A Survey on Deepfake Detection Techniques using Deep Learning and Convolutional Neural Networks

Prof. Rahul V Dagade, Narayan Ekhande, Vishvajeet Chandanshiv, Radha Lohar
Smt. Kashibai Navale College of Engineering, Vadgaon, Pune, India

Abstract: *In today's digital landscape, the proliferation of deepfake technology—synthetically altered media generated using deep learning—has raised critical concerns regarding the authenticity and integrity of visual content. Deepfakes pose threats in domains such as misinformation, political manipulation, identity theft, and digital fraud. This study presents a CNN-based deepfake detection system capable of identifying manipulated images and videos by analyzing spatial inconsistencies introduced during synthetic content generation. The model is trained on labeled datasets and evaluated using standard metrics including accuracy, precision, recall, and F1-score. A user-friendly GUI is developed using Tkinter to facilitate interactive real-time detection. By integrating robust machine learning techniques with practical implementation, this work contributes to ongoing efforts to safeguard digital trust.*

Keywords: Deepfake Detection, CNN, Digital Media Forensics, Image Manipulation, Misinformation, Deep Learning

I. INTRODUCTION

1. The exponential growth of artificial intelligence, particularly in the fields of deep learning and generative modeling, has given rise to advanced techniques capable of synthesizing highly realistic fake media—commonly known as deepfakes. These artificial videos or images are generated primarily using models such as Generative Adversarial Networks (GANs), which can manipulate or replace human faces, voices, and actions in a way that is often indistinguishable to the human eye. While these technologies offer innovative applications in entertainment, education, and virtual reality, they also present significant risks, including misinformation propagation, defamation, political manipulation, and digital impersonation.

2. The deceptive realism of deepfakes poses a major threat to public trust in visual content. The ability to fabricate convincing digital media undermines the reliability of online information, challenging societal norms around truth, evidence, and accountability. As a result, the development of robust and scalable detection systems has become a critical area of research in computer vision and multimedia forensics.

3. Among the various approaches to detecting deepfakes, Convolutional Neural Networks (CNNs) have demonstrated significant promise due to their ability to learn spatial hierarchies of features in visual data. These models can detect subtle anomalies and inconsistencies—such as unnatural facial movements, lighting mismatches, or compression artifacts—introduced during synthetic media generation. Detection systems typically involve stages like dataset collection, preprocessing, feature extraction, model training, and evaluation.

4. This survey explores existing research on deepfake detection with a specific emphasis on CNN-based techniques. It reviews the architecture, methodology, strengths, and limitations of various proposed systems while highlighting the challenges posed by increasingly sophisticated generative models. The paper aims to offer a comprehensive understanding of the current landscape and future directions in deepfake detection technologies.

DOI: 10.48175/568

II. RELATED WORK

A literature survey consists of different learning techniques to retrieve ontology from data as follows:

With the advent of AI-driven generative technologies, deepfakes have emerged as a modern threat to digital media authenticity. Several researchers have approached this issue with the goal of improving detection accuracy, robustness across datasets, and adaptability to unseen manipulation techniques. This section presents a detailed overview of notable research efforts and existing methodologies related to deepfake detection, particularly focusing on visual content manipulated through AI-based methods.

The work of Deng Pan, Lixian Sun, Rui Wang, Xingjian Zhang, and Richard Sinnott examines the use of deep learning classifiers for deepfake video detection. Their study explores two powerful CNN models—Xception and MobileNet—applied to the FaceForensics++ dataset. These models are well-known for their feature extraction capabilities in image classification tasks. The authors focused on classifying fake vs. real media by learning subtle facial irregularities introduced during synthetic content creation. Their study demonstrates high classification accuracy, highlighting that transfer learning on specific datasets can significantly boost detection performance. However, they also noted challenges in generalizing these models to deepfakes created with different algorithms or production techniques not seen in training data.

