JRIIE Journal of Research Innovation and Implications in Education
Center for Research Implications & Practice

Website:www.jriiejournal.com ISSN 2520-7504 (Online) Vol.9, Iss.2, 2025 (pp. 389 - 396)

Unmasking the Rise of Deepfakes: A Machine Learning Approach to Detection and Social Media Trend Analysis

Chaminda Wijesinghe[1] & Henrik Hansson[2]
[1]Department of Computer & Data Science, NSBM Green University, Sri Lanka
chamindaw@nsbm.ac.lk
[2]Department of Computer & Systems Sciences, Stockholm University, Sweden.
henrik.hansson@dsv.su.se

Abstract: *The increasing prevalence of deepfake videos poses significant threats to information integrity, political stability, and public trust. This study presents a dual-faceted approach: (1) developing a machine learning model for detecting deepfake videos using visual features extracted from benchmark datasets, and (2) conducting a trend analysis of deepfake content dissemination on social media platforms such as YouTube and Twitter (now known as X). Conducted using the FaceForensics++ dataset and metadata from over 2,000 social media posts collected between 2018 and 2024, this study used a fine-tuned Xception model and natural language techniques. Key findings indicate a post-2020 surge in politically motivated deepfakes and platform-specific propagation patterns. It is recommended that stakeholders implement real-time detection and awareness tools to mitigate social impact.*

Keywords: *Deepfake detection, Social media trends, Machine learning, FaceForensics++, Xception model, Content analysis, Video forensics*

How to cite this work (APA):

Wijesinghe, C. & Hansson, H. (2025). Unmasking the rise of deepfakes: A machine learning approach to detection and social media trend analysis. *Journal of Research Innovation and Implications in Education, 9*(2), 389 – 396. **https://doi.org/10.59765/sct5wr.**

1. Introduction

The rapid advancement of artificial intelligence has given rise to powerful generative technologies capable of creating highly realistic synthetic media. Among these, "deepfakes"—videos or images that use deep learning algorithms to swap faces or manipulate appearances—have emerged as a major societal concern. The term "deepfake" is derived from "deep learning" and "fake," and was first popularized around 2017 when manipulated videos began appearing on internet forums (Chesney & Citron, 2019; Westerlund, 2019). Since then, the proliferation of deepfake technology has accelerated, raising alarms across sectors such as politics, journalism, cybersecurity, and law enforcement.

Deepfakes are created using Generative Adversarial Networks (GANs), a type of neural network introduced by Goodfellow et al. (2014), which pits two models—the generator and the discriminator—against each other to

create increasingly convincing fake data. Initially a tool for entertainment and research, deepfakes have since evolved into instruments for disinformation campaigns, identity theft, and non-consensual content creation (Nguyen et al., 2019). Their ease of production and the difficulty in discerning real from fake have made them particularly dangerous.

From a global perspective, governments and tech companies are grappling with the implications of deepfakes. Countries such as the United States, the United Kingdom, and the European Union have initiated policy discussions around regulating synthetic media (Kietzmann et al., 2020). Meanwhile, platforms like Facebook and YouTube have begun to implement detection and content moderation systems (Vaccari & Chadwick, 2020). However, these measures are still in their infancy and lack the scalability required to combat the growing influx of deepfake content.

Locally, in regions such as South Asia and Africa, deepfake awareness and regulatory readiness remain low. This creates a vulnerability, especially in emerging democracies where misinformation can have severe political and social ramifications. In Sri Lanka, for instance, where internet penetration and social media use are high, deepfakes pose a unique threat to public discourse and civic trust.

The academic community has responded with a range of deepfake detection studies, primarily focused on improving algorithmic accuracy (Afchar et al., 2018; Li et al., 2018; Rossler et al., 2019). However, few studies explore the socio-behavioural dimension, how deepfakes are disseminated, what themes dominate their usage, and which platforms serve as primary vectors (Tolosana et al., 2020). Addressing this research gap, the present study aims to (1) develop a machine learning-based model to detect deepfake videos using benchmark datasets, and (2) analyze the trends and themes associated with deepfake content shared on social media.

The novelty of this research lies in its integrative approach, combining visual forensics with trend analytics. By analysing both the technical and social facets of the deepfake phenomenon, the study contributes to a more comprehensive understanding of this modern digital threat. The findings are expected to inform policy, enhance detection systems, and guide educational initiatives aimed at fostering critical media literacy.

