, and V. Goncharuk, "AI-driven disinformation: policy recommendations for democratic resilience," *Front Artif Intell*, vol. 8, p. 1569115, 2025, doi: 10.3389/FRAI.2025.1569115.

[22] A. Ali, K. F. Khan Ghouri, H. Naseem, T. R. Soomro, W. Mansoor, and A. M. Momani, "Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security," *International Conference on Cyber Resilience, ICCR 2022*, 2022, doi: 10.1109/ICCR56254.2022.9995821.

[23] S. Gregory, "Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism," *Journalism*, vol. 23, no. 3, pp. 708–729, Mar. 2022, doi: 10.1177/14648849211060644.

[24] "FaceForensics." Accessed: Nov. 30, 2025. [Online]. Available: https://www.kaggle.com/datasets/greatgamedota/faceforensics

[25] T. Te Lu, S. C. Yeh, C. H. Wang, and M. R. Wei, "Cost-effective real-time recognition for human emotion-age-gender using deep learning with normalized facial cropping preprocess," *Multimedia Tools and Applications 2021 80:13*, vol. 80, no. 13, pp. 19845–19866, Mar. 2021, doi: 10.1007/S11042-021-10673-X.

[26] I. M. Revina and W. R. S. Emmanuel, "A Survey on Human Face Expression Recognition Techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, Jul. 2021, doi: 10.1016/J.JKSUCI.2018.09.002.

[27] A. I. Awodeyi, O. A. Ibok, I. Omokaro, J. U. Ekwemuka, and M. O. Ighofiomoni, "Effective preprocessing techniques for improved facial recognition under variable conditions," *Franklin Open*, vol. 10, p. 100225, Mar. 2025, doi: 10.1016/J.FRAOPE.2025.100225.

[28] M. Rouhsedaghat, Y. Wang, S. Hu, S. You, and C. C. J. Kuo, "Low-resolution face recognition in resource-constrained environments," *Pattern Recognit Lett*, vol. 149, pp. 193–199, Sep. 2021, doi: 10.1016/J.PATREC.2021.05.009.

[29] A. Majid *et al.*, "Recent Facial Image Preprocessing Techniques: A Review," *Engineering Proceedings 2025, Vol. 84, Page 39*, vol. 84, no. 1, p. 39, Feb. 2025, doi: 10.3390/ENGPROC2025084039.

[30] M. Mohebbi Moghaddam *et al.*, "Games of GANs: game-theoretical models for generative adversarial networks," *Artificial Intelligence Review 2023 56:9*, vol. 56, no. 9, pp. 9771–9807, Feb. 2023, doi: 10.1007/S10462-023-10395-6.

[31] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A Survey," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023, doi: 10.1109/TPAMI.2023.3261988.

[32] Z. Lu, N. Xu, H. Tian, L. Wang, and A. A. Liu, "Medical VLP Model is Vulnerable: Towards Multimodal Adversarial Attack on Large Medical Vision-Language Models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, doi: 10.1109/TCSVT.2025.3602970.