UNIVERSITETI POLIS

ICCSM 2025

BOOK OF PROCEEDINGS

1ST INTERNATIONAL CONFERENCE
COMPUTER SCIENCES AND MANAGEMENT

WHERE DIGITAL & BUSINESS BECOME HUMAN

26-27 June 2025 | Tirana, Albania

ISBN 9789928347123

DOI 10.37199/c41000300

CONFERENCE CHAIR

Assoc. Prof. Merita Toska, POLIS University

PARTNER UNIVERSITIES

POLIS University, Albania
Université Lyon 2, France
Università Telematica San Raffaele, Italy
University of Western Macedonia, Greece
International University of Sarajevo, Bosnia & Herzegovina
Mother Teresa University, North Macedonia
Gebze Technical University, Turkey
Public International Business College, Kosovo
Rochester Institute of Technology – RIT Global Campus, Kosovo
Co-PLAN, Institute for Habitat Development, Albania
AI Hub Albania, Albania
Luralux, Albania

ORGANISING COMMITTEE

Dr. Blerta Mjeda
Dr. Emiliano Mankolli
Msc. Sonila Murataj
Msc. Andia Vllamasi
Msc. Klejda Hallaci
Msc. Erilda Muka
Msc. Armela Reka

SCIENTIFIC COMMITTEE

Prof. Dr. Jérôme Darmont, Université Lumière Lyon 2 (France)
Prof. Dr. Lydia Coudroy de Lille, Université Lumière Lyon 2 (France)
Prof. Dr. Jim Walker, Université Lumière Lyon 2 (France)
Prof. Dr. Besnik Aliaj, POLIS University, (Albania)
Prof. Dr. Daniela Sica, San Raffaele Roma University, (Italy)
Prof. Dr. Stefania Supino, San Raffaele Roma University, (Italy)
Prof. Dr. Arbana Kadriu, South East European University (North Macedonia)
Prof. Dr. Ing. Lejla Abazi-Bexheti, South East European University (North Macedonia)
Prof. Dr. Yusuf Sinan Akgül, Gebze Technical University (Turkey)
Assoc. Prof. Dr. Galia Marinova, Technical University of Sofia (Bulgaria)
Assoc. Prof. Dr. Vasil Guliashki, Technical University of Sofia (Bulgaria)
Assoc. Prof. Mehmet Göktürk, Gebze Technical University (Turkey)
Assoc. Prof. Yakup Genç, Gebze Technical University (Turkey)
Assoc. Prof. Habil Kalkan, Gebze Technical University (Turkey)
Assoc. Prof. Dr. Godiva Rëmbeci, POLIS University (Albania)
Assoc. Prof. Dr. Xhimi Hysa, POLIS University (Albania)
Assoc. Prof. Dr. Merita Toska, POLIS University (Albania)
Assoc. Prof. Dr. Sotir Dhamo, POLIS University (Albania)
Dr. Gennaro Maione, San Raffaele Roma University, (Italy)
Dr. Nicola Capolupo, San Raffaele Roma University, (Italy)
Dr. Benedetta Esposito, San Raffaele Roma University, (Italy)
Dr. Venera Demukaj, Rochester Institute of Technology (Kosovo)
Dr. Emil Knezović, International University of Sarajevo (BiH)
Dr. Šejma Aydin, International University of Sarajevo (BiH)
Dr. Azra Bičo, International University of Sarajevo (BiH)
Dr. Šejma Aydin, International University of Sarajevo (BiH)
Dr. Azra Bičo, International University of Sarajevo (BiH)
Dr. Hamza Smajić, International University of Sarajevo (BiH)
Dr. Panagiotis Kyratsis, University of Western Macedonia (Greece)
Dr. Delina Ibrahimaj, Minister of State for Entrepreneurship and Business Climate (Albania)
Dr. Elona Karafili, POLIS University (Albania)
Dr. Emi Hoxholli, POLIS University (Albania)
Dr. Shefqet Suparaku, POLIS University (Albania)
Dr. Manjola Hoxha, POLIS University (Albania)
Dr. Elsa Toska, POLIS University (Albania)
Dr. Emiliano Mankolli, POLIS University (Albania)

