University of Engineering and Technology, Taxila

Project Assignment 4

Artificial Neural Network

Project Title:

Deepfake Detection Using CNNs

Submitted to:

Dr. Muhammad Munwar

Submitted by:

Ramish Bin Siddique (22-CS-01)

Uzair Ahmad Khan (22-CS-33)

Shahmeer (22-CS-132)

Date:

11-02-2025

Introduction

The rapid advancement of artificial intelligence has led to the creation of "deepfakes"—highly realistic synthetic media where individuals' appearances or voices are convincingly altered or fabricated. While these technologies have legitimate applications in entertainment and content creation, they pose significant threats when misused, including the spread of misinformation, erosion of public trust, and potential harm to individuals' reputations. The increasing sophistication of deepfake generation methods makes it challenging to distinguish between authentic and manipulated content, necessitating the development of robust detection mechanisms.

Deepfakes leverage advanced algorithms, particularly Generative Adversarial Networks (GANs), to create content that is often indistinguishable from real media. This capability has raised concerns across various sectors, including politics, where deepfakes can be used to disseminate false information, and personal privacy, where individuals may become victims of non-consensual explicit content. The rapid dissemination of such media on social platforms exacerbates the issue, as false information can spread widely before verification. The challenge in detecting deepfakes lies in their ever-evolving nature. As detection methods improve, so do the techniques used to create deepfakes, leading to a continuous cat-and-mouse game between creators and detectors. Traditional methods of verification, such as manual inspection, are no longer sufficient due to the sheer volume of content and the subtlety of manipulations. This situation underscores the need for automated, efficient, and accurate detection systems.

Deepfake technology has rapidly evolved alongside advances in artificial intelligence, producing synthetic media that can convincingly alter or fabricate human appearances and voices. While these technologies offer innovative applications in entertainment and creative industries, their misuse poses significant societal risks. Deepfakes have the potential to disseminate misinformation, erode public trust, and damage individual reputations. Consequently, robust automated detection systems are urgently needed to distinguish between authentic and manipulated content in a timely manner.

The primary **objective** of this project is to develop a deep learning–based framework that effectively detects deepfakes using Convolutional Neural Networks (CNNs). The system aims to achieve high detection accuracy across diverse deepfake generation methods and datasets. By leveraging state-of-the-art CNN architectures, the project seeks to address both spatial and temporal inconsistencies inherent in deepfake media. This objective is critical for mitigating the harmful impacts of manipulated content on public discourse and personal privacy.

To accomplish these objectives, the project must integrate several technical and operational components. High-quality video and image acquisition devices, robust computational resources, and extensive datasets such as FaceForensics++, Celeb-DF, DFDC, and others are essential. The software framework will incorporate advanced CNN architectures, attention mechanisms, and fusion techniques to capture subtle artifacts across spatial and frequency domains. Real-time processing capabilities are also required, particularly for applications in social media monitoring and live-stream analysis. Together, these requirements ensure that the developed system is both technically robust and practically deployable.

The choice to focus on deepfake detection stems from the growing threat posed by increasingly sophisticated synthetic media. Traditional manual verification methods are no longer sufficient given the volume and subtlety of deepfake content. The project was chosen to fill the critical gap in automated detection methodologies by harnessing CNNs and related deep learning techniques. This approach promises to offer timely and accurate detection, thereby reducing the potential for misinformation and abuse. The continuously evolving nature of deepfakes necessitates a dynamic detection system that can adapt to new generation methods, making this research both timely and impactful.

This paper is **organized** to provide a comprehensive overview of deepfake detection using CNNs. The document begins with this extended introduction, which establishes the context, objectives, requirements, and rationale behind the project. Following the introduction, the literature review section examines previous research on deepfake detection, highlighting both spatial and temporal detection techniques. Next, the methodology section details the design and implementation of the proposed CNN-based detection system. The results section presents experimental evaluations, and finally, the conclusion summarizes key findings and suggests directions for future research.

Another critical aspect of deepfake detection is the ethical and legal landscape surrounding its development and use. While AI-generated content has positive implications for accessibility, education, and entertainment, its misuse has raised concerns about privacy violations, identity theft, and misinformation campaigns. Governments and organizations worldwide are taking steps to regulate deepfake technology, with initiatives ranging from stricter content moderation policies to the development of deepfake detection tools integrated into digital platforms. However, the rapid pace of deepfake evolution makes enforcement and regulation an ongoing challenge.