John Lewis, Imad Eddine Toubal, and Helen Chen proposed a multimodal deep learning framework, emphasizing the detection of inconsistencies in the spatial, spectral, and temporal domains of deepfake videos. This research took a more layered approach by combining information across frames, rather than relying on single-frame classification, which is common in earlier CNN-based approaches. Their system integrates Discrete Cosine Transform (DCT) techniques to extract frequency domain features, which are often difficult for generative models to replicate authentically. The result is a hybrid detection method that captures both short-term and long-term anomalies across frames. Their paper notes that although spectral inconsistencies can be subtle, they provide a unique signal that improves classification reliability—especially in low-quality or compressed video files.

Expanding the scope of deepfake detection into audio forgeries, Shwetambari Borade and team applied a Support Vector Machine (SVM) classifier, trained on Mel-Frequency Cepstral Coefficients (MFCCs) derived from the Fake-or-Real (FoR) dataset. The system they developed achieved an impressive 97.28% classification accuracy, suggesting that even traditional machine learning techniques, when paired with meaningful and well-engineered features, can perform strongly. Their work demonstrates that while deep learning dominates image and video-based detection, audio detection may still benefit from lightweight, interpretable models with reduced training complexity.

Another important contribution comes from Asad Malik and Minoru Kuribayashi, who conducted a comprehensive survey of face-based deepfake detection methods. Their work categorized detection techniques into various families, including CNN-based, GAN-based, and hybrid approaches. They reviewed deepfake generation methods such as FaceSwap, Face2Face, and NeuralTextures, and matched them with detection strategies employed in various studies. The survey emphasized that a key gap in current research is the lack of generalized detection frameworks—most models perform well on specific datasets but fail when exposed to novel deepfake types. They also discussed the importance of dataset diversity and the evolution of benchmark datasets in keeping pace with advances in deepfake generation.

Paloma Cantero-Arjona and Alfonso Sánchez-Macian addressed the issue from a computational resource standpoint, recognizing that many detection systems are too demanding for deployment on mobile or edge devices. Their research evaluated the performance of CNN-based models in low-resource environments, exploring trade-offs between accuracy and model size. They suggest that efficient architectures—such as MobileNet or quantized CNNs—are more suitable for real-time, device-level deployment, especially in contexts like media verification apps or law enforcement tools where cloud processing may be unavailable or insecure.

The paper by Prof. Aparna Bagde and colleagues emphasizes the real-world impact of deepfakes and reviews various challenges and detection mechanisms through a societal lens. Their work elaborates on how deepfake videos have been used for cyber harassment, fake news dissemination, and character assassination. They explain that CNN-based detection models are particularly suited to identifying localized facial distortions, which often go unnoticed by the human eye. They also argue that the development of user-accessible detection tools is necessary to empower common users and journalists to verify content credibility independently.

Another extensive review by Neeraj Guhagarkar et al. looked into a variety of detection techniques ranging from traditional machine learning approaches like SVMs to more complex neural networks such as CNNs integrated with LSTM layers. Their review was unique in that it highlighted the temporal dynamics of facial movements as a critical component of deepfake detection, especially when detecting videos where subtle inconsistencies in eye blinking or lip-sync can reveal tampered content. The inclusion of LSTM components helps in learning these sequential patterns more effectively than static frame-based models.

Siwei Lyu's work adds to the discussion by highlighting the existing challenges and future directions in deepfake detection research. He identified several limitations in current systems, such as poor cross-dataset generalization, lack of interpretability, vulnerability to adversarial attacks, and computational overhead. Lyu proposes that multimodal approaches, integrating visual, audio, and behavioral signals, may hold the key to building resilient detection frameworks capable of evolving with the technology. His recommendations point toward a future where deepfake detection will need to keep pace not only with technological changes but also with ethical, legal, and societal implications.

In summary, the body of research in deepfake detection is rich and diverse, encompassing a wide range of techniques and application domains. While CNNs remain a foundational tool due to their strength in spatial feature extraction, it is evident that multimodal, hybrid, and efficient models are the way forward. With deepfake technologies advancing rapidly, future detection methods must focus not only on accuracy but also on real-world adaptability, interpretability, and ethical deployment.

