Deppfake Image Detection Using CNN

Prof. Mamatha A
Assistant Professor
Computer Science and Engineering
Dayananda Sagar Academy of Technology & Management
Bengaluru, India mamatha-

Tejas N Yadav
Student, 4th Year, B.E
Computer Science and Engineering
Dayananda Sagar Academy of Technology & Management
Bengaluru, India

Shashank Kumar E
Student, 4th Year, B.E Computer Science and Engineering
Dayananda Sagar Academy of Technology & Management Bengaluru, India

Jainath Y S
Student, 4th Year, B.E
Computer Science and Engineering
Dayananda Sagar Academy of Technology & Management
Bengaluru, India

Vinuta R
Student, 4th Year, B.E Computer Science and Engineering
Dayananda Sagar Academy of Technology & Management Bengaluru, India

I. ABSTRACT

The authors present a new deep learning architecture designed specifically for Deepfake image detection using CNNs. Their methodology is based on the concept of training a CNN model with a carefully compiled dataset consisting of genuine and fake images, which the authors have extracted very carefully from the Kaggle competition and datasets website. Following the training phase, the model undergoes a critical transfer learning phase, leveraging the powerful representations encoded within the Xception architecture, pre-trained on the vast and diverse ImageNet dataset. The crux of this approach lies in the model's ability to discern intricate patterns and nuanced features that distinguish authentic imagery from its synthetic counterparts. Through the systematic study of underlying structures and pixel- level subtleties, the CNN architecture learns to deconstruct subtle disparities present in Deepfake images and hence, strengthen its predictability. Preliminary results from the experimentation phase show the validity of the proposed CNN-based approach in the task of detecting and reporting the counterfeit image with an appreciable accuracy level. In all fairness, strides have been taken so far. However, this is still something that the entire research community strives to further polish and improve performance metrics of the model. As the iterations continue and the improvements are made, the goal remains to reach unprecedented levels of precision and robustness in Deepfake detection, making our defenses much stronger against falsified multimedia proliferation in the digital space.

Keywords- Deepfake, Facial Forgery, Fake Media
Synthetic media, AI-generated material, pattern recognition image classification, computer vision, data augmentation, overfitting, underfitting, model training, and evaluation.

II. INTRODUCTION

Deepfakes are a new form of artificially manipulated multimedia content generated through advanced neural network algorithms. Deepfakes have emerged as a significant technological

The ability to produce ultra-realistic video and image files based on the deep learning and GANs has emerged through Deepfakes. Through such methods, Deepfakes present a significant source of risks where this technology holds immense potential with a possibility of negative usage.

The increasing sophistication of Deepfake technology has led to its exploitation in several malicious activities such as spreading misinformation, financial fraud, and violations of privacy. The ability to forge realistic digital content undermines the integrity of digital media, erodes public trust, and creates a pressing need for effective detection mechanisms.

However, identifying Deepfakes is a different matter. Techniques for generating them are constantly changing and surpass traditional detection methods. Moreover, the lack of diversified and labeled datasets for training machine learning models further limits the efficiency of the detection systems. Therefore, the field of Deepfake detection has become an important area of research and industry with the need for adaptive and accurate solutions.

This report on the Deepfake Image Detection system makes use of Convolutional Neural Networks. By the strong feature extraction capabilities of advanced architectures in deep learning, such as the pre-trained Xception model, a high accuracy in detection and classification of Deepfake images is promised. The proposed system integrates state-of-the-art methodologies, such as data augmentation, transfer learning, and optimized training strategies, to ensure robust detection across all datasets.

The project's real-world significance spans multiple domains:

- Social media platforms can adopt Deepfake detection tools to curb misinformation and protect user trust.
- Journalists and news agencies can leverage these tools for content verification.
- Legal and forensic fields can ensure the integrity of digital evidence.
- Cybersecurity teams can mitigate risks associated with identity theft and phishing attacks.
- The political and entertainment industries can protect against unethical manipulation and uphold authenticity.

By dealing with the issues related to Deepfake detection, this project is contributing to strengthening the reliability of digital media and countering the threats of this emerging technology. The following sections of the report will explore

The subsequent sections of this report will explore the problem statement, existing solutions, proposed methodology, system architecture, implementation details, testing, and results in relation to the conclusions.