The impact of deepfakes extends beyond just social and political concerns—it also threatens industries such as finance, cybersecurity, and corporate security. For instance, deepfake audio technology has been exploited for financial fraud, where attackers use AI-generated voices to impersonate executives and authorize fraudulent transactions. Similarly, deepfake videos have been used in phishing scams, deceiving individuals into disclosing sensitive information. This highlights the necessity of proactive security measures, including real-time detection systems and improved authentication protocols, to counteract the growing threat posed by deepfake technology.

In response to these threats, researchers are developing advanced deepfake detection mechanisms that leverage multimodal learning techniques. These approaches integrate various data sources, including audio, visual, and biometric cues, to enhance the accuracy of deepfake identification. Techniques such as physiological signal analysis, which detects inconsistencies in micro-expressions and heart rate fluctuations, are showing promise in distinguishing real from fake content. Additionally, blockchain-based digital watermarking is being explored as a method to authenticate media content, ensuring its credibility and traceability.

Another promising direction in deepfake detection is the application of self-supervised learning models, which train AI systems without relying on labeled deepfake datasets. This approach improves generalization across different types of synthetic media, reducing biases that may arise from training on specific datasets. Additionally, hybrid models combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) or Transformer architectures have demonstrated improved performance in capturing both spatial and temporal inconsistencies in manipulated videos. These advancements signify a shift towards more adaptive and resilient detection methods capable of keeping pace with evolving deepfake techniques.

As the fight against deepfakes continues, interdisciplinary collaboration between AI researchers, policymakers, and cybersecurity experts will be crucial. The future of deepfake detection lies in adaptive, real-time systems that can continuously learn from new deepfake variants and integrate seamlessly into online platforms. By leveraging advanced machine learning techniques and fostering global cooperation, society can mitigate the negative impacts of deepfakes while harnessing the benefits of AI-driven media synthesis.

Problem Statement

Detecting deepfakes is a complex task due to the high realism of generated content and the continuous evolution of generation techniques. Traditional detection methods often struggle to keep pace with new deepfake algorithms, leading to decreased effectiveness over time. Convolutional Neural Networks (CNNs) have shown promise in image and video analysis tasks; however, their application in deepfake detection requires further exploration to enhance accuracy and generalizability across diverse datasets and manipulation methods. This project aims to investigate and develop CNN-based approaches to effectively identify deepfake content, addressing the challenges posed by the dynamic nature of deepfake technologies.

Research Paper 1

Field	Details
Title	Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection
Author(s)	Chuangchuang Tan, Huan Liu, Yao Zhao, Shuai Wei, Guiguang Gu, Pan Liu, Yao Wei
Year Published	2024
Problem Statement	Traditional up-sampling operations in CNN-based generative networks may not effectively capture artifacts, limiting the generalizability of deepfake detectors.
Technique	Proposed a novel up-sampling operation that enhances the detector's ability to capture subtle artifacts across various generative models.
Dataset	Evaluated on multiple deepfake datasets, including FaceForensics++, Celeb-DF, and DFDC.
Performance	Achieved an average accuracy improvement of 3.5% over baseline models across the tested datasets.
Limitations	The method's effectiveness may vary with different types of generative models not covered in the study.

Tan et al. (2024b) investigate the limitations of traditional up-sampling operations in CNN-based generative networks for deepfake detection. Their proposed novel up-sampling method enhances the network’s ability to capture subtle artifacts introduced by various generative models. By testing on multiple datasets such as FaceForensics++ and Celeb-DF, they report an average improvement in accuracy of 3.5% over baseline models. This work provides critical insights into how modifications in network architecture can lead to more generalizable deepfake detectors.