Dr. Albina Toçilla, POLIS University (Albania)
Dr. Sonia Jojic, POLIS University (Albania)
Dr. Ilda Rusi, POLIS University (Albania)
Dr. Ledian Bregasi, POLIS University (Albania)
Dr. Klodjan Xhexhi, POLIS University (Albania)
Dr. Endri Duro, POLIS University (Albania)
Dr. Remijon Pronja, POLIS University (Albania)
Dr. Vjosë Latifi, International Business Collage Mitrovica (Kosovo)
Dr. Agron Hajdari, International Business Collage Mitrovica (Kosovo)

Table of Contents

16

SAFEGUARDING DIGITAL AUTHENTICITY AND WOMEN'S IDENTITY THROUGH DEEPFAKE DETECTION

DOI: 10.37199/c41000316

Livia IBRANJ
POLIS University (Tirana, Albania)
livia_ibranj@universitetipolis.edu.al

Abstract

Deepfake technology, the algorithmic manipulation of images and videos, is renowned for its ability to create highly realistic and stimulating content. They leverage deep learning and train generative neural architectures to map voices and faces onto another person's body. Using Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) enables accurate manipulation of voices, facial features, and expressions to create images and videos that closely resemble real people. While this may appear as a technological achievement, deepfakes have enabled profound harms, including identity theft, harassment, and non-consensual explicit imagery, with women comprising 96% of victims.

This paper explores multimodal detection approaches that combine deep learning features with forensic analysis to differentiate AI-generated images from authentic photographs. Our methodology integrates complementary detection strategies: deep semantic features via EfficientNet-B3 and CLIP models (2,304 dimensions), frequency-domain analysis detecting spectral anomalies, noise residual statistics, Local Binary Pattern texture descriptors, and facial forensics—totalling 2,339 features per image. An ensemble classifier combining Gradient Boosting and Logistic Regression was trained on 200 images (100 authentic photographs, 100 AI-generated from Midjourney, Stable Diffusion, and DALL-E), achieving 85% accuracy with 98.25% ROC-AUC. Performance analysis reveals asymmetric characteristics: 95% recall for authentic images versus 75% recall for AI-generated content, while maintaining 93.75% precision on synthetic detection. The 25% false negative rate underscores that technical detection alone cannot solve deepfake abuse—comprehensive protection requires platform accountability, legislative frameworks, and victim support systems. This study contextualises technical findings within the social crisis of digital sexual violence, examines documented psychological impacts on victims, identifies critical legal gaps, and outlines future research directions, including larger datasets, temporal analysis, and hybrid human-AI detection systems.

Keywords: Deepfake Detection, Generative Adversarial Networks, Artificial Intelligence (AI), Multimodal Feature Extraction

I. INTRODUCTION

The rapid development of artificial intelligence has ushered in a new era of synthetic media. Deepfake technology, which uses algorithms to produce incredibly lifelike but fake images and videos, is a prime example of this. Leveraging deep learning architectures such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), deepfakes manipulate voices, facial features, and expressions to produce content that is often indistinguishable from authentic media. Although these technological advancements hold many exciting potential benefits, there are also significant concerns about misinformation, privacy breaches, and digital security.

The consequences of deepfake abuse are not distributed equally across society. Research indicates that 96% of deepfakes are non-consensual explicit content, with the major target being women (Sensity, 2023). High-profile cases, such as the 2018 targeting of journalist Rana Ayyub and the mass victimisation of South Korean women in 2024, demonstrate how this technology amplifies existing patterns of gender-based violence and harassment. The psychological impact on victims mirrors that of revenge pornography, including trauma, anxiety, depression, and social withdrawal, even though victims never consented to the original imagery.

This paper investigates methods for distinguishing AI-generated images from real photographs using pixel-level analysis, feature extraction, and machine learning classification. By focusing on the technical challenges of detection, this research aims to contribute tools that safeguard authenticity in digital media. This study also emphasises how deepfake technology affects security and privacy, underscoring the need for trustworthy detection techniques. As generative models become more complex, it is imperative to improve detection precision and robustness to shield people, especially women and children, and society from the growing dangers posed by synthetic media.

II. LITERATURE REVIEW

II.1 Evolution and Technical Foundations of Deepfake Technology

What are deepfakes and why do they matter? Deepfakes are synthetic audio, image, or video outputs generated using advanced machine learning (ML) techniques, most commonly Generative Adversarial Networks (GANs). These outputs replicate human appearance, voice, and behaviour

with a level of realism that often renders manual detection difficult. As the technology has become more accessible, concerns have increased regarding misinformation, reputational harm, and the exploitation of individuals who appear in non-consensual synthetic media.

