National College of Ireland

Deepfake Detection in AV1 Compressed Videos with EfficientNet and Stacked Bi-LSTM Model

MSc Research Project
Data Analytics

Aniket Suryakant Ghadge
Student ID: x23106786

School of Computing
National College of Ireland

Supervisor: Dr David Hamill

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Aniket Suryakant Ghadge
Student ID:	x23106786
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr David Hamill
Submission Due Date:	12/08/2024
Project Title:	Deepfake Detection in AV1 Compressed Videos with EfficientNet and Stacked Bi-LSTM Model
Word Count:	6686
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Aniket Suryakant Ghadge
Date:	12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

MSc Research Project

Your Name/Student Number	Course	Date
x23106786	MSc. Data Analytics	12/08/2024

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
NA	NA	NA

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

NA	
NA	
NA	NA

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

NA

Additional Evidence:

NA

Deepfake Detection in AV1 Compressed Videos with EfficientNet and Stacked Bi-LSTM Model

Aniket Suryakant Ghadge
x23106786

Abstract

The rise of Deepfake video content online jeopardises the integrity of digital media as they are optimised to spread disinformation, sow doubt, and facilitate confusion. The following research investigates the effects of AV1 lossy compression on deepfake video detection, highlighting the need for models that adapt to different compression levels. To capture temporal inconsistencies in video frames, the author proposed an ensemble model that combines a three-layered bidirectional LSTM network with EfficientNet-B0 for feature extraction. This experiment tested the model on raw and AV1-compressed videos at 250 kbps and 1024 kbps bitrates using the FaceForensics++ dataset. The key experimental findings show that our model achieves over 90% accuracy in all formats, with the best performance in terms of accuracy, recall, and fewer false positives and negatives being seen in high-bitrate AV1 videos. On the other hand, low-bitrate compression adds complexity by hiding fake artefacts, which degrades model performance with a higher number of false positives. This study highlights the challenges and importance of adapting deepfake detection models to handle various levels of lossy compression effectively.

1 Introduction

The expeditious adoption of social and digital media has dramatically changed how we both share and use information. But this change has also raised serious questions about the truthfulness of the content, particularly regarding the development of deepfake technology. Deepfakes are a type of artificial intelligence (AI) art in which highly realistic videos and images of people are produced, frequently imitating and subsequently offering the impression that they are actively participating in such activities, which is not the case. As a consequence, there are significant risks to public safety, individual privacy, and even international security from such capability. A deepfake video of a world leader, for example, may cause controversy or conflict, underscoring the urgent need for reliable detection systems.

The detection of such deepfakes has become an increasingly difficult process due to the growing competence of AI-driven tools adopted to create them. The initial focus of research has been related to detecting deepfake detection in images using various machine learning models. With the advancement of technology, increasingly intricate deep learning models were created that analyze both temporal and spatial inconsistencies to detect deepfakes in videos. The efficacy of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks in these tasks can be attributed to their respective strengths in feature extraction and sequential data processing.

Even with these improvements, one important problem is still generally not solved: identifying deepfakes in videos that have been compressed with the AV1 codec. The AV1 codec is a lossy type of codec that is known for its efficiency and royalty-free licensing over other compression techniques. The deepfake detection models rely on subtle inconsistencies, which can be hidden by artefacts and noise introduced by the AV1 codec, which is preferred due to its high efficiency in video compression. Previous researches Kuang et al. (2022); Guan et al. (2023); Wu et al. (2023); Chen et al. (2022) has examined and focused on the effects of alternative codecs, i.e. H.264 and H.265, however, there is a clear research gap relating to the use of the AV1 codec.

Therefore, to address such a gap, this research study suggests adopting a unique fusion model that incorporates a stacked Bi-LSTM network to identify temporal inconsistencies in AV1 compressed videos and EfficientNet-B0 for feature extraction. Bi-LSTM, which is good at identifying long-term dependencies in video frames, and EfficientNet-B0, which is well-known for its exceptional feature extraction performance, work well together for this particular task. This fusion model's performance will be evaluated at various compression bitrates and compared to that of uncompressed video in the study.