III. PROPOSED WORK

The proposed work of the study including the dataset description, the proposed methodology and algorithmic detail:

The proposed work aims to develop a robust deepfake detection system leveraging deep learning techniques, specifically Convolutional Neural Networks (CNNs), to identify manipulated facial images and videos. The goal is to build an efficient, scalable, and user-friendly framework capable of accurately distinguishing between real and fake content, addressing the increasing challenges posed by sophisticated deepfake generation methods. This section elaborates on the dataset preparation, preprocessing pipeline, system architecture, algorithm design, and objectives underpinning the detection framework.

A. Dataset Description

Data Collection:

The foundation of the proposed system is a carefully curated dataset comprising real and deepfake facial images and video frames. Multiple publicly available datasets were employed, including FaceForensics++, Deepfake Detection Challenge (DFDC), and Celeb-DF. These datasets encompass a variety of deepfake generation techniques such as face swapping, lip-syncing, and neural rendering, ensuring the system learns to detect a wide spectrum of manipulations.

Data Annotation and Labeling:

Each sample within these datasets is annotated with binary labels: real or fake. For videos, individual frames were extracted and treated as separate image instances, enabling frame-level detection granularity. This approach allows the detection system to identify fake segments even if only a portion of the video is manipulated.

Class Distribution:

To ensure balanced learning and prevent bias towards any class, equal representation of real and fake samples was maintained during training. This balance mitigates overfitting and enhances the model's generalization capability across diverse manipulation types.

B. Preprocessing Techniques

Robust preprocessing is critical to enhance model performance and ensure the quality of inputs. The following preprocessing steps were applied systematically:

• Face Detection and Cropping:

Each image or video frame undergoes face detection using MTCNN, isolating the facial region to remove background clutter and irrelevant content, focusing the model on pertinent features.

• Image Resizing:

Detected face regions are resized uniformly to 224×224 pixels, conforming to the input size requirements of the CNN architecture.

• Normalization:

Pixel intensity values are normalized to a range of [0,1] to standardize the input data, facilitating stable and faster convergence during training.

• Data Augmentation:

To improve model robustness and prevent overfitting, augmentation techniques such as random horizontal flips, rotations (up to ±15 degrees), zooming, and brightness adjustments were applied. This increases the effective size of the training dataset and introduces variability.

• Frame Sampling for Videos:

Video frames were extracted at regular intervals (e.g., every 5 frames) to reduce redundancy and ensure efficient training without losing temporal diversity.

These preprocessing steps collectively improve the model's ability to learn discriminative features while mitigating noise and irrelevant variation in the data.

C. Proposed System Architecture

The proposed system architecture integrates multiple functional modules to achieve seamless deepfake detection:

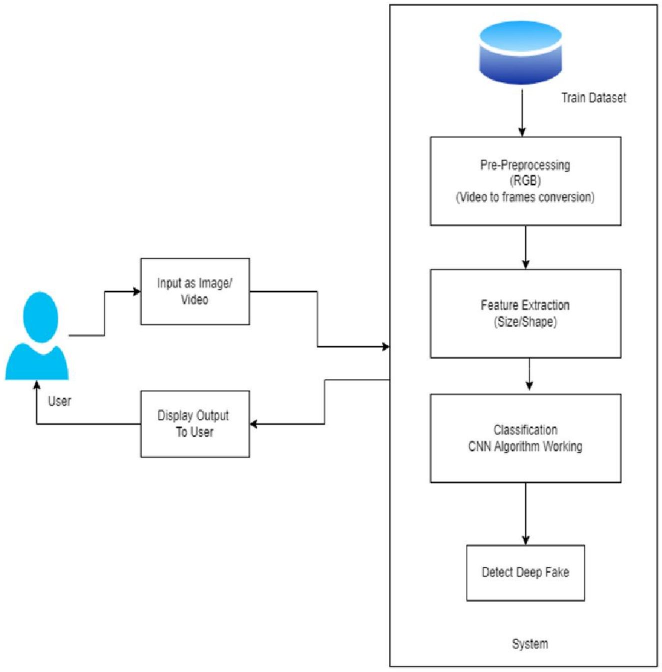


Fig: 3.1. System Architecture :

• User Interface Module:
A Tkinter-based graphical user interface (GUI) enables users to upload images or videos for analysis. The GUI is designed for ease of use, catering to both technical and non-technical users by providing straightforward interaction and real-time feedback.