Deepfake technology emerged publicly in 2017 when an anonymous Reddit user posted AI-manipulated pornographic videos of celebrities. By April 2018, the phenomenon gained mainstream attention through a Buzzfeed demonstration featuring a synthetic video of President Obama, illustrating both the technology's capabilities and its potential for misinformation (Facebook, 2018). The rapid proliferation was striking: by late 2018, 96% of the 14,678 identified deepfakes online were pornographic, with women disproportionately targeted (Deeptrace, 2018). Applications like DeepNude (2019) and Telegram bots (2019) further democratised the creation of non-consensual explicit imagery, affecting over 45,000 users within months (Hao, 2020).

II.1.1 Technical Foundations: Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are neural network-based architectures in which two models, the Generator and the Discriminator, are trained simultaneously in an adversarial framework (Goodfellow et al., 2014, p. 1). In this collaboration, the Generator produces synthetic images, whereas the Discriminator evaluates whether an input image is authentic or generated. The analysis begins with very basic features and progresses to increasingly more high-level ones, each layer compressing the input more. During training, the Generator iteratively improves its output, making it increasingly difficult for the Discriminator to distinguish synthetic images from real ones. Conversely, the Discriminator improves its ability to detect artefacts and inconsistencies. This competitive process drives both components toward higher performance and results in outputs that closely approximate the distribution of the training data (de Vries, 2020, p. 2113).

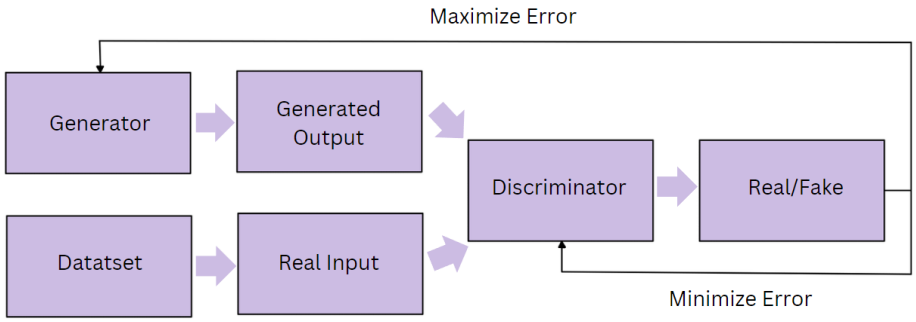

Figure 1. GAN: Neural Networks 1

Source: Author's processing

II.1.2 Machine Learning Concepts Underlying Deepfakes

The foundation of Machine Learning (ML) relies on a preference for implicit rather than explicit programming. Unlike traditional AI methods, where a computer system is given a predefined model by its designer, ML systems are designed to autonomously generate models from sets of examples (de Vries, 2020, p. 2111). Consequently, the computer does not receive direct instructions but learns independently, hence the term Machine Learning. The distinction between supervised and unsupervised learning is the type of input used and the resulting output (de Vries, 2020, p. 2112).

Supervised ML systems operate on labelled examples. In Supervised Machine Learning (ML), it is akin to having a teacher guide you through a set of well-labelled textbooks. For instance, if you want to distinguish among different bird species, the teacher gives you a collection of bird images, each labelled with its name. Additionally, you learn to recognise and classify various bird species based on the provided labels. On the contrary, Unsupervised Machine Learning is comparable to exploring a library filled with unmarked books. In this analogy, imagine delving into a section dedicated to birds, but without any labels on the books. In this case, you will not be explicitly taught about different bird species. Instead, you observe common themes and patterns across the books, leading to a broader understanding of birds and their characteristics. This exploration does not focus on specific classifications but on generating output that resembles their training data.

II.2 Psychological Dimensions of Digital Sexual Violence

The act of distributing sexual imagery is as old as time itself. Only in the last century has it been monetised and transformed into a billion-dollar industry. While it has become the norm to engage in this behaviour, we are still unable to profile when consuming this media becomes pathological. This section reviews the neuroscience of porn addiction and how we react to deepfakes.

The utilisation of online pornography, also referred to as Internet pornography, is recognised as a potentially addictive behaviour specific to the Internet. This behaviour encompasses the use of the Internet for engaging in various sexually gratifying activities, predominantly involving the consumption of pornography (Cooper, A. 2004). Research has shown that regular use of pornography is linked to issues such as objectification, sexual deception, and sexual aggression. In a study, researchers discovered that the consumption of reality TV, sports programming, and pornography was correlated with increased acceptance of objectification of women and more frequent engagement in acts of sexual deception (Seabrook et al. 2019).

