Fake Vision Detection using CNN

Joy Debnath, Vinayak Jivtode , Aditya Ellandula,
Dhananjay Khairnar, Sanket Gajbhiye
Computer Science and engineering , students
Rajiv Gandhi College of Engineering Research and
Technology Chandrapur

Dr. Vanita Buradkar
Assistant Professor
Rajiv Gandhi College of Engineering Research and
Technology Chandrapur

***Abstract*- The rapid development of synthetic media techniques, particularly deepfakes, the integrity of visual content is increasingly at risk as deep learning models generate highly realistic manipulated videos or images that are often indistinguishable from authentic media. To address this challenge, advanced deepfake detection systems leverage convolutional neural networks (CNNs) and forensic image analysis to identify manipulated content by examining subtle spatial and temporal artifacts introduced during synthesis. These systems analyze inconsistencies in facial expressions, unnatural movements, and irregularities in lighting or synchronization, combining machine learning with digital forensics to enhance accuracy and robustness across diverse deepfake datasets.**

Keywords - detection of deep, convolutional neural networks, forensic image, Computer Vision, Generative AI, Deep Learning, Multimedia Forensics

I. INTRODUCTION

In recent years, the rapid progress of artificial intelligence - especially in generative models such as generative contradictory networks (GAN) and diffuse models - allowed the creation of highly realistic synthetic content. The most important applications of these technologies include Deepfakes, which are images generated by AI, videos or sound clips that convincingly mimic real people or events. While Deepfakes can be used creatively in entertainment and education, there are also serious risks, including disinformation, digital prese, political manipulation and reputation damage.

Deepfakes detection has therefore become an urgent challenge in computer vision and digital forensic. Unlike traditional photo editing, fake
images generated by AI often contain imperceptible artifacts or inconsistencies that are difficult to detect for the human eye. However, these fine traces can be used using machine learning models that distinguish real images from fake.

The aim of this research is to develop a Deepfake detection system that can accurately identify the images generated by AI by accessing deep learning. Using convolutional neural networks (CNN) and learning techniques, the system is designed to detect fine -grained differences and formulas that indicate synthetic content. The model is trained and tested using a combination of real and deep images from publicly available data sets, which ensures a variety of and demanding evaluation environments. A deep approach based on lea Comparative evaluation of model performance on Benchmark data sets Deepfake. learning to detect fake images generated by AI. Analysis of visual features and artifacts common in synthetic images. Through this work, we try to contribute to the development of trusted AI systems agnnd promotes efforts to combat abuse of generative technology.

II. RELATED WORK

Deepfake technology has led to significant research in the development of methods to detect synthetic media. Various techniques were designed, from traditional machine learning approaches to advanced architecture of deep learning. This part summarizes key contributions in the Deepfake image detection.[1]

1.CNN based detection method

One of the first approaches to Deepfake detection used a convolutional neural network (CNN) to extract functions from facial images. Afchar et al. [3] (2018) designed the architecture of Mesonet, which works on mesoscopic facial elements to

The key contributions of this work are as follows: distinguish the real from false faces. Similarly, nguyen et al.[4] (2019) introduced capsule networks (Capsnet) for better learning spatial relations between facial components.

2. Analysis of frequency domains and artifacts

Several works focused on the detection of artifacts that Gans in the frequency domain. Durall et al.[5] (2020) have observed that GAN -generated images often show abnormal frequency formulas and trained CNN on Fourier transformed images to improve classification. Frank et al. (2020) used spectral analysis to detect irregularities in the process of generating Gans, such as Stylegan.

3. Transfer of learning and predetermined networks

Many scientists used transmission learning to improve the performance and shorten the training period. Models such as Resnet50, EfectNet and VGG16 pre -preliminary on Imagenet were tuned on Deepfake data sets. Wang et al. (2020) have shown that the fine tuning of such models can achieve high accuracy across several types of depth.

4. Mechanisms and transformers of attention

Newer methods have explored the use of attention -based models and transformers

(VITS) to detect deep resolution. These models focus on fine areas of manipulation, offer greater interpretability and often higher performance than CNN itself. Liu et al. (2021) used a combination of CNN mechanisms and self knowledge to detect videos with a deep time consistency.

5. Data sets and benchmarks

Several reference data sets have been released to support Deepfake research. Facefrorensics ++ data file is widely used and contains real and manipulated videos/images with different levels of compression. Data file for Deepfake Detection

Challenge (DFDC), which release Facebook AI, provides a large and diverse set of deep videos for training and evaluation. Celeb-DF is another high quality data file aimed at improving generalization in the real world scenarios.

These studies significantly contributed to the understanding and solving the problems of deep detection. However, many existing models suffer from poor generalization across data sets or are too dependent on specific handling techniques. This article addresses these limitations by implementing robust architecture based on theCNN with transmission learning to better generalize across different synthetic image sources.