III. LITERATURE SURVEY

Deepfake technology has seen a rise in prevalence, thereby catalyzing extensive research in detection methodologies. This literature review explores key advancements in Deepfake detection, which analyzes various traditional and state-of-the-art techniques. The findings provide insights into the evolution, effectiveness, and emerging challenges of detection strategies, highlighting promising approaches and areas for future research.

1. Comparative Analysis on Different DeepFake Detection Methods and Semi-Supervised GAN Architecture

Deepfake technology proves itself as capable of advanced deep learning methods but poses severe risks, including fraud and misinformation. This paper classifies the methods for feature-based, time-based, and deep feature-based approaches.

Feature-based methods depend on detecting inconsistencies or artifacts at pixel levels.

Temporal-based methods inspect inter-frame correlation in videos. Deep feature-based methods apply deep architectures to draw these patterns. The proposed semi-supervised Generative Adversarial Network (GAN) structure introduces an architecture that makes use of minimal labeled datasets to enhance the detection accuracy. It strikes a balance between supervised and unsupervised learning while addressing the problem of lack of labeled data in Deepfake detection.

2. A Survey on Deepfake Detection Algorithms This paper highlights the social threats arising from the diffusion of Deepfake content, threatening democracy, security, and privacy. It covers detection methods for videos and images, with emphasis on: Strengths and weaknesses of existing models, such as Xception and VGG. Comparison of algorithms based on their ability to generalize over unseen data. The study emphasizes the need for reliable detection methods capable of adapting to evolving Deepfake creation techniques. The comparative insights contribute to the development of robust detection algorithms to mitigate societal harm.

3. Improving Deepfake Detection by Mixing Top Solutions of the DFDC The DeepFake Detection Challenge (DFDC) highlighted the difficulty of achieving high accuracy in identifying manipulated content. The best-performing model in the DFDC, with an 82% accuracy on curated datasets but only at 65% for unseen videos. The paper deals with ensemble techniques including boosting, bagging, and stacking for multi-model fusion in deep learning to boost. It shows that integrating complementary models does improve detection accuracy; it reduces error rates and hence opens up further solutions for reliable solutions.

4. A Heterogeneous Feature Ensemble Learning-Based Deepfake Detection Method The paper here presents a heterogeneous feature ensemble learning approach to overcome the shortcomings of the single-model detection systems.

The method extracts various features, such as gray gradients, spectrum patterns, and textures, which are integrated into an ensemble feature vector.

5. Deepfake Generation and Detection - An Exploratory Study

Deepfakes, created with deep learning algorithms, create ethical and social implications since they allow the convincing manipulation of multimedia content.

The work reviews the state of techniques both for Deepfake creation and detection and offers a balanced view of how they interact with each other.

It further discusses the usage of benchmark datasets such as Face Forensics++ and DFDC for assessing detection performance. The paper focuses on the multi-faceted role of deep learning in creating and detecting Deepfakes and underlines the critical necessity for continuous research efforts to combat growing risks.

IV. ARCHITECTURE AND SYSTEM DESIGN

Based on the proposed task of classifying an image as deepfake or non - deepfake, the model is a CNN architecture tailored for this task. The proposed architecture utilizes the Xception model in particular, which is a pre-trained CNN with its weights being trained on the very extensive ImageNet dataset. Using the Xception model as the base helps establish a more advanced feature-extracting capability in the proposed architecture. The model starts with the convolutional part of Xception model's top layers for preliminary feature extraction, then follows with a sequence of fully connected layers to refine the features extracted. The first fully connected layer has 512 units with ReLU activation to let the network learn intricate nonlinear relationships between the data. To reduce overfitting risk, a dropout layer with 0.5 is added immediately following the initial fully connected layer; dropout deactivates part of the input units at each stage of training according to an unbiased coin flip while forcing the network to produce representations that are stronger and more general. Next, another fully connected layer that contains 128 units and a ReLU activation is added; another dropout of rate 0.5 will be added to the end. This latter dropout helps with further regularization in the model, and finally a fully connected layer consisting of 64 units and activation by ReLU is added on the end. The output layer is a single unit with sigmoid activation, which produces a probability score that indicates the likelihood of an image being classified as a deepfake. The sigmoid activation function constrains the output between 0 and 1, which allows for an interpretable probability interpretation.

V. METHODOLOGY

1. System Objectives

The purpose of the proposed system is to implement a robust deep learning architecture in detecting Deepfake images with a Convolutional Neural Network. The system used a pre-trained Xception model for feature extraction and classification through the use of transfer learning and data augmentation, with the intent to improve generalization and accuracy in the model.