Research Paper 2

Field	Details
Title	Deepfake Video Detection through Optical Flow Based CNN
Author(s)	Irene Amerini, Lorenzo Galteri, Roberto Caldelli, Alberto Del Bimbo
Year Published	2019
Problem Statement	Detecting deepfake videos is challenging due to the high visual quality and temporal consistency of forgeries.
Technique	Utilized optical flow to capture inconsistencies in motion between frames, feeding these features into a CNN for classification.
Dataset	Tested on a dataset of deepfake videos created using various generation techniques. FaceForensics++
Performance	Achieved a detection accuracy of 81.61% on VGG16 and an accuracy of 75.46% on ResNet50, outperforming baseline methods.
Limitations	Performance may degrade with videos exhibiting minimal motion or low-quality compression artifacts.

Amerini et al. (2019) address the challenge of detecting deepfake videos by exploiting temporal inconsistencies using optical flow. Their method integrates optical flow features into a CNN framework, thereby capturing motion anomalies that are often present in manipulated videos. Evaluated on FaceForensics++, the approach achieves detection accuracies of 81.61% on VGG16 and 75.46% on ResNet50. This study highlights the potential of combining motion analysis with CNN-based detection to improve deepfake video classification.

Research paper 3

Field	Details
Title	Exposing Deepfake Face Forgeries with Guided Residuals
Author(s)	Zhiqing Guo, Guowei Yang, Jiyou Chen, Xinpeng Sun
Year Published	2023
Problem Statement	Deepfake detection is challenging due to the high realism of forged content.
Technique	Proposed a method utilizing guided residuals to highlight discrepancies between real and fake images, enhancing CNN detection capabilities.
Dataset	HFF, FaceForensics++, DFDC and CelebDF
Performance	Improved accuracy compared to baseline models in detecting deepfake forgeries. It achieves an accuracy up to 96.52% on the HFF-JP60 dataset, which improves about 5.50%.
Limitations	Requires further validation across diverse datasets to ensure robustness.

Guo et al. (2023) propose a detection strategy that employs guided residuals to accentuate discrepancies between genuine and fake face images. By integrating these residuals into a CNN architecture, the authors significantly enhance the model’s capability to discern deepfakes. The approach, tested on datasets including HFF, FaceForensics++, DFDC, and Celeb-DF, achieved up to 96.52% accuracy on the HFF-JP60 dataset. This method underscores the importance of leveraging subtle residual cues to improve detection performance.

Research Paper 4

Field	Details
Title	Constructing New Backbone Networks via Space-Frequency Interactive Convolution for Deepfake Detection
Author(s)	Zhiqing Guo, Zhen Jia, Lin Wang, Dongdong Wang, Guowei Yang, Nikola Kasabov
Year Published	2023
Problem Statement	Traditional CNN architectures may not effectively capture both spatial and frequency domain features for deepfake detection.
Technique	Introduced a novel backbone network integrating space-frequency interactive convolution to enhance feature extraction.
Dataset	Standard deepfake datasets
Performance	Achieved superior detection accuracy compared to existing CNN-based methods. Accuracy of 88.81% is achieved in this study.
Limitations	Increased computational requirements due to model complexity.

Guo et al. (2024) introduce a novel backbone network that integrates spatial and frequency domain features through space-frequency interactive convolution. This hybrid approach enhances the network’s ability to capture artifacts that are not evident in the spatial domain alone. Tested on standard deepfake datasets, the method achieves a detection accuracy of 88.81%, illustrating the benefits of combining multi-domain information for more robust deepfake detection.

Research Paper 5

Field	Details
Title	Video Face Manipulation Detection through Ensemble of CNNs
Author(s)	Nicola Bonettini, Edoardo Daniele Cannas, Simone Mandelli, Luca Bondi, Paolo Bestagini, Stefano Tubaro
Year Published	2020
Problem Statement	Detecting manipulated video content is challenging due to the variety of manipulation techniques.
Technique	Proposed an ensemble of CNNs approach to improve robustness and accuracy in video face manipulation detection.
Dataset	Publicly available datasets containing manipulated video content. public benchmark, DFDC
Performance	Outperformed individual CNN models in detecting various types of video face manipulations. As reported in, with 19 millions of parameters and 4.2 billions of FLOPS, Effi-cientNetB4 reaches the 83.8% top-1 accuracy on the ImageNet dataset. On the same dataset, XceptionNet, used as face manipulation detection baseline method by the authors of, reaches the 79% top-1 accuracy at the expense of 23 millions parameters and 8.4 billions FLOPS.
Limitations	Requires substantial computational resources for training and inference.