Lead researcher and a psychologist at Princeton University, Susan Fiske, experimented in 2008 involving 21 men. The goal was to study male behaviour and how they respond to the exposure of scantily clad women. Brain scans conducted during the study indicated increased activity in the brain region associated with tool use when men viewed such images. Additionally, the men were more inclined to associate sexualised images of women with first-person action verbs like "I push,

I grasp, I handle," according to Fiske. These findings suggest a neural pattern consistent with dehumanisation and instrumental perception of women.

In 2018, Indian investigative journalist Rana Ayyub became the target of an online hate campaign due to her outspoken criticism of the Indian government, particularly her condemnation of the rape of an eight-year-old Kashmiri girl. This campaign included the dissemination of rape and death threats, as well as the circulation of manipulated pornographic videos featuring her. This phenomenon reached a disturbing peak in 2025, when thousands of South Korean women and minors were targeted by deepfake pornography, leading to a government investigation and public outcry (Yonhap News, 2025).

The psychological harm inflicted by deepfake pornography parallels that of revenge porn, despite the absence of any authentic explicit imagery. Victims report shame, humiliation, anxiety, depression, and social withdrawal (Harris, 2019). Neuroscientific research provides insight into the objectification underlying such abuse: Fiske's (2008) study demonstrated that viewing sexualised images of women activated brain regions associated with tool use rather than social cognition in male participants, suggesting dehumanisation at a neural level. This objectification creates an environment where creators view their work as harmless experimentation—as the anonymous DeepNude creator claimed (Motherboard, 2017)—while victims experience profound psychological violence and erosion of agency.

Alternative approaches investigate biological signal inconsistencies, such as abnormal eye blinking patterns or absence of physiological signals like pulse detection in facial videos (Waldrop, 2020). Feature-based forensic methods analyse compression artefacts and image metadata. Yet with the rapid rise of precise quality in AI image generation, research examining traditional machine learning methods as baselines for deepfake detection remains limited, despite their potential value for understanding fundamental distinguishability.

II.3 Legal and Policy Gaps

Despite growing awareness, legal frameworks have failed to keep pace with the rapid development of deepfake technology. No comprehensive federal legislation specifically addresses deepfake pornography, leaving victims with limited recourse (Harris, 2019). Some states have enacted legislation, but enforcement remains challenging due to jurisdictional issues and Section 230 immunity for platforms. The financial burden of content removal falls on victims, while perpetrators often operate anonymously across international borders. This legal vacuum enables continued exploitation, particularly of women and minors, underscoring the urgent need for both technical detection solutions and comprehensive policy responses.

III. METHODOLOGY AND DATA

This study employs a multimodal detection approach combining deep learning features with forensic analysis techniques to differentiate AI-generated images from authentic photographs. Our methodology integrates complementary detection strategies that examine both high-level semantic content and low-level statistical artefacts.

III.1 Dataset Preparation

We curated a balanced dataset of 200 images, split evenly into two classes: 100 authentic photographs and 100 AI-generated images. Real images were sourced from established photography repositories including Getty Images, Pinterest, and Google Images, ensuring diverse subjects, lighting conditions, and capture devices to represent authentic photographic content. AI-generated images were collected from multiple generators including Midjourney (versions 5-6), Stable Diffusion (various models), and DALL-E (2-3). This diversity in both real and synthetic sources increases the robustness of our detection approach by preventing overfitting to specific generator signatures or photographic styles.

Images were resized to 380×380 pixels, chosen to balance computational efficiency with preservation of fine-grained details necessary for forensic analysis. The dataset was split 80/20 into training (160 images) and testing (40 images) subsets using stratified sampling to maintain class balance.

III.2.1 Deep Semantic Features (2,304 dimensions)

We employ two pre-trained deep learning models to capture high-level semantic and visual representations:

- **EfficientNet-B3** (1,536 dimensions): Pre-trained on ImageNet, this convolutional architecture efficiently captures hierarchical visual features from low-level edges to high-level objects. We extract features from the penultimate layer before the classification head, providing a rich representation of image content that differs subtly between real photographs (which contain natural scene statistics) and AI generations (which may exhibit distributional artifacts from the generative process).
- **CLIP ViT-L/14** (768 dimensions): This vision-language model, trained on image-text pairs, provides multimodal embeddings that capture semantic meaning and visual concepts. CLIP features are particularly valuable because AI generators are often trained to optimise for semantic coherence that CLIP would recognise, potentially creating detectable patterns in this embedding space.