2. Method of Data Gathering Primary Data:

A total of 140,000 images; 70,000 real, 70,000 fake. All images from Kaggle taken at diverse scenarios and lighting conditions and with varying facial expressions to make them robust.

Secondary Data:

Pre-processed and augmented image data that improve the variability in the pattern and try to increase the efficiency of training.

3. Data Preprocessing and Augmentation

Scale all images to pixel range 0–1. Data augmentation technique includes:

Random rotations (-10° to +10°).

Horizontal and vertical shifts of up to 10% of the image size. Shear transformations, zooming, and horizontal flipping of 50%

Probability). 1.
Model Architecture:
Pre-trained Xception Model: This model is pre-trained on the ImageNet dataset for feature extraction:
Entry Flow: This extracts low-level features such as edges and textures. Middle Flow: It captures high-level, abstract features using depthwise separable convolutions. Exit Flow: It aggregates and prepares features for classification. Custom fully connected layers:
1st Layer: 512 units with ReLU activation and 50% dropout. 2nd Layer: 128 units with ReLU and 50% dropout.
3rd Layer: 64 units with ReLU.
Output Layer: It uses a single unit with sigmoid activation for the output of a binary classification - either real or fake.
2. Training and Optimization:
Transfer Learning: Xception model fine-tuned over the Deepfake dataset using custom layers.
Optimization: gradient descent and techniques that minimize the loss to boost performance, with dropout to help the model regularize and not overfit.
3. Evaluation Metrics and Testing
Model evaluation will be on all three - train, validation, and test - datasets. The metrics will include:
Accuracy Precision Recall
F1 Score.
Confusion Matrix is used to evaluate false positives, false negatives, and correct classifications.
Results Obtained:
Training Accuracy: 98.56%
Validation Accuracy: 95.11%
4. Deployment:
The trained model is deployed in a Python-based application using Flask. Real-time image uploads are processed to classify images as real or Deepfake.
5. System Workflow:
Input: Image is uploaded for analysis.
Preprocessing: Image is normalized and augmented. Model Inference: The CNN processes the image to classify it as real or fake.
Output: Probability score showing the chance of being a Deepfake.
6. Hardware and Software Integration:
Hardware:
• GPU: Nvidia GeForce RTX 3070 or similar for model training.
• Processor: Intel i5 minimum or AMD equivalent. Software:
• Python 3.x with libraries such as TensorFlow, Keras, and OpenCV.
• Flask for application deployment.

VI. EXPERIMENTATION AND RESULT

Experimentation

The evaluation results yield strong validation of the proposed deepfake detection model, where its effectiveness for deepfake images was confirmed, having achieved precise accuracy in correctly classifying such images. Both "real" and "fake" classes saw very impressive precision, recall, and F1-scores from 0.94 to 0.95. These metrics indicate how well the model correctly identifies true positives, or correct classifications of "real" or "fake," minimizing both false positives, where real images are misclassified as fake, and false negatives, where fake images are misclassified as real.

The confusion matrix is a useful tool for evaluating the performance of a classification model, such as a deepfake detection model, as it provides a detailed breakdown of the model's predictions compared to the actual labels in the test dataset.

This is how the confusion matrix usually looks:

- True Positives (TP): Instances where the model correctly predicts a sample as positive (e.g., correctly identifying a deepfake image as "fake").

- True Negatives (TN): Instances where the model correctly predicts a sample as negative (e.g., correctly identifying a real image as "real").

- False Positives (FP): Instances where the model incorrectly predicts a sample as positive when it is actually negative (e.g., incorrectly classifying a real image as "fake").

False Positives (FP): The cases in which the model inaccurately predicts a sample to be positive when it is actually negative, for example, incorrectly classifying a fake image as "real."

The confusion matrix organizes these predictions in a matrix format that provides a clear and concise summary of the model's performance across different classes.

Interpreting the confusion matrix requires analyzing the distribution of predictions within each quadrant and deriving meaningful insights into strengths and weaknesses of the model. For example, a balanced distribution of true positives and true negatives may suggest that the model correctly classifies deepfake and non-deepfake image. False positives or false negatives tend to inform and suggest areas where a model needs improvement in sensitivity or specificity

Results:

It also achieved an overall accuracy of 95% on the test dataset. This high accuracy shows that the model is rugged and reliable enough in differentiating between deepfake and non- deepfake images. The performance of the model on unseen data not used in training or validation further confirms its generalization capabilities and effectiveness in real- world scenarios.