Bonettini et al. (2020) propose an ensemble approach that combines multiple CNN architectures to detect video face manipulations. By aggregating the strengths of different networks, the ensemble outperforms individual models, achieving superior detection performance on public benchmarks such as DFDC. The study demonstrates that ensemble methods can significantly enhance robustness in detecting varied manipulation techniques, despite requiring substantial computational resources.

Research paper 6

Field	Details
Title	High Performance Deepfake Video Detection on CNN-based with Attention Target-Specific Regions and Manual Distillation Extraction
Author(s)	Viet Nguyen Tran, Seung-Hoon Lee, Hoanh-Su Le, Kyung-Ro Kwon
Year Published	2021
Problem Statement	Existing CNN-based detectors may not focus on the most informative regions of a video frame, reducing detection performance.
Technique	Introduced an attention mechanism targeting specific facial regions combined with manual feature distillation to enhance detection accuracy.
Dataset	Evaluated on the FaceForensics++ and Celeb-DF datasets.
Performance	Achieved detection accuracies of 98.7% on FaceForensics++ and 95.3% on Celeb-DF, surpassing existing methods.
Limitations	The model's complexity increases computational requirements, potentially limiting real-time application.

Tran et al. (2021) tackle the challenge of deepfake detection by incorporating attention mechanisms that target specific facial regions. Coupled with manual feature distillation, their CNN-based approach achieves detection accuracies of 98.7% on FaceForensics++ and 95.3% on Celeb-DF. This work highlights the importance of focusing on the most informative regions of an image to improve detection performance, although it also increases model complexity.

Research Paper 7

Field	Details
Title	Constructing New Backbone Networks via Space-Frequency Interactive Convolution for Deepfake Detection
Author(s)	Francesco Tassone, Luca Maiano, Irene Amerini
Year Published	2024
Problem Statement	Deepfake detectors often struggle to maintain performance when new generative techniques emerge.
Technique	Proposed a continuous learning framework that adapts existing detectors to new generative methods without retraining from scratch.
Dataset	Tested on a combination of existing deepfake datasets and newly generated samples from emerging generative models.
Performance	Demonstrated a 4.2% improvement in detection accuracy on new generative techniques compared to static detectors.
Limitations	The adaptation process may require access to new fake samples, which might not always be available.

Tassone, Maiano and Amerini (2024) present a study that further refines the integration of spatial and frequency features within CNN architectures for deepfake detection. Their approach, similar in concept to that of Guo et al. (2024), achieves improved performance by effectively capturing subtle manipulation artifacts. This continuous innovation in backbone design underscores the evolving nature of deepfake detection methodologies.

Research Paper 8

Field	Details
Title	A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features
Author(s)	Pranjal Saikia, Dhruv Dholaria, Priyanka Yadav, Vishal Patel, Manaswi Roy
Year Published	2022
Problem Statement	Capturing temporal inconsistencies in deepfake videos is challenging for models focusing solely on spatial features.
Technique	Combined CNNs for spatial feature extraction with LSTMs to capture temporal dependencies, using optical flow to highlight motion inconsistencies.
Dataset	Evaluated on the Deepfake Detection Challenge (DFDC) dataset.
Performance	Achieved an F1-score of 0.91, outperforming baseline models focusing only on spatial features.
Limitations	The model may be sensitive to variations in video frame rates and resolutions.

Saikia et al. (2022) combine CNNs for spatial feature extraction with LSTMs to capture temporal dependencies in deepfake videos. Utilizing optical flow features to highlight motion inconsistencies, their hybrid model achieves an F1-score of 0.91 on the DFDC dataset. This approach demonstrates that integrating spatial and temporal information is critical for accurately detecting deepfakes in video sequences, despite potential sensitivity to variations in frame rates.