Research Question: How well does the fusion of EfficientNet-B0 and stacked Bi-LSTM models adapt to AV1 compressed videos at different bitrates and impact the model performance compared to uncompressed videos?

Research Objectives:

- To evaluate the effectiveness of the EfficientNet-B0 and stacked Bi-LSTM fusion model in detecting deepfakes in AV1 compressed videos.
- To analyze the impact of different AV1 compression bitrates on the performance of the proposed model.
- To compare the performance metrics (accuracy, precision, recall, F1-score, cross entropy loss) of the proposed model against uncompressed videos.

In the next section, this study will analyse the recent research available from within this domain.

2 Related Work

This section provides an overview of the past and current research in developing deepfake detection models, by focusing on key areas such as datasets, detection categories, the use of machine learning and deep learning models, and the effects of various compression techniques. Later, the study will highlight the necessity of adapting to AV1 compression techniques.

2.1 Datasets

In evaluating the proposed model, it is imperative to select both a significant and unbiased dataset. The advancement in the generation of fake videos has substantially increased with the help of powerful deepfake generation techniques such as FaceSwap (FS), Face2Face (F2F), DeepFakes (DF), and NeuralTextures (NT). Research by Hussain et al. (2022) discusses how deepfake generation techniques impact the model's performance.

The prominent datasets that have been predominantly used in the majority of previous research experiments, which proved promising results by setting a benchmark, are the Deep Fake Detection Challenge dataset (DFDC), the Deep Fake Detection dataset (DFD), Face Forensics, Face Forensics++, and Celeb-DF datasets. The research papers Guan et al. (2023); Kuang et al. (2022); Chen et al. (2022); Wu et al. (2023) found that the use of the random sampling technique on the FaceForensics++ dataset was effective in selecting a significant number of random videos that would eventually represent the entire dataset. K et al. (2023) found that using the Ffmpeg library to convert videos into defined frames was an effective method for extracting frames from the videos. The alternative datasets—Deep Fake Detection dataset (DFD), Deep Fake Detection Challenge dataset (DFDC), and Celeb-DF—were also used as supplementary datasets for improving the model training capabilities Guan et al. (2023); Kuang et al. (2022); Wu et al. (2023). The author's own research project experimented with the FaceForensics++ dataset, containing a total of 1000 real videos and 3000 fake videos. FaceForensics++ helps standardise the model evaluation process using the above four core manipulation techniques. The remaining aforementioned datasets have limitations in the wider range of complexity adopted for generating deepfake videos. Guan et al. (2023) stated the use of these four fake video generation algorithms makes FaceForensics++ suitable and provides a core foundation for training and evaluating deepfake detection models by offering a comprehensive range of manipulated videos created using various powerful techniques; by using this dataset, the author's research project aims to develop a robust model that can effectively detect deepfakes, particularly in terms of AV1 compressed videos.

2.2 Deepfake detection categories

Deepfake detection can be broadly categorized based on the approach and focus areas i.e. image-based and video-based detection. Image-based detection methods focus on identifying artefacts or inconsistencies within a single frame. Alternatively, video-based detection methods aim to capture temporal inconsistencies across frames, making them more complex but potentially more effective. Deepfake detections are further drilled down into specific sub-categories including:

- **Temporal-Based Inconsistencies:** These focus on detecting anomalies in the sequence of frames, such as unnatural movement or temporal coherence issues.
- **Spatial-Based Inconsistencies:** These methods identify irregularities within individual frames, such as inconsistencies in lighting, texture, and facial features.
- **Biological-Based Inconsistencies:** These techniques look for physiological signs that are difficult to fake, such as eye blinking patterns, facial expressions, lip-sync, and heart rates.

The predominance of prior research studies has noted the importance of using both spatial and temporal-based inconsistencies for identifying artifacts in the data. Studies by K et al. (2023)and Kuang et al. (2022) have highlighted the significance of combining these inconsistencies by proposing a state-of-the-art approach using fusion techniques with a dual-branch neural network. This approach leverages the strengths of both spatial and temporal detection methods, enhancing the overall accuracy and robustness of deepfake detection systems.