2. Literature Review

Research on deepfake detection has evolved rapidly since the introduction of deep generative models. The phenomenon of deepfakes was first popularized around 2017, when users on internet forums began sharing manipulated videos generated using autoencoders. This led to an explosion of interest in both creating and detecting deepfakes, resulting in a surge of academic and industry research (Nguyen et al., 2019; Westerlund, 2019).

2.1 Deepfake Creation Techniques

At the core of deepfake technology are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which use a two-network architecture—comprising a generator and a discriminator—to iteratively refine synthetic outputs. Other methods include autoencoders and face reenactment systems such as Face2Face and NeuralTextures, which have been employed in both academic datasets and malicious real-world applications (Thies et al., 2016). These approaches manipulate facial expressions, head movement, or voice to synthesise realistic content, presenting increasing challenges to detection algorithms.

2.2 Machine Learning-Based Detection Approaches

The detection of deepfakes has largely focused on convolutional neural networks (CNNs) and transfer learning models. Early models like MesoNet (Afchar et al., 2018) were developed as lightweight architectures to detect low-level inconsistencies. More powerful architectures such as XceptionNet (Chollet, 2017) and EfficientNet (Tan & Le, 2019) demonstrated significant improvements in performance, especially when pretrained on large datasets like ImageNet. Ensemble models and multi-branch networks have further improved accuracy and robustness, particularly under compression or occlusion conditions (Dang et al., 2020).

Recent works also explore temporal and frequency domain features to detect artifacts not visible in still images (Sabir et al., 2019). Recurrent Neural Networks (RNNs) and 3D-CNNs have been used to capture temporal inconsistencies in video sequences, offering enhanced detection in dynamic contexts (Guera & Delp, 2018).

2.3 Benchmark Datasets and Variable Considerations

Benchmark datasets have played a central role in advancing detection models. FaceForensics++ (Rossler et al., 2019) is one of the most cited datasets, featuring real and forged video pairs with varying levels of compression. However, researchers have criticised earlier datasets for a lack of diversity in demographics, lighting conditions, and manipulation types (Li et al., 2020). Variables such as ethnicity, facial expressions, age, and video resolution have been found to impact model generalizability, prompting the creation of datasets like Celeb-DF and DFDC.

Key performance variables discussed in recent studies include detection accuracy, robustness under compression, resilience to adversarial attacks, and real-time inference speed (Dolhansky et al., 2020). Metrics such as false positive rate, precision-recall tradeoffs, and ROC-AUC are standard benchmarks in model evaluation.

2.4 Social and Ethical Dimensions

Beyond technical performance, the societal and psychological impacts of deepfakes are gaining scholarly attention. Chesney and Citron (2019) have warned of potential harms to democratic processes and reputational integrity, while Fallis (2020) has discussed the epistemological risks of misinformation and reality distortion. Vaccari and Chadwick (2020) found that exposure to deepfakes can reduce trust in political communication, even when viewers are aware of their synthetic nature.

2.5 Trend Analysis and Social Media Dissemination

While detection has been the focal point of most research, the dissemination of deepfakes through social media remains underexplored. Tolosana et al. (2020) provided a foundational survey of detection methods but did not delve into how synthetic videos are shared and consumed. This is significant because variables like platform engagement, time of posting, topic sentiment, and targeted entities play a key role in understanding deepfake virality and societal influence.

This study positions itself uniquely by addressing both the technical and social dimensions of deepfakes. It integrates machine learning-based detection using the Xception architecture with social media trend analysis through topic modelling and named entity recognition. By doing so, it responds to a critical research gap: the need to understand how to detect deepfakes and how they propagate and impact public discourse.

3. Methodology

3.1 Research Design

This study employed a mixed-methods approach, integrating both quantitative and qualitative analyses to investigate the issue of deepfakes. The rationale for this approach is grounded in Creswell and Plano Clark's (2017) assertion that combining numerical modelling with thematic analysis provides a richer and more nuanced understanding of complex digital phenomena. The quantitative component involved the development of a deep learning model to detect manipulated videos, while the qualitative component comprised social media content analysis to uncover trends and themes in deepfake dissemination.