III.2.2 Frequency Domain Analysis (3 dimensions)

We perform Fast Fourier Transform (FFT) analysis on grayscale images to examine frequency spectrum characteristics. AI generators often produce images with unnatural frequency

distributions—particularly "spectral holes" or dead zones in mid-to-high frequencies where real cameras produce noise and textural detail. We compute: (1) presence of low-energy spectral regions excluding DC component, (2) magnitude spectrum standard deviation, and (3) mean magnitude, capturing the overall frequency distribution profile.

III.2.3 Noise Residual Statistics (3 dimensions)

Authentic photographs contain sensor noise patterns specific to camera hardware, while AI-generated images exhibit algorithmic noise from the generative process. We extract high-frequency noise residuals using a Laplacian kernel and compute statistical measures (mean, standard deviation, median absolute value) that characterise these noise patterns. Differences in noise structure provide forensic evidence of image origin.

III.2.4 Local Binary Patterns (26 dimensions)

Local Binary Patterns (LBP) capture micro-texture information at the pixel level. We compute uniform LBP histograms with radius 3 and 24 sampling points, providing rotation-invariant texture descriptors. AI generators may produce textures that appear realistic at first glance but exhibit subtle statistical regularities detectable through LBP analysis.

3.2.5 Facial Forensics (3 dimensions)

For images containing faces (detected via MTCNN), we extract face-specific features: (1) eye region symmetry—measuring pixel-level differences between left and right eye regions, as AI often produces unnaturally symmetric facial features, (2) teeth whiteness extremity—detecting unnaturally bright teeth common in AI-generated portraits, and (3) number of detected faces. For images without faces, these features are set to zero.

III.2 Evaluation Metrics

Model performance is evaluated using multiple metrics to provide comprehensive assessment: accuracy (overall correctness), precision (positive predictive value for each class), recall (sensitivity or true positive rate), F1-score (harmonic mean of precision and recall), and ROC-AUC (area under the receiver operating characteristic curve, measuring discrimination ability across all classification thresholds). These metrics collectively assess both the model's classification accuracy and its ability to provide reliable confidence estimates for practical deployment.

IV. RESULTS

The confusion matrix for our test set (n=40) reveals the distribution of correct and incorrect classifications:

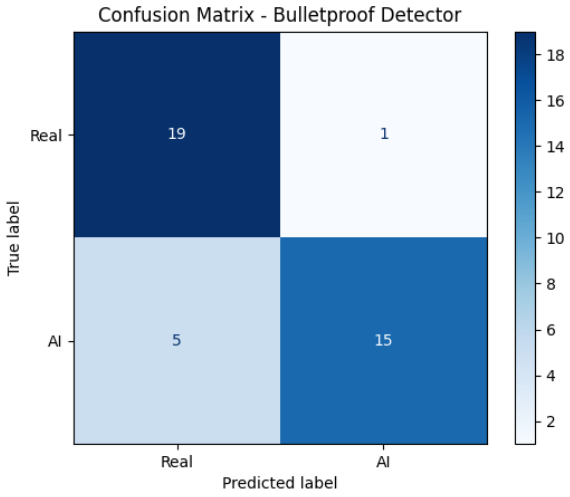

Figure 2. Confusion matrix

Source: Author's processing

True Negatives (19): The model correctly identified 19 of 20 authentic photographs, demonstrating strong capability to recognise genuine content.

False Positive (1): Only one real image was misclassified as AI-generated. This low false positive rate (5%) is critical for practical deployment, as incorrectly flagging authentic content could discredit legitimate evidence or unfairly censor real photographs.

True Positives (15): The model correctly detected 15 of 20 AI-generated images, showing that the multimodal feature approach captures distinguishing characteristics of synthetic content.

False Negatives (5): Five AI-generated images evaded detection and were classified as real. This 25% miss rate represents the primary limitation of the current approach and likely reflects cases where modern generators (particularly Midjourney v6 and Stable Diffusion XL) successfully replicate natural photographic statistics across all examined feature modalities.