Understanding these categories and their associated methods is essential for developing robust deepfake detection systems. By combining spatial and temporal inconsistencies, researchers can create more effective models, enhancing the accuracy and reliability of detecting manipulated content.

2.3 Evolution in deepfake detection models

Unfortunately, there has been consistent progress development in building of robust, efficient, and effective deepfake video detection techniques. In this section, literature around the evolution of deepfake detection models by using machine learning algorithms to complex deep learning models is discussed in detail. These models employ a variety of techniques to identify subtle inconsistencies or artefacts from the manipulated videos by focusing on both spatial and temporal inconsistencies.

2.3.1 Early Approaches with Machine Learning Models

In the Initial stage, researchers relied heavily on basic machine learning models for developing deepfake detection models. These models have typically used feature extraction techniques to identify irregularities in images and videos, followed by classification algorithms such as Random Forest, KNN, Naive Bayes, decision tree and SVM to distinguish between real and fake content. However, these early methods have limitations in their ability to capture the complex patterns present in deepfake videos which results in lower accuracy and reliability Rana et al. (2021); Hamza et al. (2022); Raveena et al. (2023).

2.3.2 Transition to Deep Learning Models

As the advancement in the generation of deepfake technology has increased by adopting new complex deep learning techniques, researchers began to build additional complex deep learning models as the core foundation of complex deepfake video detection. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, are frequently used due to their powerful capabilities in feature extraction and sequence modelling while handling long-term dependencies. In the following section a number of inquiries within prominent research papers are reviewed, providing useful, appropriate techniques for building the proposed deepfake detection model.

One significant development in recent research is the adopting of hybrid models that combine both CNNs and RNNs. For example, recent research by Suratkar and Kazi (2023) presents a hybrid model that combines transfer learning with CNNs and RNNs. The authors examined pre-trained models on datasets like FaceForensics++ and the DeepFake Detection Challenge (DFDC) by using EfficientNet, ResNet V2, and VGG 16 Inception. According to their results, EfficientNet performed better than other models in terms of accuracy, recall, F-measure, precision, and AUC curve when combined with a transfer learning approach. However, it must be noted that inter-frame relationships which are crucial for identifying temporal inconsistencies were not adequately addressed by this model. Similarly, another study employed CNN and long short-term memory (LSTM) to evaluate spatial and temporal inconsistencies within the frames in order to detect fake videos using DFDC and face forensics datasets. This study by Khedkar et al. (2022) evaluated models on four baseline pre-trained models: Inception V3, DenseNet

121, ResNet-50 v2, and VGGNet-19, where DenseNet performed better than the other models. The study had a series of weaknesses including its neglect to take into consideration other effective pre-trained models (EfficientNet) and furthermore failed to detect the temporal and spatial inconsistencies that had been mentioned in the research question.

The novel approach put forward by K et al. (2023) for identifying the presence of artefacts in compressed deepfake videos (under compression rate 40) using the spatial and temporal approaches aims at addressing the drawbacks of Khedkar et al. (2022) and Suratkar and Kazi (2023). The model was evaluated on the FaceForensics++ dataset. In the temporal approach, the author used the CoViAR method to generate residual features of compressed videos, while in the spatial approach, the MesoNet model was used, with network pruning increasing the model performance. The ResNet model receives these features in order to identify if the videos are legit or fake. The final step in the detection of fake videos takes into account all of the results from the spatial and temporal evaluations. The study successfully handled the model overfitting, including the vanishing gradient problem. The proposed model by K et al. (2023) outperformed earlier methods but unfortunately included limitations as it only considered a compression factor of 40. The model evaluation can also be performed using other performance matrices, except accuracy, as the primary matrix. Research by Hassan et al. (2024) incorporated an ensemble deep learning model in the research study to identify suspicious activities in real time. The authors implemented pre-trained convolutional neural network models such as MobileNet V2 to extract features from the collected videos. Subsequently, these extracted features were passed through the complex using a stacked bi-directional LSTM model for real-time classification of human activities and observed to deliver promising results. These proposed techniques can be helpful in this research project for identifying artefacts from fake compressed videos.