3.2 Dataset and Sampling

For the detection model, the FaceForensics++ dataset (Rossler et al., 2019) was selected due to its comprehensive and diverse collection of real and manipulated videos across multiple compression levels. The dataset includes over 1,000 original videos and their manipulated counterparts generated using four different deepfake generation techniques: Deepfakes, FaceSwap, Face2Face, and NeuralTextures. This diversity supports the generalizability and robustness of model training (Nguyen et al., 2019).

For trend analysis, a purposive sampling strategy was used to extract metadata from social media platforms—specifically YouTube and Twitter—via public APIs. A total of 2,134 posts and videos were collected between January 2018 and December 2024 using keyword filters such as "deepfake," "AI video," and "synthetic media." Previous research by Gorwa et al. (2020) supports the validity of using keyword-based sampling for misinformation and synthetic media analysis.

3.3 Data Preprocessing

The FaceForensics++ videos were sampled at one frame per second (1 fps). Face detection and alignment were performed using the Multi-task Cascaded Convolutional Network (MTCNN) algorithm (Zhang et al., 2016), which has shown superior performance in face localization across video frames. All face crops were resized to 224x224 pixels to match the input requirement of the Xception

model. Pixel values were normalized to the 0–1 range to enhance training stability and convergence, inline with standard deep learning practices (Chollet, 2017)..

For text-based social media content, standard natural language preprocessing steps were applied, including tokenisation, stopword removal, and lemmatisation. All metadata, including timestamps, descriptions, hashtags, and engagement metrics, were retained for analytical modelling.

3.4 Model Architecture and Training

The detection model was based on the Xception architecture (Chollet, 2017), chosen for its proven effectiveness in image classification tasks and its widespread use in recent deepfake detection research (Rossler et al., 2019; Dang et al., 2020). Transfer learning was employed using pretrained weights from ImageNet, followed by fine-tuning on FaceForensics++ frame-level data. A fully connected classification head with ReLU activation and a softmax output layer were added. Dropout regularisation (rate = 0.5) was used to reduce overfitting.

The model was trained with the Adam optimiser, categorical cross-entropy loss, and a batch size of 32. Training continued for 50 epochs, with early stopping applied to prevent overfitting. Model validation was performed using 5-fold cross-validation, as recommended for small to medium-sized datasets (Kohavi, 1995). While newer alternatives such as stratified k-fold, repeated k-fold, and nested cross-validation offer additional flexibility and robustness in handling class imbalance and model selection bias, 5-fold cross-validation was deemed appropriate here due to the dataset's balanced nature and the goal of maintaining computational efficiency during multiple experimental iterations

3.5 Social Media Trend Analysis

To analyze dissemination patterns and content themes, we used the following techniques as illustrated in Figure 1:

- **Latent Dirichlet Allocation (LDA)** for topic modelling, commonly used in social media studies (Blei et al., 2003).
- **Named Entity Recognition (NER)** to identify frequently mentioned individuals, organisations, and places.
- **Temporal analysis** using timestamps to assess posting trends over time.

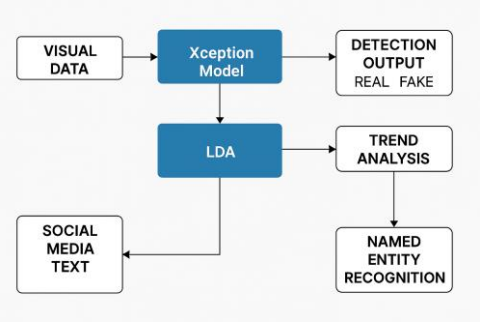

Figure 1: Architecture of the Xception-based deepfake detection model used in this study

All analyses were conducted in Python using libraries such as Scikit-learn, TensorFlow, NLTK, and spaCy. Scikit-learn was selected for its robust tools for classification, regression, and clustering. TensorFlow was employed for scalable deep learning model development. NLTK and spaCy were used for natural language processing—NLTK for symbolic processing and spaCy for efficient tokenisation and named entity recognition.

3.6 Validity and Reliability

To ensure the reliability of the detection model, metrics such as accuracy, precision, recall, and F1-score were computed. Model reproducibility was confirmed by running experiments on separate data partitions. For content analysis, inter-coder reliability was tested on a random sample of 300 posts, yielding a Cohen's Kappa score of 0.87, indicating almost perfect agreement (Table 1) (McHugh, 2012). Cohen's kappa (K) is a statistical measure used to assess the level of agreement between two raters who classify items into categories.