• Feature Extraction Module:
At the core lies a custom-designed CNN model consisting of several convolutional layers with ReLU activations, interspersed with max-pooling layers to progressively extract spatial hierarchies of features. The network ends with fully connected layers and a sigmoid output neuron to predict the binary classification of input samples.

• Model Training and Validation Module:
The CNN is trained with a binary cross-entropy loss function, optimized using the Adam optimizer. The dataset is split into 80% training and 20% validation subsets to monitor learning progression and avoid overfitting. Evaluation metrics such as accuracy, precision, recall, and F1-score are tracked to assess performance rigorously.

• Prediction and Output Module:
Upon inference, the system produces a confidence score indicating the likelihood of the input being fake. For videos, frame-wise predictions are aggregated via majority voting or averaging to provide an overall classification, improving robustness against transient misclassifications.

D. Algorithm Overview

The proposed algorithm operates through the following procedural stages:

1. Data Loading and Preprocessing:
Load the real and fake image/video datasets, apply face detection, resizing, normalization, and augmentation to prepare standardized input tensors.

2. CNN Model Definition:
Construct the CNN architecture with convolutional layers for feature extraction, followed by fully connected layers for classification, concluding with a sigmoid activation for binary output.

3. Training:
Train the CNN using the preprocessed dataset, minimizing binary cross-entropy loss via the Adam optimizer. Monitor validation performance to tune hyperparameters and avoid overfitting.

4. Evaluation:
Validate the trained model on unseen data, computing metrics such as accuracy, precision, recall,and F1-score to quantify detection efficacy.

5. Inference:
Accept input images or videos from the GUI, preprocess them similarly, and classify using the trained model. For videos, aggregate frame-wise results.

6. Result Presentation:
Display the prediction and confidence score to the user through the GUI, facilitating informed decisions.

E. System Objectives and Advantages

• High-Accuracy Detection:
Exploit CNNs to capture subtle visual artifacts and inconsistencies indicative of deepfake manipulations, achieving robust detection accuracy.

• User Accessibility:
Provide a simple, intuitive GUI for easy deployment and adoption by end-users, including those without technical expertise.

• Real-Time Performance:
Optimize model and preprocessing pipeline for real-time or near real-time inference on images and short videos.

DOI: 10.48175/568

- **Modular and Scalable Design:**

Architect the system to allow future enhancements such as temporal analysis, audio-visual fusion, or integration with social media platforms for content moderation.

- **Baseline for Further Research:**

Establish a foundational framework that can be extended or fine-tuned as new deepfake generation methods emerge and datasets evolve.

IV. RESULTS AND DISCUSSIONS

The results and discussions involve the accuracy, loss noticed in the model:

1. Model Accuracy Analysis

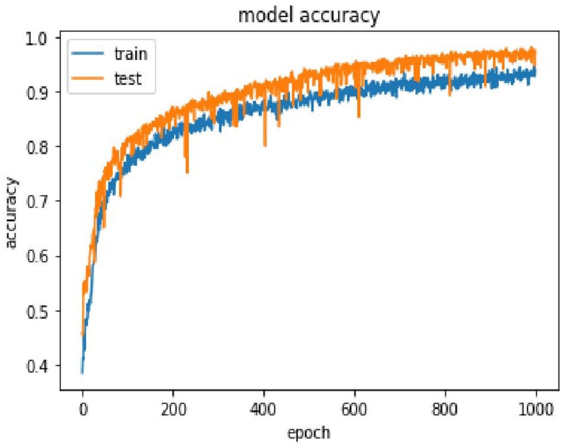


Fig 4.1 Accuracy Graph

The fig 4.1 shows accuracy of the CNN model during training testing is shown in the first figure through 1000 epochs. Important findings include: When the training accuracy and test accuracy are on the rise, it is a good sign that learning has taken place. When test accuracy is higher than training accuracy, it may indicate strong generalization ability, but not necessarily so. It also happens due to test dataset benefiting more from data augmentation or regularization.