The Receiver Operating Characteristic (ROC) curve plots true positive rate (sensitivity) against false positive rate (1-specificity) across all possible classification thresholds. Our model achieves an Area Under the Curve (AUC) of 0.9825, indicating excellent discrimination ability.

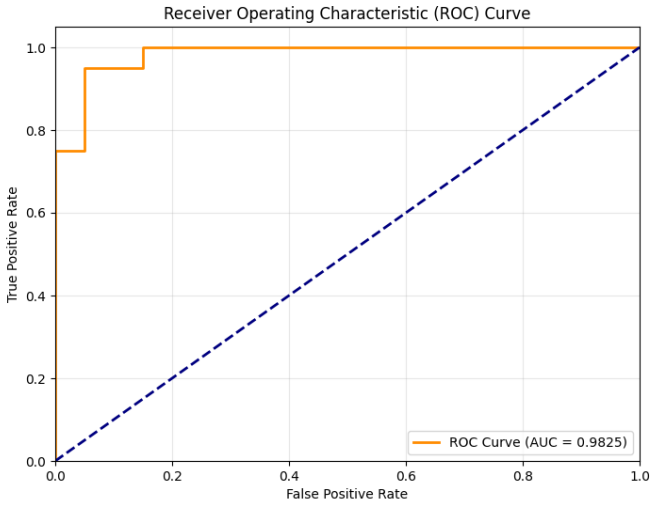

Figure 3. ROC curve

Source: Author's processing

The high AUC, despite 85% accuracy at the 0.5 threshold, suggests that most misclassifications occur near the decision boundary, where the model exhibits uncertainty. For practical applications with different risk tolerances, the threshold could be adjusted:

Stricter threshold (e.g., 0.7): Reduces false positives further, flagging only high-confidence deepfakes. Useful for automated content moderation where false accusations are particularly harmful.

Looser threshold (e.g., 0.3): Increases recall on AI images, catching more deepfakes at the cost of more false alarms. Appropriate for initial screening in high-risk contexts.

The precision-recall curve displays the trade-off between precision (positive predictive value) and recall (sensitivity) across different thresholds. This metric is particularly informative for understanding performance on the positive class (AI-generated images).

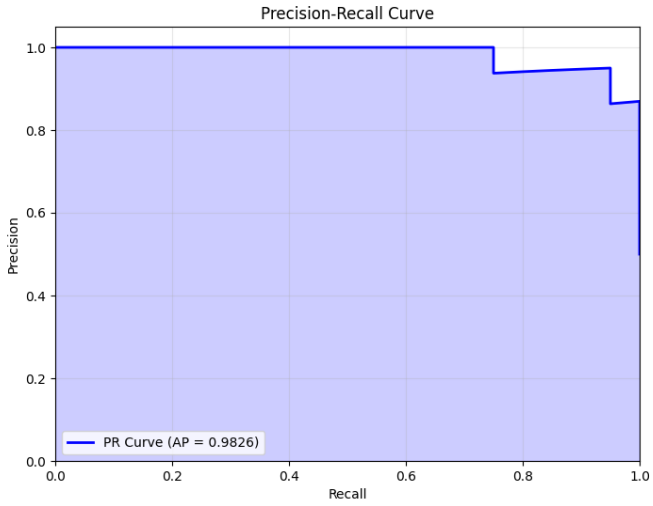

Figure 4. Precision-Recall Curve analysis

Source: Author's processing

At our default threshold (0.5), the model achieves:

- **Precision = 93.75%** on AI class: When the model predicts "AI-generated," it is correct 15 out of 16 times
- **Recall = 75%** on AI class: The model detects 15 out of 20 actual AI images

The curve's overall shape and the maintained high precision across varying recall levels confirm that the model has learned meaningful discriminative patterns rather than exploiting dataset artefacts or noise.

IV. DISCUSSIONS

Our multimodal approach demonstrates that combining deep semantic features with forensic analysis yields strong detection performance, with an AUC of 98.25% indicating excellent class separation. However, the 75% recall on AI images reveals significant limitations requiring careful consideration.

With 96% of non-consensual deepfakes targeting women, imperfect detection (25% false negative rate) remains problematic given consequences including trauma, harassment, and reputational harm. Technical solutions must integrate with broader protections: platform accountability for proactive moderation, comprehensive legislation criminalising non-consensual deepfakes with civil

remedies for victims, and support systems addressing psychological impacts. Our 93.75% precision on AI content enables flagging for human review without overwhelming moderators, but detection alone cannot solve deepfake abuse.

