Deep Learning for Social Security: Identifying Fake Accounts and AI-Generated Profiles

Dr. David Raj Micheal
Division of Mathematics
School of Advanced Sciences
Vellore Institute of Technology Chennai
Tamil Nadu – 600127
davidraj.micheal@vit.ac.in

Savani M Kulkarni
Division of Mathematics
School of Advanced Sciences
Vellore Institute of Technology Chennai
Tamil Nadu – 600127
savani.mkulkarni2023@vitstudent.ac.in

***Abstract*—By separating authentic accounts from fraudulent AI-generated profiles on social media sites, the study investigates a deep learning-based strategy to enhance social security. It improves identification accuracy by analyzing profile data and photos using Convolutional Neural Networks (CNN) and dense neural network architectures. While the dense neural network model handles numerical and category variables, the CNN model draws out visual patterns from profiles. The system uses both textual and visual data inputs to guide its conclusions. The efficacy of deep learning in social security applications is demonstrated by the model's overall accuracy of over 85 percent on a dataset that includes both real and AI-generated profiles. Enhancing model interpretability and diversifying datasets should be the main goals of future studies.**

***Index Terms*—AI-generated profiles,social security,CNN,neural network,deep learning.**

I. Introduction

Global connections have been transformed by digital social platforms, but they also bring with them new problems, like the emergence of AI-generated profiles and phony accounts. These accounts endanger user safety and confidence by disseminating false information, phishing scams, and financial frauds. Automated, data-driven techniques are required to combat advanced AI-powered fakes because traditional manual methods are insufficient. By integrating Convolutional Neural Networks (CNN) and dense neural network designs to assess image and profile data, this study investigates a deep learning-based method to detect phony accounts and AI-generated profiles. The system can categorize profiles with over 85 percent accuracy after being trained on labeled datasets from both synthetic and actual profiles. By offering a workable and scalable method for identifying and flagging fraudulent profiles, this study advances social security.

II. Problem Definition

Traditional techniques for identifying fraudulent accounts are inadequate due to the growing complexity of AI-generated profiles. The question this study attempts to answer is whether deep learning models can successfully differentiate between real and artificial intelligence (AI)-generated social media profiles.

III. Objective

- Create a deep learning model to determine whether social media profiles are authentic or fraudulent.
- To improve model correctness, make use of both textual and visual account characteristics.
- Evaluate the model's performance for social security purposes in actual situations.
- Describe the main elements that go into spotting phony accounts.

IV. Dataset

The dataset of real and fake social media profiles used in this project was obtained from publicly accessible information on websites such as Instagram. The dataset contains both visual and linguistic information, such as the length of the username, the number of followers, the frequency of posts, and the specifics of the profile description. It covers the use of convolutional neural networks for image-based categorization. For objective assessment, the dataset is divided into training, validation, and testing sets.

V. Literature Review

[1]Identity fraud is a significant issue on online social networks, and researchers are working to develop technologies to detect it. A study focuses on detecting identity fraud using clustering and classification techniques. Traditional methods have limitations and suggest ways to improve their effectiveness in real-world contexts. Data from social media accounts is collected and preprocessed using techniques like Natural Language Process (NLP), vectorization, dimensionality reduction, and data normalization. Features are extracted based on behavioral analysis and profile characteristics. Clustering approaches detect real or fake profiles, while deep learning classification uses Recurrent Neural Network (RNN) for classification. The system's effectiveness is demonstrated in experimental analysis. [2]Deep learning techniques have made it easier to create fake multimedia content, but existing forensic techniques only analyze one modality at a time. This lack of multimodal detectors is due to the lack of research datasets containing multimodal forgeries. This paper proposes a new audio-visual deepfake dataset containing multimodal video

forgeries, using Text-to-Speech (TTS) and Dynamic Time Warping (DTW) techniques to synthesize deepfake speech content. The proposed dataset can be used standalone or combined with DeepfakeTIMIT and VidTIMIT video datasets for multimodal research.