The test accuracy curve can bounce around more than the training accuracy because of varying batches and models being sensitive to the unseen data.

2. Model Loss Analysis

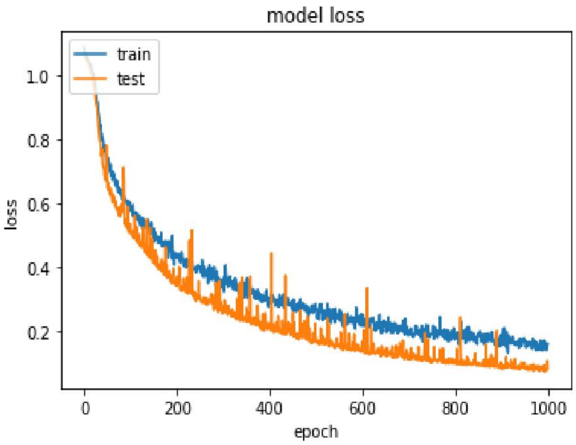
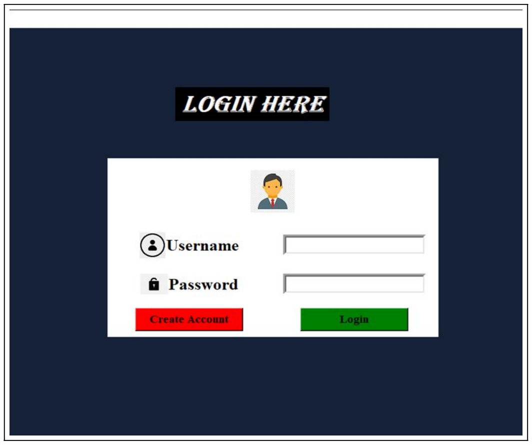


Fig 4.2 Loss Graph

The fig 4.2 show the model's loss changes.

Over time, both the training as well as test loss decrease which indicates good convergence. Abnormally low-test loss means less than training lossgenerallyindicatesthataproblemmaybetaking place such as data leakage, case in which test data maybeunintentionallyaffectingthemodel.Theloss doesn't oscillate between epochs: The stabilization in loss suggests that the model is fine-tuning its decision boundaries and learning steadily.

3. Login Page


LOGIN HERE

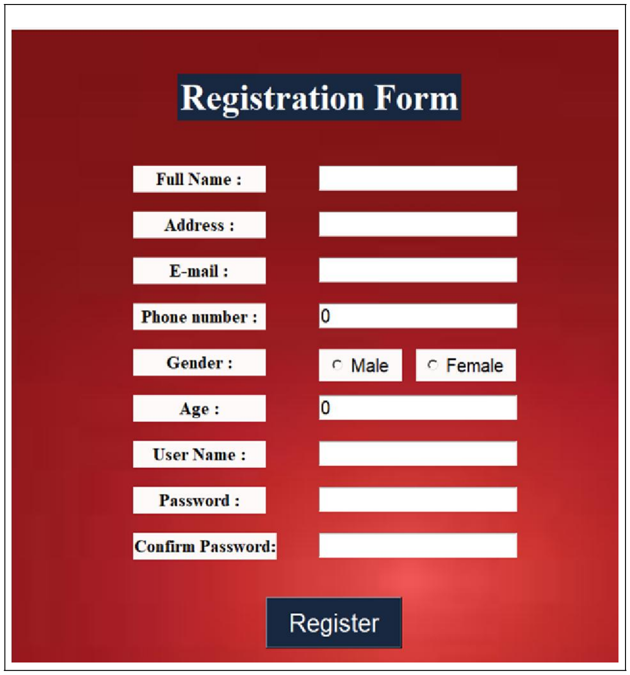
Username

Password

Fig4.3 Login Page

The Login Page GUI is a user interface designed using Python's Tkinterlibrary . It serves as the entry point for users to access the main application . The interface includes labeled fields where users can enter their username and password . When the user clicks the "Login" button the input credentials are validated .