The above results clearly demonstrate that this model is potentially very applicable for real-world implementations, like moderation in social media content, verification of news articles, and cyber security. Thus, this model's correct identification of deepfake content serves the broader objective of efforts in mitigating the spreading of misinformation and fraud using altered media. On the whole, the above evaluation results demonstrate significant advancement in techniques to detect deepfakes.

VII. CONCLUSION

A deep learning-based approach for predicting deepfake images using CNNs was proposed in the research paper. The CNN architecture was trained on a dataset comprising real and fake images. Utilizing transfer learning, the researchers applied the Xception model pre-trained on the ImageNet dataset to enhance the model's ability to discern patterns and features specific to each class. This means the study had shown promising results in terms of fake image prediction using the approach based on CNN. However, there is more to do since it may improve the results even further with further improvements. One possible improvement is to test the dataset given using other pre-trained models available, such as Xception. This way, one can make a comparative study about the results that will show which model works the best for this task of predicting deepfakes. The study can also guide the researcher with respect to where they might require a stronger and weaker pre-trained model for better application in that specific model architecture. Further, the researchers have suggested that generalization of the model could be improved by trying to obtain samples from multiple sources. Increasing the diversity of the dataset includes using images from multiple sources and contexts. This strategy will likely result in a much more robust and reliable deepfake detection system that is able to accurately capture manipulated media across all manner of different scenarios and environments.

In summary, the proposed research is to improve deepfake detection techniques by investigating alternative pre-trained models and enhancing generalization of the CNN-based approach through incorporation of diverse datasets. The comparative studies and dataset expansion will help improve the performance and reliability of deepfake prediction models, which can contribute to ongoing efforts to combat misinformation and fraudulent activities in digital media.

the performance and reliability of deepfake prediction models, which can contribute to ongoing efforts to combat misinformation and fraudulent activities in digital media

VIII. REFERENCES

1. Dr C Nandini ,Antara Mukherjee, Bhoomika “Smart Health Prediction System Using Machine Learning Techniques.”, International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.10, Issue 4, pp.e73-e78, April 2022, Impact factor: 7.97 – Q4
2. Nagaraj M Lutimath, Jhanavi Oza, Khushi NB, Maithili Joshi, Prarthana P, “Brain Tumor Classification Using Deep Learning Technique”, International Journal of Research and Analytical Reviews (IJRAR), May 2023, Volume 10, Issue 2, www.ijrar.org (E-ISSN 2348-1269, P- ISSN 2349-5138), pp. 978-985.
3. Nagaraj M. Lutimath, Chandra Mouli, B. K. Byre Gowda, K. Sunitha, “Prediction of Heart Disease Using Hybrid Machine Learning Technique”, Springer Book Series "Transactions on Computer Systems and Networks”, with title "Paradigms of Smart and Intelligent Communication, 5G and Beyond", Springer Nature, Singapore,2023, pp 277–293.
4. Priya Nandihal, Vijay Shetty, Tapas, Piyush Pareek “ Giloma Detection Using Improved Artificial Neural Network in MRI Images”, 2022 IEEE 2nd Mysore Sub Section International Conference(MysuruCon), 2022, pp 1-9.
5. Sumanth Reddy, S., D R, M., S, J., & C, N. (2024). A Comprehensive Review of Machine Learning Approaches in Livestock Health Monitoring. *Journal of Big Data Technology and Business Analytics*, *3*(3), 11–19.
6. D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.
7. X. Chang, J. Wu, T. Yang and G. Feng, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 7252-7256.
8. Huaxiao Mo, Bolin Chen, and Weiqi Luo. 2018. Fake Faces Identification via Convolutional Neural Network. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '18). Association for Computing Machinery, New York, NY, USA, 43–47.
9. Deepfake Video Detection through Optical Flow Based CNN, Irene Amerini, Leonardo Galteri, Roberto Caldelli, Alberto Del Bimbo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 0-0.
10. Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee. 2020. "Deep Fake Image Detection Based on Pairwise Learning" Applied Sciences 10, no. 1: 370.
11. Hasin Shahed Shad, Md. Mashfiq Rizvee, Nishat Tasnim Roza, S. M. Ahsanul Hoq, Mohammad Monirujjaman Khan, Arjun Singh, Atef Zaguia, Sami Bourouis, "Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network", Computational Intelligence and Neuroscience, vol. 2021, Article ID 3111676, 18 pages, 2021.