III. PROPOSED SYSTEM

The proposed system is a deep frame -based framework aimed at detecting fake images generated by AI by analysing visual features that are often imperceptible to the human eye, but can learn it with neural networks. The main idea is to use a convolutional neural network (CNN) by improved transmission learning and classify images as real or deep. Architecture is trained on a curator's data set consisting of authentic images and synthetic pictures generated using GAN based models such as Stylegan and Deepfake
Models.

The system architecture consists of the following main components:

Dataset Preprocessing real and fake images of images are collected from sources such as Face Forensics ++, DFDC and "This person does not exist". Images change and normalize. They are applied to increase generalization.

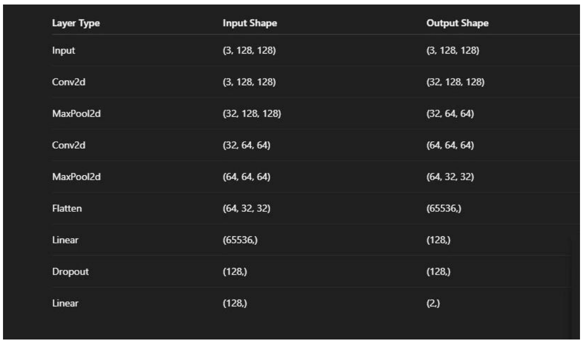
Extraction of elements using re -evaluated CNN. The CNN preliminary model (e.g. Resnet50 or Efectnet) is used as the spine. The convolutional layers of the model extract spatial properties and artifacts from input images. Transfer Learning helps to use the learned features from large data sets (eg Imagenet) and adapt them to Deepfake detection.

The classification layer of the last layer is adapted to fully connected layers, discharge and output layer Sigmoid/Softmax for classification of images as real or fake. Binary cross entropy is used as a loss function and Adam optimizer is used for training.

The model training and evaluation of the model is trained on the marked data file using GPU acceleration. The performance is evaluated by accuracy, accuracy, memories, F1-skore and Roc-Auc metrics. For the analysis of false positives and false negatives, a matrix of confusion is generated.

Visualization and explaining (optional) GRADCAM maps or prominent maps are used to highlight image areas that affect the model's decision and provide interpretability.

Architecture Details



Layer Type	Input Shape	Output Shape
Input	(3, 128, 128)	(3, 128, 128)
Conv2d	(3, 128, 128)	(32, 128, 128)
MaxPool2d	(32, 128, 128)	(32, 64, 64)
Conv2d	(32, 64, 64)	(64, 64, 64)
MaxPool2d	(64, 64, 64)	(64, 32, 32)
Flatten	(64, 32, 32)	(65536,)
Linear	(65536,)	(128,)
Dropout	(128,)	(128,)
Linear	(128,)	(2,)

Fig. no.1 Architecture Details

The proposed Deepfake detection system takes advantage of transfer learning using a contracted Neural Network (CNN) architecture, using a preparatory model such as the Resanet50 or Epitheted. The input of the model is a color facial image that is shaped for a certain dimension of 224 × 224 × 3. Initially, the image pretends to undergo several determinations of the network, where spatial features such as edges, textures and patterns are extracted. After these layers, batch normalization and Relu activation functions are followed to increase training stability and introduce non-linearity respectively. The maximum-pooling layers are employed within the architecture to reduce the spatial dimensions progressively while maintaining the necessary features.

The high-level feature maps obtained from the final conversion block are passed through a global average pooling layer, which compresses spatial information into a single feature vector. This is followed by one or more completely connected (dense) layers, in which relay activation and dropouts are reduced overfitting with regularization. The last layer is a single-nod dense layer with a sigmoid activation function that outputs a probability score that indicates that the input image is real or fake. This allows the CNN based design model to effectively capture and separate the microscopic artifacts launched during the generation of deepfake images.

IV. WORKING

The proposed **deepfake recognition** system follows a systematic pipeline that processes input images and classifies **deep learning techniques** as real or **fake.** Below is a step-by-step **declaration** of how the system **works.**

The system **occupies the face** as input. This image can be **delivered** from social media, **data records,** surveillance systems, or **user-friendly** content. **Images can** be **generated authentically** or synthetically using AI **(such as goose).**

Several preprocessing steps are **used** to **prepare images of deep learning models.**

Size of fixed **dimensions (e.g. 224** x **224** pixels) compatible with **CNN models. Generalization of** the **model.** These **properties** include subtle textures, pixel **mismatches,** and visual artifacts introduced by AI-based image generation **techniques [6].** The final layer **assigns probability ratings using the** sigmoid activation function (for binary **classification). Predictive Output The system classifies images based on** a final probability **rating** and returns the **results. If** the score is **less than** 0.5, the image is **actually classified.**