[3]Fake accounts on social media have become a significant concern, spreading fake news, rumors, spam, and unethical harassment. Detecting these accounts manually is time-consuming and challenging, making the need for automated approaches to detect them more significant. This article explores various methodologies to detect fake accounts and proposes a generalized deep learning model to detect them using multimodal data. The model uses a combination of textual, visual, and network-based features to capture the characteristics of fake accounts. The model was evaluated on a publicly available dataset of Twitter accounts and achieved state-of-the-art performance in detecting fake accounts with an F1 score of 0.96. Experiments were conducted to demonstrate the effectiveness of each feature and the combination of the three features. [4]The issue of fake accounts on social media, known as "impostors," has gained attention for their role in spreading misinformation, manipulating opinions, and interfering with elections. A GAN-based framework was studied to study this problem. The impostor aims to create realistic posts mimicking the target person's, while the detector identifies the posts. The model co-trained the impostor and detector until an equilibrium was reached. The method was applied to a Twitter dataset, and results showed the model is promising in generating and detecting impostors' posts.

[5]Deep learning advancements have made it difficult to differentiate between authentic and manipulated facial images and videos. DeepFake, a technology that manipulates facial appearances through deep generative approaches, has led to numerous malicious face manipulation applications. To reduce the impact of DeepFake creations, other techniques are needed to assess digital visual content integrity. A large body of research on DeepFake creation and detection aims to develop more robust approaches to deal with advanced DeepFake in the future. This study presents challenges, research trends, and directions related to DeepFake creation and detection techniques. [6]Deep learning algorithms are revolutionizing the production of audiovisual media, often referred to as "deepfakes." These synthetic audiovisual media, often indistinguishable from real sounds and images, are becoming increasingly trivial to produce. However, ethical concerns have been raised about their use. This article focuses on the concept of synthetic audiovisual media, its place within the audiovisual media taxonomy, and how deep learning techniques differ from traditional methods. The author argues that deepfakes and related synthetic media not only offer incremental improvements but also challenge traditional taxonomical distinctions, paving the way for new audiovisual media.

[7]The ethical debate surrounding face recognition and identification algorithms has led to the introduction of three face access models in a hypothetical social network. These models replace unapproved faces with quantitatively dissimilar deepfakes, using new metrics for this task. The access models are based on strict data flow and discuss their impact on privacy, usability, and performance. The system is evaluated on the Facial Descriptor Dataset and two synthetic datasets, with MFMC reducing average accuracy by 61 percent when running seven SOTA face recognizers. The study also analyzes similarity metrics, deepfake generators, and datasets in structural, visual, and generative spaces to support design choices and verify quality.

[8]Fake images and videos, created by digital manipulation, are a major public concern. DeepFake, a deep learning technique, has sparked interest in fake face detection using biometric anti-spoofing and data-driven deep learning. Traditional methods, like in-camera and out-camera fingerprints, are insufficient against unseen conditions, especially in social media. Recent surveys provide detailed reviews of facial manipulation techniques and benchmarks for fake detection methods.

[9]Research in security and identity management focuses on detecting false attributes in contested ecosystems. A new method using proximity-based methods and order-of-consensus-calculation is presented for detecting multiple formats of fake attributes, both known and unknown. The strength lies in investigating differences between natural and human activities and improving precision through forgery detection. However, challenges such as data quality and computational complexity persist. Synthetic identity fraud is tackled using a GCN-based approach, but limitations include data quality and scalability. Social network identity fraud is addressed using certified social profiles and decentralized trust computation. Research on object stiffness perception shows visual feedback plays a leading role, with behavioral adaptation techniques and outcome measures aiding in understanding multisensory integrations.

[10]The research on online social media platforms has grown, but recognizing anonymous, identical users remains a challenge. Current methods, such as text mining and location-based matching, are fragile due to the impersonation of users. The Friend Relationship-Based User Identification (FRUI) algorithm was proposed to address this issue. FRUI calculates match degrees for all candidate User Matched Pairs (UMPs), identifying only top-ranking UMPs as identical users. Experiments show FRUI performs better than current network structure-based algorithms, addressing the problem of recognizing anonymous users among multiple SMNs.