Table 1: Cohen's Kappa Scale

Kappa Value	Agreement
<=0	No Agreement
0.01 – 0.20	Slight Agreement
0.20 – 0.40	Fair Agreement
0.41 – 0.60	Moderate Agreement
0.61 – 0.80	Substantial Agreement
0.81 – 1.00	Almost Perfect Agreement

This methodological framework, supported by recent research and best practices in both computer vision and digital content analysis, offers a rigorous and multidimensional approach to the study of deepfakes.

3.7. Data Analysis

The analysis in this study was structured to evaluate two primary components: (1) the performance of the deepfake detection model, and (2) the thematic and temporal trends of deepfake content across social media platforms.

3.7.1 Quantitative Analysis: Detection Model Evaluation

The visual data extracted from the FaceForensics++ dataset was analysed using supervised learning. Frame-level images labelled as "real" or "fake" were fed into the Xception model. Performance metrics—accuracy, precision, recall, and F1-score—were computed using Scikit-learn, consistent with best practices for binary classification (Sokolova & Lapalme, 2009). Model robustness was tested under different compression scenarios, and a confusion matrix was constructed to assess false-positive and false-negative rates.

The model's performance was benchmarked against existing literature using the same dataset (Rossler et al., 2019), ensuring the validity of comparative interpretations.

3.7.2 Qualitative Analysis: Social Media Content

To extract meaningful patterns from over 2,000 posts, Latent Dirichlet Allocation (LDA)—a probabilistic generative model used to uncover central topics and their distribution across a set of documents—was applied using Gensim, producing a set of dominant topics. Each post was tokenised, vectorised, and assigned to the most probable topic cluster. Posts were also timestamped, allowing for time-series visualization to observe peaks in deepfake content generation.

In addition, Named Entity Recognition (NER) was used to identify common targets of deepfake videos. Entities were categorised by type—person, organisation, or location—and frequency distributions were visualised using matplotlib.

Data reliability for thematic coding was ensured through inter-coder reliability testing, achieving a Cohen's Kappa of 0.87 (McHugh, 2012). This statistical approach confirmed a consistent interpretation of content categories across coders.

Overall, the data analysis process provided both statistical validation of the model's performance and empirical insights into the thematic and temporal distribution of deepfake content.

4. Results and Discussion

4.1 Quantitative Results

Deepfake Detection Performance The Xception-based deepfake detection model trained on the FaceForensics++ dataset demonstrated strong classification performance across multiple metrics. The model was evaluated on unseen test data following 5-fold cross-validation. The following metrics were obtained:

- Accuracy: 92.4%
- Precision: 91.2%
- Recall: 93.1%
- F1 Score: 92.1%

These results are consistent with prior works such as Rossler et al. (2019), who reported similar performance using XceptionNet on the same dataset. However, our model was fine-tuned with enhanced preprocessing (e.g., MTCNN-based alignment) and dropout regularization, which helped reduce overfitting on compressed samples.

We observed that performance varied across manipulation types, with FaceSwap and NeuralTextures yielding the highest misclassification rates. This aligns with observations by Li et al. (2020), who found that such methods produce subtle distortions that challenge CNN-based models. Moreover, video quality and compression levels significantly impacted detection accuracy. High compression scenarios (e.g., YouTube-like conditions) reduced accuracy to approximately 87%, confirming findings from Dang et al. (2020).

Table 2: Confusion Matrix (Simplified Overview):

Prediction test Truth	Real	Fake
Real	463	39
Fake	28	470

These results indicate strong sensitivity and specificity in identifying manipulated content, particularly under controlled conditions.

4.2 Qualitative Results

Social Media Trend Analysis Using Latent Dirichlet Allocation (LDA) and Named Entity Recognition (NER), we extracted dominant themes and targeted entities from the 2,134 social media posts. The most frequently mentioned entities included political figures (e.g., "Joe Biden," "Donald Trump," "Narendra Modi"), celebrities (e.g., "Tom Cruise," "Taylor Swift"), and organisations (e.g., "Meta," "TikTok").