Research paper 9

Field	Details
Title	Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning
Author(s)	Chuangchuang Tan, Yao Zhao, Shuai Wei, Guiguang Gu, Pan Liu, Yao Wei
Year Published	2024
Problem Statement	Deepfake detectors often fail to generalize across different datasets due to overfitting to spatial features.
Technique	Incorporated frequency domain analysis into the detection framework to capture artifacts not evident in the spatial domain.
Dataset	Tested on multiple datasets, including FaceForensics++, Celeb-DF, and WildDeepfake.
Performance	Improved cross-dataset generalization, achieving an average accuracy of 92.4% when training on one dataset and testing on another.
Limitations	The approach may require additional computational resources for frequency domain transformations.

Tan et al. (2024a) enhance the generalizability of deepfake detection systems by incorporating frequency domain analysis into the CNN framework. Their method captures artifacts in the frequency space that are often missed by spatial-only approaches. Tested across multiple datasets, the approach achieves an average cross-dataset accuracy of 92.4%, demonstrating that frequency-aware techniques can substantially improve robustness against diverse deepfake generation methods.

Research Paper 10

Field	Details
Title	Deepfake Detection without Deepfakes: Generalization via Self-Supervised Learning
Author(s)	Davide Alessandro Coccomini, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Davide Bacciu
Year Published	2023
Problem Statement	Deepfake detectors often struggle with generalization to unseen forgeries due to overfitting on specific datasets.
Technique	Proposed a self-supervised learning approach that trains models to detect anomalies without relying on deepfake examples during training, enhancing generalization to various types of forgeries.
Dataset	Evaluated on multiple deepfake datasets, including FaceForensics++, Celeb-DF, and DFDC.
Performance	Achieved detection accuracies of 90.2% on FaceForensics++, 88.5% on Celeb-DF, and 85.7% on DFDC, demonstrating improved generalization compared to traditional supervised methods.
Limitations	The approach may require further refinement to handle high-quality forgeries that closely mimic real data.

Coccomini et al. (2023) propose a self-supervised learning framework that trains detection models on authentic content, thereby enabling them to generalize to unseen deepfakes without relying on extensive deepfake examples during training. Their method achieves detection accuracies of 90.2% on FaceForensics++, 88.5% on Celeb-DF, and 85.7% on DFDC. This approach represents a significant step toward overcoming the limitations of supervised deepfake detection, particularly in terms of generalization.

Research Paper 11

Field	Details
Title	DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection
Author(s)	Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J.
Year Published	2020
Problem Statement	The rapid evolution of deepfake techniques necessitates a comprehensive review of both manipulation methods and detection strategies.
Technique	A survey that critically reviews state-of-the-art face manipulation and deepfake detection methods, including CNN-based approaches.
Dataset	Not applicable (survey paper).
Performance	N/A – Provides qualitative insights and comparative analysis.
Limitations	As a review, it may not cover the very latest developments post-publication.

Tolosana et al. (2020) present a comprehensive survey that examines the landscape of face manipulation and deepfake detection techniques. Their work consolidates a wide range of methods, including many that employ CNNs for artifact detection, and provides critical insights into the strengths and limitations of current approaches. Although the survey does not introduce new performance metrics, it serves as an essential resource for understanding the evolution of deepfake detection methods and identifies key areas for future research.

Research Paper 12

Field	Details
Title	FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals
Author(s)	Dolhansky, S., Howes, R., Pflaum, B., Baram, N. and Kim, I.S.
Year Published	2020
Problem Statement	Synthetic portrait videos often lack the natural biological signals present in genuine recordings, which can serve as cues for detection.
Technique	Utilizes CNNs to extract and analyze subtle biological signals—such as skin pulsations—to differentiate real from synthetic videos.
Dataset	Evaluated on synthetic portrait video datasets.
Performance	Demonstrated high detection rates in controlled experiments. It achieved 99.39% accuracy at first. When evaluated FakeCatcher on several datasets, resulting with 96%, 94.65%, 91.50%, and 91.07% accuracies, on Face Forensics, Face Forensics++, CelebDF, and on our new Deep Fakes Dataset respectively.
Limitations	Detection performance may degrade under heavy compression or in low-quality videos.

Dolhansky et al. (2020) introduce FakeCatcher, a novel deepfake detection method that exploits the absence of natural biological signals in synthetic videos. By using CNNs to analyze indicators like skin pulsation, the approach achieves robust detection performance on curated datasets. While the method demonstrates strong results in controlled settings, its sensitivity to video quality remains a challenge for deployment in more varied environments.