2.4 Effects of various compression techniques

In this subsection, the effects of various compression techniques that are applied using H.264 and H.265 codecs are discussed using previously supported research studies. All the below-mentioned challenges that occur to simple deep learning models help in understanding the need for adaptation of compressed fake videos.

Kuang et al. (2022) introduced a dual-branch approach that utilizes both temporal and spatial techniques on the FaceForensics++ and Celeb-DF datasets. In the initial preprocessing phase, the use of the EfficientNet model on both branches proved efficient and effective by performing feature extraction. The next phase, the Bi-LSTM neural network model, improved the model performance by identifying temporal and spatial inconsistencies within the frames. The proposed model achieved 90% accuracy on both datasets but has limited its research by only using H.264 (compression rate of 40).

Another different approach was executed by Chen et al. (2022), where the authors executed a dual-branch technique for detecting fake compressed images, where one branch helps in identifying similarities using the loss function as the evaluation parameter and the other for classifying (binary classification) using the cross entropy. At the final model evaluation, to determine whether an image is real or fake, the evaluations from the two branches are combined. The datasets used in this experiment are compressed under three major qualities: high quality (HQ), medium quality (MQ), and low quality (LQ). The results observed were that HQ and MQ have achieved more accuracy than LQ. This helps in addressing the need for the adaptation of deep learning models to low compression

quality.

A novel framework brought forward by the authors Guan et al. (2023) is the few-shot scenario technique for identifying inconsistencies within the frames (shots). In the pre-processing phase, the use of EfficientNet-b4 was made for extracting features along with the fusion of RGB. Two suggested methods—cross-forgery and cross-dataset methods—were used to test this model on the FaceForensics++ and Google DFD datasets. Guan et al. (2023) achieved better results than other algorithms with stable performance at a high compression rate (C40) after evaluating the model at C23 and C40 with the comparison baseline of the previous six different algorithms. The model can lead to biased predictions as the number of shots is considered only 10 for long-duration videos.

Additionally, Wu et al. (2023) presents a framework for identifying inconsistencies in videos using CNN and LSTM networks. The approach uses a novel channel groups loss technique to extract features for both original and forgery content. This model adopts a channel-wise attention mechanism and combines CG-loss and cross-entropy loss. Wu et al. (2023) model outperformed current models on the FaceForensics++ and Celeb-DF datasets, showing better efficiency and accuracy, particularly when using compressed videos (C23 and C40). In particular, the model adopted by Wu et al. (2023) successfully outperformed Guan et al. (2023) 77.1% AUC score on the FaceForensics++ dataset, achieving 83.5%. These proposed models, however, fail to tackle AV1 compressed videos, indicating a potential research gap in the literature.

2.5 Comparison between various codecs

The previous highlighted research studies clearly demonstrate how the use of various compression techniques affects the model performance in detecting the fake videos Chen et al. (2022); Kuang et al. (2022); Guan et al. (2023); K et al. (2023); Wu et al. (2023). The following section will help us to understand the comparison between various compression techniques (codecs).

The earlier research study focused on the data quality and efficiency for streaming of AV1 codec videos. In their comparison of AV1, VP9, and H.265 research by Akyazi and Ebrahimi (2018) concluded that AV1 was superior because it was royalty-free and required no licensing fees. Their results further demonstrated that, at various bitrates, AV1 and H.265 performed better in terms of PSNR than VP9. This work was later extended by Uhrina et al. (2024), when comparing metrics such as PSNR, SSI, and VMAF between H.266/VVC, H.265/HEVC, AV1, and H.264/AVC. They concluded that AV1 offered notable bitrate savings with lower computational demands, while H.266/VVC offered the best bitrate savings but required more processing time. In answering a key component of the research question, this study highlights the need for AV1 to be modified for compressed video.

In the subsequent sections, this study will focus on describing the methodology followed and the design architecture of the proposed model with its step-by-step implementations. Later, an evaluation of the achieved results will be discussed.