4. Registration Form :


Registration Form

Full Name :

Address :

E-mail :

Phone number : 0

Gender : ○ Male ○ Female

Age : 0

User Name :

Password :

Confirm Password:

Fig 4.4 . Registration Form

The registration pages are essential components of any application that manages user accounts. The registration page allows new users to create an account by providing details such as username, email, and password. This information is securely stored in a database after validation. Once registered, users can
access the login page, where they enter their credentials to authenticate themselves. If the credentials match the stored data, access is granted to the application's protected areas. These pages help ensure that only authorized users can access specific features, maintaining the system's security and personalization.

5. Output

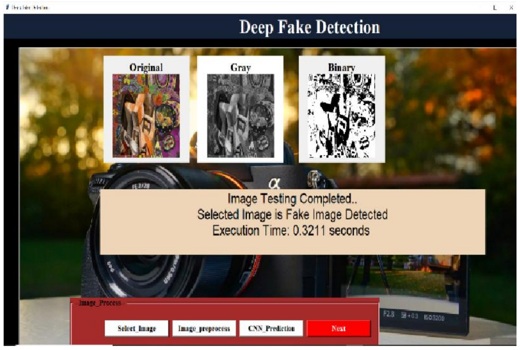


Image Testing Completed..
Selected Image is Fake Image Detected
Execution Time: 0.3211 seconds

Fig. 4.5. Output

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown remarkable effectiveness in detecting deepfake content. The preprocessing steps, such as resizing, grayscale conversion, and noise reduction, played a vital role in enhancing input quality, enabling the model to focus on relevant features. The training phase involved feeding the CNN with both real and fake samples, allowing it to learn distinguishing patterns and subtle inconsistencies introduced by manipu- lation techniques. Despite challenges posed by high-quality and realistic forgeries, the trained model exhibited robust performance in classifying content accurately.

V. CONCLUSION

We have laid the groundwork for the development of deep fake detection models designed to identify synthetic media, encompassing images, generated through deep learning techniques. While the models themselves have yet to be constructed, our research has progressed through the initial phases, establishing a clear methodology and setting research objectives to guide our future endeavors. Our current progress involves an in-depth literature review, offering valuable insights into existing meth- ods and the challenges associated with deep fake detection. In conclusion, while we are currently in the initial stages of our research, the methodology and research framework we have established are robust and well-defined. We anticipate the forth- coming phases with enthusiasm as we embark on the development and evaluation of the deep fake detection models, ultimately contributing to the preservation of trust and authenticity in the digital information ecosystem.

REFERENCES

[1]. H.Farid, Image forgery detection, IEEE Signal Process. Mag., vol. 26, no. 2, pp. 1625, Mar. 2021.

[2]. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2022, pp. 19.

[3]. P. Baldi, Autoencoders, unsupervised learning, and deep architectures, in Proc. ICML Workshop Unsupervised Transf. Learn., 2022, pp. 3749.

[4]. T. Karras, S. Laine, and T. Aila, A style-based generator architecture for gen- erative adversarial networks, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2023, pp. 44014410.

[5]. Y. Mirsky and W. Lee, The creation and detection of deep fakes: A survey, ACM Comput. Surv., vol. 54, no. 1, pp. 141, Jan. 2022.

DOI: 10.48175/568

[6]. M.Masood,M.Nawaz,K.M.Malik,A.Javed,andA.Irtaza, Deepfakes generation and detection: State-of-the-art, open challenges, countermea sures, and way forward, 2021, arXiv:2103.00484.

[7]. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Inf. Fusion, vol. 64, pp. 131148, Dec. 2020.

[8]. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, Deep learning for deepfakes creation and detection: Asurvey, 2019, arXiv:1909.11573.

[9]. L. Verdoliva, Media forensics and DeepFakes: An overview, IEEE J. Sel. Top- ics Signal Process., vol. 14, no. 5, pp. 910932, Aug. 2020.

[10]. K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cy- bern., vol. 36, no. 4, pp. 193202, Apr. 2021

DOI: 10.48175/568