[11]This paper addresses the challenge of analyzing large, correlated social graph data while protecting users' privacy. It first protects users' data using local differential privacy, then designs a correlation-based privacy protection approach. A K-means algorithm is applied to perturbed local data, and synthetic graphs are generated. Experiments on the Facebook and Enron datasets show the proposed approach outperforms state-of-the-art methods in accuracy and utility evaluation criteria.

[12]The rapid development of synthetic media tools has blurred the lines between human-created and AI-generated

content, posing risks to digital media authenticity. This paper explores the security and legal implications of synthetic media creation, as well as the limitations of current detection methods. It suggests that a combination of neural networks and blockchain technology can help mitigate these risks by detecting and preserving media authenticity. Once an image is detected as human-created, its hash is stored on the blockchain, thereby preserving digital media authenticity. This work contributes to public resources for wider digital media authenticity detection and preservation.

[13]The rise in digital media usage has led to a rise in manipulation of visual media, including spreading false news and misinformation. This has made it challenging to maintain the authenticity of images and videos. This paper provides a comprehensive analysis of current techniques for detecting tampering, including machine learning-based methods, digital signatures, and statistical techniques. It also discusses emerging techniques like deep learning and blockchain-based approaches. The paper provides guidelines for identifying and preventing tampering, ensuring the authenticity of images and videos, and educating the public about the harmful effects of fake visual media.

[14]Advancements in artificial intelligence (AI) are enabling synthetic media, which can produce fake videos, photos, and writing. This technology has raised concerns about spreading political disinformation and financial harm. While the financial threat from synthetic media is low, experts differ on how to address it. This paper presents ten scenarios illustrating how criminals could use synthetic media to inflict financial harm on various targets, based on current technology and financial crime realities. [15]Artificial intelligence (AI) has produced synthetic images, particularly those from Generative Adversarial Networks (GAN) and Diffusion Models (DM), which are crucial for spreading misinformation on Online Social Networks (OSNs). However, current solutions struggle to identify these images accurately. A study suggests that selecting the right features and a deep learning-based classification model can improve synthetic image detector performance under challenging conditions. Two innovative solutions, a Gradient-based method and a novel Sine Transform Feature-based Network, achieve over 99 percent accuracy in detecting synthetic images and 91 percent in challenging scenarios.

VI. Methodology

A. Dataset Acquisition

- Downloaded datasets from Kaggle include both user account and image data. The Instagram dataset provides labels for fake and genuine accounts, while the real vs. fake faces dataset contains images labeled as ”real” or ”fake.”

B. Data Preprocessing

- Loaded and inspected datasets to understand feature distributions and missing values.
- Preprocessed numerical and categorical account features, scaling them to standardize inputs for the neural network.
- Applied data transformations to the image dataset, scaling pixel values to a 0-1 range.

C. Model Design

- **Profile Data Classification**: A Sequential Neural Network model was built to classify accounts as real or fake based on profile attributes like username length, post count, follower count, and profile picture indicators.
  - The model includes multiple dense layers with ReLU activation functions, interspersed with dropout layers to prevent overfitting.
- **Image Classification**: A Convolutional Neural Network (CNN) was used to classify real and fake images.
  - The CNN architecture includes convolutional layers followed by max-pooling layers for feature extraction and downsampling, followed by dense layers for final classification with a sigmoid activation in the output layer.

D. Training and Evaluation

- **Account Classification Model**: Used categorical cross-entropy as the loss function and trained the model with a 90-10 train-validation split. Achieved accuracy improvements across epochs, monitored using validation accuracy and loss metrics.
- **Image Classification Model**: Split data into training, validation, and test sets. Trained the CNN model for 10 epochs, using binary cross-entropy as the loss function. Validation accuracy was tracked to assess model generalization.
- **Evaluation Metrics**: Performance was evaluated using accuracy, precision, recall, and confusion matrices to measure effectiveness.

E. Result Visualization

- Plotted training and validation loss progression to observe learning patterns and detect overfitting.
- Visualized accuracy across epochs for both models to confirm stability in performance improvements.