3 Methodology

The following section firstly includes the collection and selection of appropriate datasets with effective pre-processing techniques, secondly, the selection of models and training, and lastly, fine-tuning using various evaluation metrics are reviewed. Figure 1 below

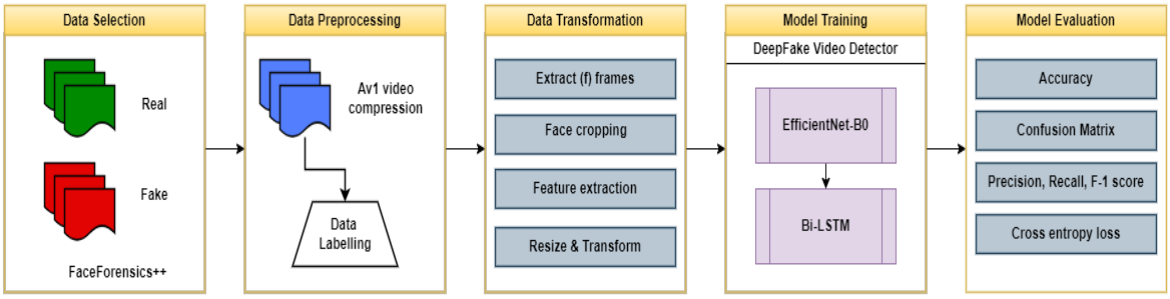

Figure 1: Methodology

illustrates an overview of the methodology adopted in this research study. This approach is similar to the KDD methodology, where the proposed approach is carried out in five phases: data selection, data preprocessing, data transformation, model training, and model evaluation.

3.1 Data Selection

The selection of appropriate datasets is one of the crucial processes throughout the research study to address the objectives of the proposed research questions. Section 2.1 highlights the availability of sources of datasets considered in previous research studies. The most commonly endorsed datasets that provided benchmark model performance were the Deep Fake Detection Challenge dataset (DFDC), Deep Fake Detection dataset (DFD), Face Forensics, Face Forensics++, and finally, Celeb-DF datasets. Also observed by Chen et al. (2022); Khedkar et al. (2022); Kuang et al. (2022); Guan et al. (2023); K et al. (2023); Suratkar and Kazi (2023); Wu et al. (2023), the FaceForensics++[1] dataset offers promising results due to its wider range of complex techniques used for generating deepfake videos, while the other datasets lack these variations but can be considered on the later stage as the supplementary datasets. This research focuses on building a simulation model for examining the effects of the AV1 codec on the deepfake detection model with respect to low and high bitrates. For this research study, a FaceForensics++ dataset is favored, which consists of 3000 fake videos that are generated by using various deepfake generation techniques along with 1000 real videos. Due to the imbalance in the datasets, the random sample selection technique used by Guan et al. (2023); Kuang et al. (2022); Chen et al. (2022) is being utilised, where 280 real and 280 fake videos from all four techniques are selected. Later, these selected videos are transferred to the next data preprocessing phase.

3.2 Data Preprocessing

Following selection of the FaceForensics++ dataset, it is essential to perform preprocessing techniques. In the initial phase, the raw FaceForensics++ videos are compressed into an AV1 codec with a selection of parameters such as bitrates and frames per second (FPS) using an openly available tool called HandBrake[2]. The videos are compressed at

[1]https://github.com/ondyari/FaceForensics
[2]https://handbrake.fr/

two different bitrates—250 kbps for a low bitrate and 1024 kbps for a high bitrate—in order to conduct a comparison study and respond to the aforementioned research question. The alternative method for compressing videos into the AV1 codec is by using another openly available platform called ffmpeg[3], which provides a libaom library for building video compression pipelines that require higher computational power (CPU). The second technique used was assigning the appropriate labelling to the videos. For this, a Python script for allocating binary labelling is performed. For example, real videos are (1) and fake videos are (0).

Figure 2: Data Labelling: Real (1) and Fake (0)

3.3 Data Transformation

In this phase of data transformation techniques such as extraction of frames, Face detection and Face cropping, Features extractions and resizing the videos, are implemented. Each of these techniques is explained in detail as follows.