False positives (5 real images misclassified as AI) could discredit authentic evidence in journalism or legal contexts. Applications requiring high confidence should adjust thresholds to trade recall for precision. Moreover, detection systems may inadvertently accelerate the improvement of generators as developers specifically target detection vulnerabilities.

Promising avenues include temporal analysis for video deepfakes, adversarial training paradigms, hybrid human-AI review systems, standardised benchmarks tracking detection performance across evolving generators, and expanded forensic modalities examining lighting consistency, physically impossible reflections, and semantic anomalies.

V. CONCLUSIONS

This research investigated multimodal deepfake detection combining deep learning features (EfficientNet-B3, CLIP) with forensic analysis (frequency domain, noise statistics, texture patterns, facial forensics) to address non-consensual synthetic media that disproportionately targets women. Our ensemble achieved 85% accuracy and 98.25% ROC-AUC on images from modern generators (Midjourney, Stable Diffusion, DALL-E), demonstrating that AI-generated content exhibits detectable differences across multiple feature modalities.

However, 75% recall on synthetic images reveals that one quarter of deepfakes evade detection, indicating significant challenges remain for high-stakes deployment. This technical limitation underscores that deepfake detection is fundamentally a sociotechnical challenge requiring responses spanning technology, policy, and social norms. Detection tools provide necessary but insufficient protection; comprehensive solutions demand platform accountability, legislation criminalising non-consensual deepfakes, victim support systems, and cultural shifts prioritising digital consent.

Despite some advancements in detection and growing public discourse, legal systems lag far behind. Social norms continue to diminish the seriousness of digital sexual violence, and there is no federal criminal framework that particularly targets deepfake pornography. Deepfakes will continue to pose an unchecked threat until legal and technological safeguards are improved, as well as until society recognises the agony that victims have endured. Meaningful change requires not only code and legislation but a shift in cultural values that prioritises consent, safety, and digital integrity.

REFERENCES

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of deepfakes: Landscape, threats, and impact*. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Ayyub, R. (2018, November 21). I was the victim of a deepfake porn plot intended to silence me. *HuffPost*. https://www.huffpost.com/entry/deepfake-porn_n_5bf2c126e4b0f32bd58ba316

BBC News. (2021, March 15). Mother "used deepfake to frame cheerleading rivals." https://www.bbc.com/news/technology-56404038

Cooper, A. (2004). Online sexual activity in the new millennium. *Contemporary Sexuality, 38*(3), 1–7.

Deeptrace. (2019). *The state of deepfakes: Landscape, threats, and impact*. https://deeptracelabs.com/

de Vries, K. (2020). Deepfakes, deep harms? An analysis of the gendered harms of synthetic media. *Journal of Global Ethics, 16*(2), 147–165. https://doi.org/10.1080/17449626.2020.1810165

Fiske, S. T., Cuddy, A. J., & Glick, P. (2008). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27, pp. 2672–2680). MIT Press.

Hao, K. (2020, October 20). A deepfake bot is being used to "undress" underage girls. *MIT Technology Review*. https://www.technologyreview.com/2020/10/20/1011116/deepfake-ai-telegram-bot-undresses-women-and-underage-girls/

Harris, D. (2019). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law and Technology Review, 17*(1), 99–127.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Ovadya, A. (2018, February 11). Infocalypse now: P0wnage of the public sphere, and what to do about it. *Medium*. https://medium.com/@virgilgr/infocalypse-now-p0wnage-of-the-public-sphere-and-what-to-do-about-it-b8bfc5ce9ee7

Seabrook, R. C., Ward, L. M., & Giaccardi, S. (2019). Less than human? Media use, objectification of women, and men's acceptance of sexual aggression. *Psychology of Violence, 9*(5), 536–545. https://doi.org/10.1037/vio0000198

Sensity. (2023). *The state of deepfakes 2023: Reface in the top*. https://sensity.ai/reports/

Vice. (2017, December 11). AI-assisted fake porn is here and we're all fucked. *Motherboard*. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Waldrop, M. M. (2020, March 16). Synthetic media: The real trouble with deepfakes. *Knowable Magazine*. https://doi.org/10.1146/knowable-031620-1

Yonhap News Agency. (2024, August 28). S. Korea launches investigation into deepfake sex crimes. https://en.yna.co.kr/view/AEN20240828007251315