VII. Results

With an accuracy rate of more than 85 percent in validation testing, the suggested model shows good accuracy in differentiating between real and fraudulent profiles. With a noticeable improvement in recognizing AI-generated profiles, the combination of textual and visual features works better than single-modality features. These findings demonstrate deep learning's potential for use in social security applications.

VIII. Applications and Future Work

Applications for this concept can be found in a number of fields, such as online identity verification, cybersecurity, and social media moderation. By using a wider range of datasets and investigating alternative deep learning architectures, such transformer models, we hope to increase the model's resilience in the future. Future research should also focus on lowering

categorization biases and improving interpretability for a better understanding of model decisions.

IX. Conclusion

Our study shows that deep learning can be an effective method for spotting AI-generated profiles and phony accounts. Accurate classification is made possible by the combination of textual and visual data, which is crucial for improving social security on digital platforms. This strategy may become a regular tool for social media networks to protect user authenticity as deep learning technology advance.

References

[1] Bharat S. Borkar, Dipak R. Patil, Ashok V. Markad, and Manish Sharma. Real or fake identity deception of social media accounts using recurrent neural network. In *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, pages 80–84, 2022.

[2] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C Stamm, and Stefano Tubaro. Timit-tts: A text-to-speech dataset for multimodal synthetic media detection. *IEEE access*, 11:50851–50866, 2023.

[3] Bharti Goyal, Nasib Singh Gill, Preeti Gulia, Om Prakash, Ishaani Priyadarshini, Rohit Sharma, Ahmed J. Obaid, and Kusum Yadav. Detection of fake accounts on social media using multimodal data with deep learning. *IEEE Transactions on Computational Social Systems*, pages 1–12, 2023.

[4] Masnoon Nafees, Shimei Pan, Zhiyuan Chen, and James R Foulds. Impostor gan: Toward modeling social media user impersonation with generative adversarial networks. In *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*, pages 157–165. Springer, 2021.

[5] Sm Zobaed, Fazle Rabby, Istiaq Hossain, Ekram Hossain, Sazib Hasan, Asif Karim, and Khan Md Hasib. Deepfakes: Detecting forged and synthetic media content using machine learning. *Artificial Intelligence in Cyber Security: Impact and Implications: Security Challenges, Technical and Ethical Issues, Forensic Investigative Challenges*, pages 177–201, 2021.

[6] Raphaël Millière. Deep learning and synthetic media. *Synthese*, 200(3):231, 2022.

[7] Umur A Ciftci, Gokturk Yuksek, and Ilke Demir. My face my choice: Privacy enhancing deepfakes for social media anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1379, 2023.

[8] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

[9] Okunola Orogun, Lanre Ogungbe, Niyi Adegboye, Tolu Adetuyi, and Samuel Alabi. Strategies for combating synthetic identity fraud: The role of machine learning and behavioral analysis in enhancing financial ecosystem security.

[10] Xiaoping Zhou, Xun Liang, Haiyan Zhang, and Yuefeng Ma. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE transactions on knowledge and data engineering*, 28(2):411–424, 2015.

[11] Xin Ju, Xiaofeng Zhang, and William K Cheung. Generating synthetic graphs for large sensitive and correlated social networks. In *2019 IEEE 35th international conference on data engineering workshops (ICDEW)*, pages 286–293. IEEE, 2019.

[12] Liam Kearns, Abu Alam, and Jordan Allison. Synthetic media authentication threats: Detection using a combination of neural network and blockchain technology. *Available at SSRN 4658121*.

[13] Mahejabi Khan, Samta Gajbhiye, and Rajesh Tiwari. Fighting fake visual media: A study of current and emerging methods for detecting image and video tampering. In Amit Kumar and Stefan Mozar, editors, *Proceedings of the 6th International Conference on Communications and Cyber Physical Engineering*, pages 545–556, Singapore, 2024. Springer Nature Singapore.

[14] Jon Bateman. *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace., 2022.

[15] Tanusree Ghosh and Ruchira Naskar. Gan and dm generated synthetic image detection in the age of misinformation. In *International Conference on Applied Cryptography and Network Security*, pages 225–229. Springer, 2024.