3.3.1 Extraction of Frames

Research by Chen et al. (2022) describes how compressed images of high-quality (HQ), low-quality (LQ), and medium-quality (MQ), can be applied for the extraction of temporal inconsistencies between each image. Similarly, this approach can be further adopted to break the videos into individual images and extract frames from the videos to identify temporal inconsistencies within frames. Other research studies by Kuang et al. (2022); Guan et al. (2023); K et al. (2023); Suratkar and Kazi (2023) have proven how the extraction of the (f) number of frames helps in detecting temporal inconsistencies for convolutional neural network models. The authors' research study extracts 10 frames of equal time from each video. Figure 3 demonstrates an example of frames extracted from videos.

3.3.2 Face Detection and Face cropping

The face-swapping technique is one of the most commonly used deep-fake generation techniques known. The research study by Wu et al. (2023) has stated how deepfake videos are generated with the inclusion of face-cropping and face-swapping techniques. One of the challenges faced while building a deepfake video detection model is the storing and

[3]https://www.ffmpeg.org/

Figure 3: Extracting (f) Frames (videos to images).

managing of data due to its significant file-size. The preprocessing of videos is usually focused on extracting features from the facial area. Cropping the frames by selecting only essential areas from the frames helps in building an efficient and effective model. The previous state-of-the-art models by Kuang et al. (2022); Suratkar and Kazi (2023) introduced the face cropping technique rather than passing the entire data to the model training phase, which achieved promising results in predicting deep fake videos. The authors' own research has used a similar approach for detecting faces using the dlib package from face_recognition library and later cropping the faces from the frames.

Figure 4: Face Cropping.

3.3.3 Features Extraction

In the initial preprocessing phases, the extraction of features from the frames helps in tracking the changes that occur within initial frames and end frames. A few image processing techniques offered by the scikit-image library assist in extracting those features. In this section, this research study focuses on achieving features such as hidden noise from the frames, entropy, which helps in performing error level analysis (ELA) of compressed images, and an unwrapped phase to analyse the inconsistencies in the colour distribution throughout the frames which can additionally help in predicting fake videos. Figure 5 below illustrates the feature extraction visualisation using the scikit-image library. As discussed by K et al. (2023) the hidden noise from the compressed video frames can affect

the model's performance in the research study and focusing on the need to adapt these features is implemented in this study.