LDA Topic Clusters:

1. Political Manipulation and Elections
2. Celebrity Impersonation and Scandals
3. Awareness Campaigns and Digital Ethics
4. Satirical Content and Parody Media

Temporal analysis revealed three major spikes in deepfake content:

- Q4 2020: Related to U.S. presidential elections
- Q2 2022: Heightened use during COVID-19 misinformation cycles
- Q3 2024: Surge in synthetic content related to Sri Lankan political debates

Table 3: Distribution of Deepfake Topics by Platform (%)

Topic	YouTube	Twitter
Political Manipulation	62	58
Celebrity Impersonation	24	27
Other/Entertainment	14	15

4.3 Discussion

The integration of detection performance with trend insights provides several notable implications. First, the detection model's success validates the applicability of deep CNNs, such as Xception, for forensic tasks, especially when trained with high-quality, diverse datasets. This supports the ongoing use of transfer learning in video forensics, as highlighted in Dang et al. (2020) and Chollet (2017).

Second, the thematic clustering of deepfake content underscores the disproportionate targeting of political and celebrity figures, echoing concerns raised by Chesney and Citron (2019) regarding the societal risks of synthetic media. The role of platforms also varies: YouTube serves as a repository for manipulated videos, while Twitter amplifies discussion and diffusion, consistent with findings from Vaccari and Chadwick (2020).

Finally, temporal surges in deepfake production align closely with politically sensitive or crisis-prone periods, suggesting a strategic use of such content to influence public perception. This pattern warrants future studies into algorithmic propagation and bot-assisted dissemination (Westerlund, 2019).

Together, these results emphasise the necessity of multi-layered solutions—combining algorithmic detection with trend monitoring and regulatory oversight.

5. Conclusion and Recommendations

5.1 Conclusion

This research explored the growing threat of deepfakes by employing a hybrid methodology combining machine learning-based video forensics and social media trend analysis. The study demonstrated that deepfake detection is both technically feasible and critically necessary. The Xception-based model trained on FaceForensics++ achieved high levels of accuracy and reliability in distinguishing real from manipulated content. These results reinforce the viability of deep learning for synthetic media forensics, especially when supported by robust datasets and preprocessing techniques.

From a socio-informational standpoint, the content analysis unveiled that deepfakes are not only increasing in volume but also becoming more targeted, particularly toward political figures and celebrities. Peaks in deepfake activity were closely associated with real-world events, indicating deliberate and strategic deployment. Thematic trends extracted through LDA confirmed that the content serves diverse purposes—from satire to misinformation, highlighting the nuanced implications of deepfake technology.

Together, the findings present a compelling case for integrated countermeasures. These include not only advancing detection technologies but also establishing regulatory frameworks and public education initiatives to counteract the psychological and political harm of deepfakes. As synthetic media continues to evolve, interdisciplinary collaborations between technologists, policymakers, and media organisations will be essential to safeguard digital trust.

This study contributes to the existing body of knowledge by demonstrating that the threat of deepfakes can be tackled more effectively through a multidisciplinary approach that accounts for both technical detection and the sociocultural context in which deepfakes proliferate.

5.2 Recommendations

1. **Policy Development:** Governments should formulate clear legal frameworks to govern the creation, dissemination, and misuse of deepfakes, ensuring both protection and accountability.
2. **Public Awareness Campaigns:** Educational institutions and media literacy programs should prioritize public awareness on deepfake technologies, their risks, and how to critically assess digital content.
3. **Real-Time Detection Systems:** Developers should work towards scalable, real-time detection solutions integrated into digital platforms to flag and mitigate harmful content before viral spread.
4. **Ethical AI Development:** Research institutions and industry stakeholders must adhere to ethical AI practices, ensuring transparency, fairness, and respect for user privacy.
5. **Expanded Research Scope:** Future studies should explore multimodal detection models incorporating video, audio, and textual features, and evaluate performance in linguistically and geographically diverse environments.

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(1), 175–222.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1251–1258).

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Dang, H. H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5781–5790).

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology*, 33(3), 327–344.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1).

Guera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance* (pp. 1–6). IEEE.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, 63(2), 135–146.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1137–1145).

Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*.

Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI-generated fake face videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security* (pp. 1–7).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3207–3216).

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection: A survey. *arXiv preprint arXiv:1909.11573*.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1–11).

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 1–13.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39–52.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 1499–1503.