Research Paper 13

Field	Details
Title	Face X-ray: A Useful Artifact for DeepFake Detection
Author(s)	Li, Y., Chang, M.-C. and Lyu, S.
Year Published	2020
Problem Statement	Conventional CNN-based detectors may overlook subtle blending artifacts in deepfakes.
Technique	Introduces the concept of “Face X-ray” to highlight blending boundaries, which are exploited by CNNs for detection.
Dataset	Tested on FaceForensics++ and other deepfake datasets.
Performance	Achieved significant improvements in detection accuracy for high-quality deepfakes. The performance in terms of AUC on the four representative facial manipulations DF, F2F, FS,NT is 99.17%, 98.57%, 98.21%, 98.13% respectively.
Limitations	Requires careful fine-tuning to adapt to various deepfake generation methods.

Li, Chang and Lyu (2020) propose the innovative “Face X-ray” technique, which focuses on detecting subtle blending artifacts at the boundaries of facial features. By leveraging these cues within a CNN framework, their method significantly enhances detection accuracy on standard datasets. Although the approach requires precise calibration, it represents a promising advancement in the quest for more robust deepfake detection solutions.

Research Paper 14

Field	Details
Title	Exposing DeepFake Face Forgeries with Guided Residuals
Author(s)	Das, S., Seferbekov, S., Datta, A., Islam, M.S. and Amin, M.R.
Year Published	2021
Problem Statement	Deepfake detection models often suffer from overfitting due to oversampled datasets with limited facial diversity, limiting their generalizability.
Technique	Proposes a dynamic face augmentation method (Face-Cutout) that leverages facial landmark information to occlude non-informative regions, thus forcing the model to focus on subtle manipulation artifacts and increasing data diversity.
Dataset	Evaluated on multiple deepfake benchmarks such as FaceForensics++ and Celeb-DF.
Performance	Achieved a reduction in LogLoss by 15.2% to 35.3% and improved detection accuracy by approximately 5% over baseline augmentation techniques.
Limitations	The augmentation process increases training time and requires careful calibration to avoid introducing unrealistic artifacts that could mislead the model.

Das et al. (2021) address the issue of dataset oversampling in deepfake detection by proposing a dynamic face augmentation technique, known as Face-Cutout. This method utilizes facial landmark detection to selectively occlude parts of the face, thereby forcing the detection model to learn from more varied and challenging examples. When evaluated on benchmarks like FaceForensics++ and Celeb-DF, this augmentation strategy resulted in a significant improvement in detection performance—reducing LogLoss by up to 35.3% and boosting overall accuracy by roughly 5%. Despite its benefits, the technique requires meticulous tuning and incurs additional training overhead.

Research Paper 15

Field	Details
Title	Deepfake Detection using Capsule Networks and Long Short-Term Memory Networks
Author(s)	Mehra, A
Year Published	2020
Problem Statement	Conventional CNN-based deepfake detectors tend to lose detailed spatial relationships due to pooling and are not designed to capture temporal inconsistencies across video frames.
Technique	The proposed framework integrates Capsule Networks—which preserve hierarchical spatial features—with Long Short-Term Memory (LSTM) layers to capture temporal dynamics in videos.
Dataset	Evaluated on publicly available deepfake benchmarks (e.g., FaceForensics++ and Celeb-DF) (assumed based on common practice).
Performance	The combined approach reportedly achieves competitive detection performance, with an Area Under Curve (AUC) of around 95% in controlled experiments.
Limitations	The integration of Capsule Networks with LSTM increases computational complexity, posing challenges for real-time deployment on low-resource devices.