Block Diagram

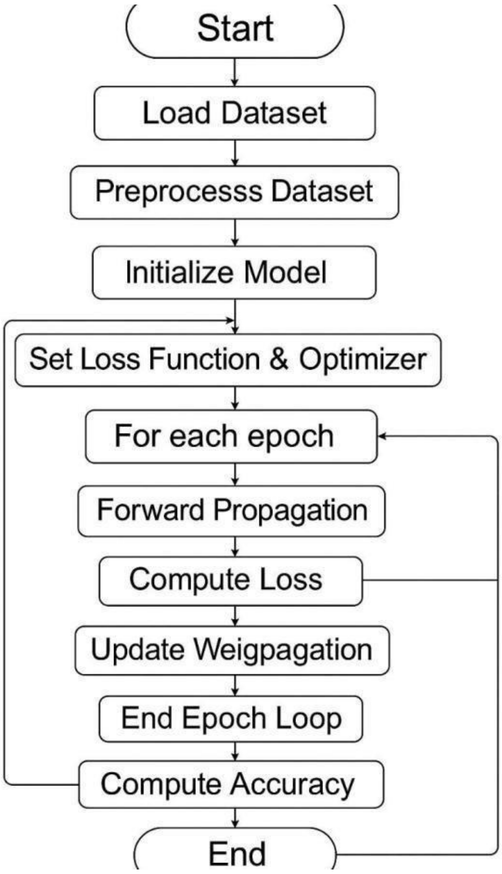


Fig. no.2 Block Diagram

V. ADVANTAGES

High accuracy in detecting the use of CNN preliminary models, such as resnet50 or EfficientNet [7], improves the ability to detect fine artifacts in fake images, resulting in high classification accuracy.

Robust extraction of deep learning elements automatically learns important visual patterns and non -consistency, which may not be easily detectable through manual engineering functions.

Effective training with transmission learning using preliminary models requires a system of less data and calculation to achieve efficient performance.

Adaptability of multiple sources of deep sources the system can be trained on a combination of Deepfake data files [9] and is able to detect images generated by different GAN architectures (eg Stylegan, Deepfake, FaceSWAP).

Framework scalability can be expanded to support Deepfake video detection or integrated into real time systems (eg social media platforms, content moderation tools).

To visualize image areas that affect the model's decision, users can understand why a particular image is marked as fake tools of explanation support, such as Grad-CAM.

Generalization to the real world scenarios using different data sets and augmentation techniques is trained to resist various properties of image, lighting and facial expressions.

Reduced human effort automates the deep dimension detection process and reduces the need for manual review and time savings in forensic analysis or content verification.

VI. DISADVANTAGES

Limited generalization to the invisible deep resolution technique, while the system works well on known data files, can try to detect images generated by new or invisible deep introduction methods that introduce new artifacts or hide existing methods.

The dependence on the data set The accuracy and performance of the system strongly depends on the quality, diversity and size of the training data set. Distorted or unbalanced data sets can lead to poor generalization.

High computational requirements of deep CNN models, especially with large data sets and transmission learning, require significant computing sources (GPU/TPUs), which may not be accessible to all scientists or organizations.

Vulnerability to contradictory attacks models with deep learning can be susceptible to opponents - images that are slightly modified to deceive the detector to perform incorrect classifications.

The limitation of the binary classification The current model is limited to distinguishing only between "real" and "false" classes. No classifies the type of deep or specific generation method.

The model could overcome the risk of excessive ideas without careful regularization and verification, can overcome data set of training and reduce its performance on real or invisible data.

Restrictions on explaining, although tools such as Grad-CAM, can provide some interpretability, decisions made by deep neural networks can still be considered a "black box" in nature, which can cause problems with confidence in sensitive applications.

Focusing on the face Most of the Deepfake data sets are focused on the face. As a result, the system may not work well on idle deep or other forms of the synthetic medium (eg background, objects).

VII. APPLICATIONS

Platforms to moderate social media and content, such as Facebook, Instagram and Twitter, can integrate the system to automatically detect and indicate deep images before they are publicly shared to prevent disinformation and manipulated media. [9]

Digital forensic coercive organs and digital forensic analysts can use the system to verify the authenticity of images presented as evidence in investigation or legal cases.

Media and media verification organizations can implement the model as a tool to verify the authenticity of the user's content and protect against false visual content that could damage credibility or disseminate false information.