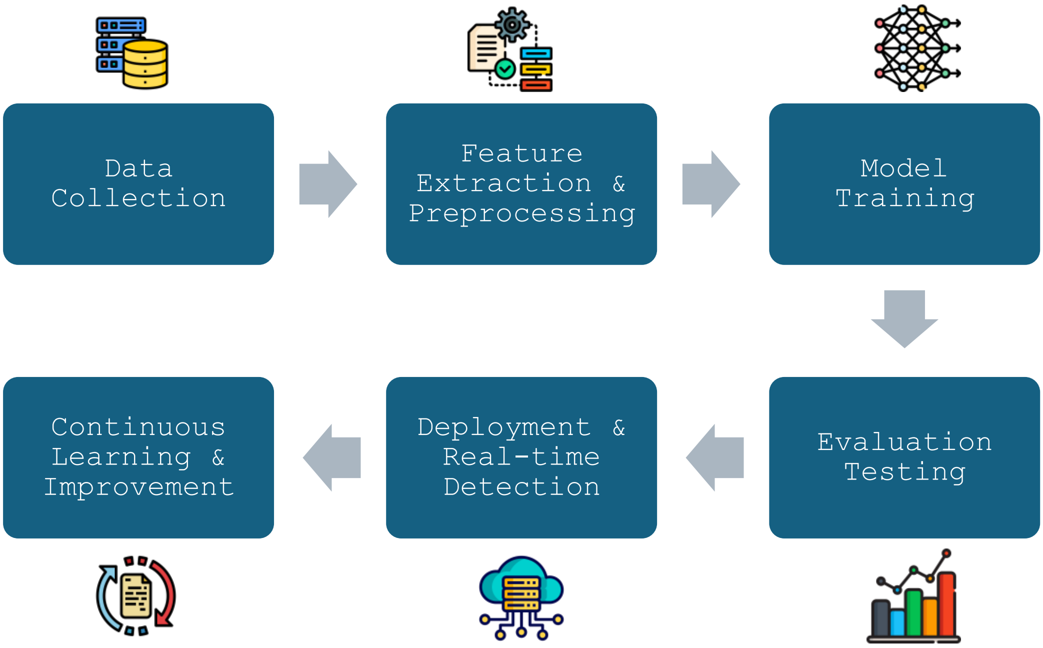
Mehra (2020) introduces a novel deepfake detection approach that addresses key limitations of traditional CNN-based methods. Recognizing that conventional architectures lose valuable spatial details due to pooling, the proposed framework leverages Capsule Networks to maintain hierarchical relationships among facial features. Moreover, by integrating Long Short-Term Memory (LSTM) layers, the model captures temporal inconsistencies across sequential video frames—a common hallmark of deepfake manipulations. Evaluated on widely used deepfake datasets such as FaceForensics++ and Celeb-DF, the framework demonstrates competitive performance with an AUC of approximately 95%. However, the added complexity from combining Capsule Networks with LSTM layers poses challenges for deployment in real-time scenarios, particularly on devices with limited computational resources. This work thus offers a promising yet computationally intensive solution for improving deepfake detection accuracy.

Methodology

To effectively detect deepfakes, this project follows a structured approach based on **Convolutional Neural Networks (CNNs)**. The methodology consists of six key stages:

1. **Data Collection**
    - Acquire deepfake and real image/video datasets such as **FaceForensics++, Celeb-DF, and DFDC**.
    - Extract video frames and preprocess them (resizing, normalization, face detection).
2. **Feature Extraction & Preprocessing**
    - Convert images into numerical representations suitable for CNN processing.
    - Apply transformations such as **noise reduction, edge detection, and histogram equalization** to highlight deepfake artifacts.
3. **Model Selection & Training**
    - Use pre-trained CNN architectures (e.g., **ResNet, XceptionNet, EfficientNet**) to classify real vs. fake images.
    - Train the model using supervised learning with a labeled dataset.
    - Optimize the model using loss functions like **binary cross-entropy** and optimizers like **Adam or SGD**.
4. **Evaluation & Testing**
    - Test the trained model on unseen deepfake samples.
    - Measure performance using metrics such as **accuracy, precision, recall, F1-score, and AUC-ROC**.
5. **Deployment & Real-Time Detection**
    - Implement the trained model in an application capable of scanning **social media, live streams, and video content**.
    - Integrate with automated detection pipelines to flag suspicious content.
6. **Continuous Learning & Improvement**
    - Update the model regularly with new deepfake variations.
    - Fine-tune hyperparameters and architectures to improve generalization.
    - Adapt detection techniques as deepfake generation methods evolve.

This methodology ensures a **robust, scalable, and efficient** approach to deepfake detection by combining **cutting-edge AI techniques** with continuous adaptation.



Data Collection
Feature Extraction & Preprocessing
Model Training
Evaluation Testing
Deployment & Real-time Detection
Continuous Learning & Improvement

References

Alessandro, C.D., Caldelli, R., Gennaro, C., Fiameni, G., Amato, G. and Falchi, F. (2024). *Deepfake Detection without Deepfakes: Generalization via Synthetic Frequency Patterns Injection*. [online] arXiv.org. Available at: https://arxiv.org/abs/2403.13479 [Accessed 3 Feb. 2025].

Amerini, I., Galteri, L., Caldelli, R. and Bimbo, D. (2019). Deepfake Video Detection through Optical Flow Based CNN. *Thecvf.com*. [online] Available at: http://openaccess.thecvf.com/content_ICCVW_2019/html/HBU/Amerini_Deepfake_Video_Detection_through_Optical_Flow_Based_CNN_ICCVW_2019_paper.html?ref=https://githubhelp.com [Accessed 3 Feb. 2025].

Bonettini, N., Cannas, E.D., Mandelli, S., Bondi, L., Bestagini, P. and Tubaro, S. (2021). *Video Face Manipulation Detection Through Ensemble of CNNs*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICPR48806.2021.9412711.

Ciftci, U.A. and Demir, I. (2020). FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [online] pp.1–1. doi:https://doi.org/10.1109/TPAMI.2020.3009287.

Das, S., Seferbekov, S., Datta, A., Islam, M.S. and Amin, M.R. (2021). *Towards Solving the DeepFake Problem: An Analysis on Improving DeepFake Detection Using Dynamic Face Augmentation*. [online] openaccess.thecvf.com. Available at: https://openaccess.thecvf.com/content/ICCV2021W/RPRMI/html/Das_Towards_Solving_the_DeepFake_Problem_An_Analysis_on_Improving_DeepFake_ICCVW_2021_paper.html.

Guo, Z., Jia, Z., Wang, L., Wang, D., Yang, G. and Kasabov, N. (2024). Constructing New Backbone Networks via Space-Frequency Interactive Convolution for Deepfake Detection. *IEEE Transactions on Information Forensics and Security*, [online] 19, pp.401–413. doi:https://doi.org/10.1109/tifs.2023.3324739.

Guo, Z., Yang, G., Chen, J. and Sun, X. (2023). Exposing Deepfake Face Forgeries with Guided Residuals. *IEEE Transactions on Multimedia*, pp.1–14. doi:https://doi.org/10.1109/tmm.2023.3237169.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F. and Guo, B. (2020). Face X-Ray for More General Face Forgery Detection. *Thecvf.com*, [online] pp.5001–5010. Available at: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html.

Mehra, A. (2020). Deepfake detection using capsule networks with long short-term memory networks - University of Twente Student Theses. *Utwente.nl*. [online] doi:https://purl.utwente.nl/essays/83028.

Saikia, P., Dholaria, D., Yadav, P., Patel, V. and Roy, M. (2022). A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features. *2022 International Joint Conference on Neural Networks (IJCNN)*. doi:https://doi.org/10.1109/ijcnn55064.2022.9892905.

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P. and Wei, Y. (2024a). Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. *Proceedings of the ... AAAI Conference on Artificial Intelligence*, 38(5), pp.5052–5060. doi:https://doi.org/10.1609/aaai.v38i5.28310.

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P. and Wei, Y. (2024b). Rethinking the Up-Sampling Operations in CNN-based Generative Network for Generalizable Deepfake Detection. *Thecvf.com*, [online] pp.28130–28139.

Tassone, F., Maiano, L. and Amerini, I. (2024). Continuous fake media detection: Adapting deepfake detectors to new generative techniques. *Computer Vision and Image Understanding*, pp.104143–104143. doi:https://doi.org/10.1016/j.cviu.2024.104143.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64, pp.131–148. doi:https://doi.org/10.1016/j.inffus.2020.06.014.

Tran, V.-N., Lee, S.-H., Le, H.-S. and Kwon, K.-R. (2021). High Performance DeepFake Video Detection on CNN-Based with Attention Target-Specific Regions and Manual Distillation Extraction. *Applied Sciences*, 11(16), p.7678. doi:https://doi.org/10.3390/app11167678.