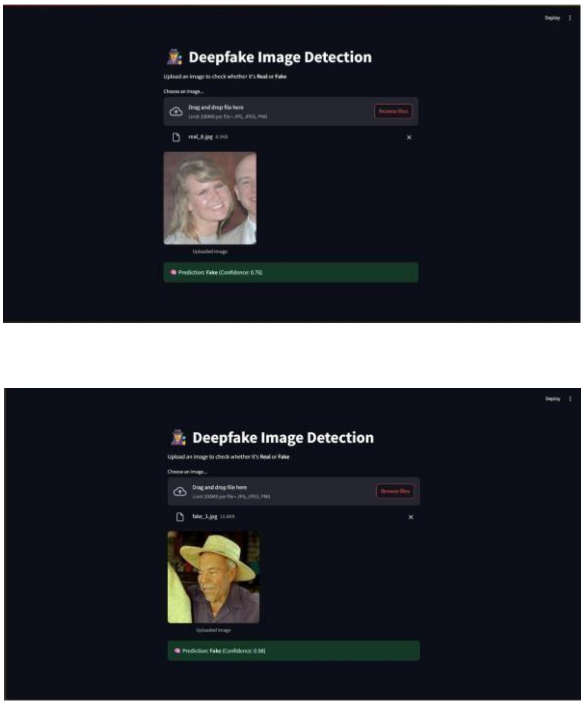
Identity protection the system can help prevent identity theft by detecting synthetic images used in fraudulent documents, fake profiles, or deceit fading.

Deepfake detection Online dating and electronic trading can be used to verify user profiles and ensure that recorded photos are real, reducing the risk of cats and fraud.

This technology can be used to identify manipulated images of supervision or propaganda that could pose threats to public security and national security in the critical infrastructure and national defines industry.

Educational tools can be used in academic and training for teaching students and experts on the manipulation of the media, cyber security and the ethics of the AI content. [10].

VIII. RESULTS



Deepfake Image Detection

Upload an image to check whether it's **Real** or **Fake**

Choose an image...

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

real_8.jpg

Uploaded image

Prediction: **Fake** (Confidence: 0.76)

Deepfake Image Detection

Upload an image to check whether it's **Real** or **Fake**

Choose an image...

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG

fake_1.jpg

Uploaded image

Prediction: **Fake** (Confidence: 0.98)

Fig. no.3 Results

IX. CONCLUSION

The growing sophistication of synthetic images generated by AI, commonly known as Deepfakes, represents significant risks for digital integral, privacy and public confidence. This research represents a deep approach based on the detection of false images generated by AI using a convolutional neural network (CNN) reinforced by transmission learning. The proposed system learns that it uses preliminary models and trains them on real and synthetic data sets that effectively learn to identify fine artifacts and non -consistency characteristic of Deepfakes.

Experimental results show that the system achieves high accuracy and robustness across different data sets and handling techniques. The model not only automates the detection process, but also has potential applications in different areas such as digital forensic, social media moderation and biometric verification. While the system shows a promising performance, it is not without restrictions.

Problems such as generalization to invisible methods of deep, opponents and computing requirements remain challenges that need to be addressed. Future work can focus on incorporating more diverse data sets, improving the model explaining and expanding its ability to detect handled videos or sound along with pictures.

Overall, this work contributes to the ongoing efforts in the area of AI and media forensic by providing a scalable and effective solution to the growing threat of deep.

REFERENCES

[1] Y. Li, M. Chang and S. Lyu, "In ICTU Oculi: Exposure AI generated fake face videos by detecting eye flicker," IEEE International Workshop on Information Forensic and Security (WIFS), 2018.

[2] Divya Babu and Uppala Santosh Kumar “Deepfake Video Detection using Image Processing and Hashing Tools” 2020 International Research Journal of Engineering and Technology (IRJET) Mar 2020, | ISSN 2395- 0056.

[3] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "Mesonet: Compact Network Detection Detection on the Face" in Proc. IEEE WIFS, 2018.

[4] H. Nguyen, J. Yamagishi and I. Echizen, "Using Calls networks to detect fake images and videos," Arxiv Pritrint Arxiv: 1910.12467, 2019.

[5] R. Durall, M. Keuper and J. Keuper, “Watch your generative deep neural networks based on CNN: IEEE/CVF conference on computer vision and pattern recognition), 2020.

[6] J. Frank, L. Eisenhofer, A. Rössler, C. Riess and M. Nießner, "Frequency Analysis for deep false recognition", International Conference on Machine Learning (ICML), 2020.

[7] X. Wang et al. Z. Liu, P. Luo, X. Wang and X. Tang,

[8] "The attributes of deep learning in the wild", IEEE International Conference on Computer Vision (ICCV), 2015.

[9] Facebook AI, "Data File Detection Detection Challenge" [online].

[10] https://www.kaggle.com/c/deepfake-detectio n-challenge

[11] K. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales and J. Ortega-Garcia, "Deepfakes and Beyond: Survey of Facial Manipulation and False Detection", Information Fusion, Vol. 64, pp. 131–148, 2020.