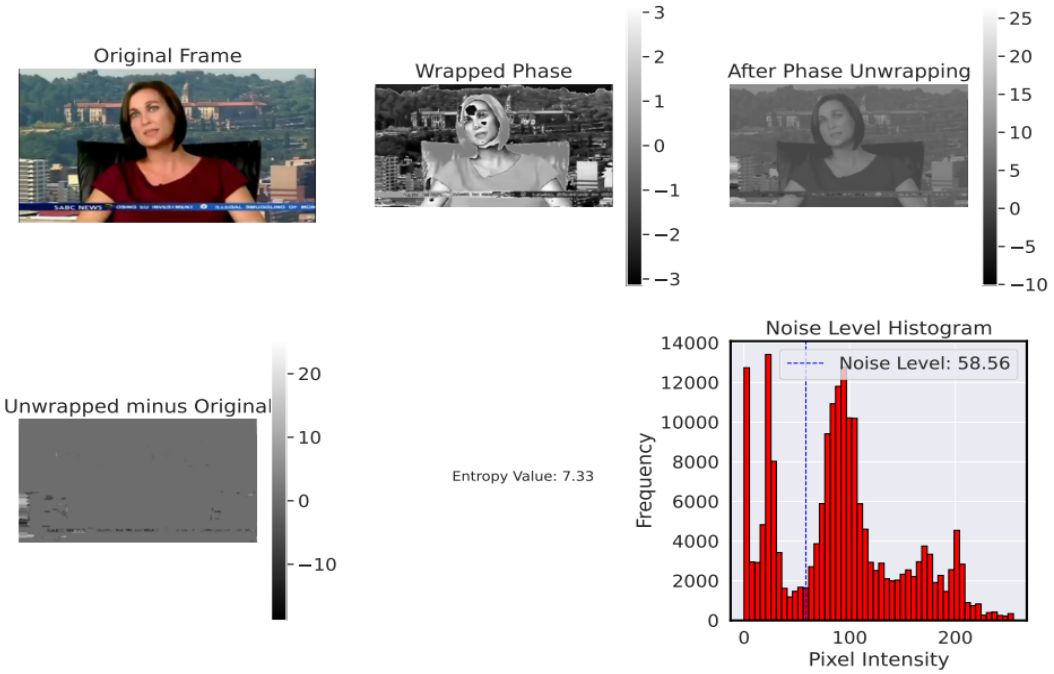

Figure 5: Features Extraction: Original frame, wrapped phase, unwrapped phase, unwrapped phase minus original frame, Entropy, and Noise level.

3.3.4 Resizing and Transformation

In this section, the unbalanced dimension of the frames is handled using the resizing feature. As discussed by Chen et al. (2022); Khedkar et al. (2022); Guan et al. (2023); K et al. (2023) the need for resizing the frames into 224 * 224 network inputs for passing the frames to the EfficientNet-B0 model is essential and requires frame augmentation techniques along with the resizing, such as flipping the frames horizontally or vertically to increase the model training capabilities.

Figure 6: Resizing the frames into 224x224.

3.4 Model Training

After successfully preprocessing the data, this section will discuss the modelling process proposed by the authors' research study. The pre-processed data is split into training and testing data at a ratio of 80:20, where 80% of the data is allocated for training the model and the remaining 20% for testing the model performance using appropriate evaluation metrics. This authors' study adopted two different neural network models—EfficientNet-B0 and Stacked Bi-directional Long Short-Term Memory (LSTM) model, for detecting compressed deepfake videos. The following section will illustrate each of these model structures:

- EfficientNet-B0: It is the pretrained model using ImageNet weights. As discussed in the above sections 2.3 and 2.4, the use of the EfficientNet model by Kuang et al. (2022); Guan et al. (2023); Suratkar and Kazi (2023) for the extraction of features from the frames using the convolutional neural network framework offered prominent model performance. This pretrained model is found to be suitable for deep learning models such as CNN, RNN, LSTM, and Bi-LSTM.

- Stacked Bi-directional Long Short-Term Memory (LSTM): This is one of the ensemble types of models, which is obtained by using two LSTM models simultaneously in opposite directions. The forward and backward LSTM mechanism helps to examine the contextual features of the frames over time. As stated by Khedkar et al. (2022); Kuang et al. (2022); Wu et al. (2023); Hassan et al. (2024), the implementation of the LSTM model for capturing temporal inconsistencies within the frames was found to be advantageous. Khedkar et al. (2022) had a few limitations in not using the pretrained EfficientNet model for extracting the features; this research gap is being covered in this proposed research model by using EfficinetNet-B0 and stacked bi-directional LSTM models. The last layer consists of SoftMax activation functions which help in predicting the probability of the videos being fake or real.

3.5 Model Evaluation

At the final stage of the proposed methodology, this section will discuss the model evaluation techniques followed by this research study to answer the research questions previously stated above. This study aims to build a compressed deepfake detection model capable of handling AV1 compressed video. The result of this model is in the form of binary classification; for example, if the video is real, it will predict real (1), and vice versa for fake videos. The most commonly used evaluation matrices are accuracy, confusion matrix, which helps in examining the false positive (FP), false negative (FN), true positive (TP), and true negative (TN) rates of the model, precision, recall, F-1 score, and cross-entropy loss. Along with all these matrices, this research study evaluated the model performance based on the training and validation losses with respect to the change in the number of epochs. The model performance is compared across three distinct groups based on all of these evaluation matrices: raw videos, videos compressed with the AV1 codec at a low bitrate (250 kbps), and videos compressed with the same codec at a high bitrate (1024 kbps). Later, model fine-tuning is performed to achieve a robust model by preventing model over and underfitting challenges. The next section will discuss the design specifications of the proposed model, which will help in understanding the framework used in detail.

4 Design Specification

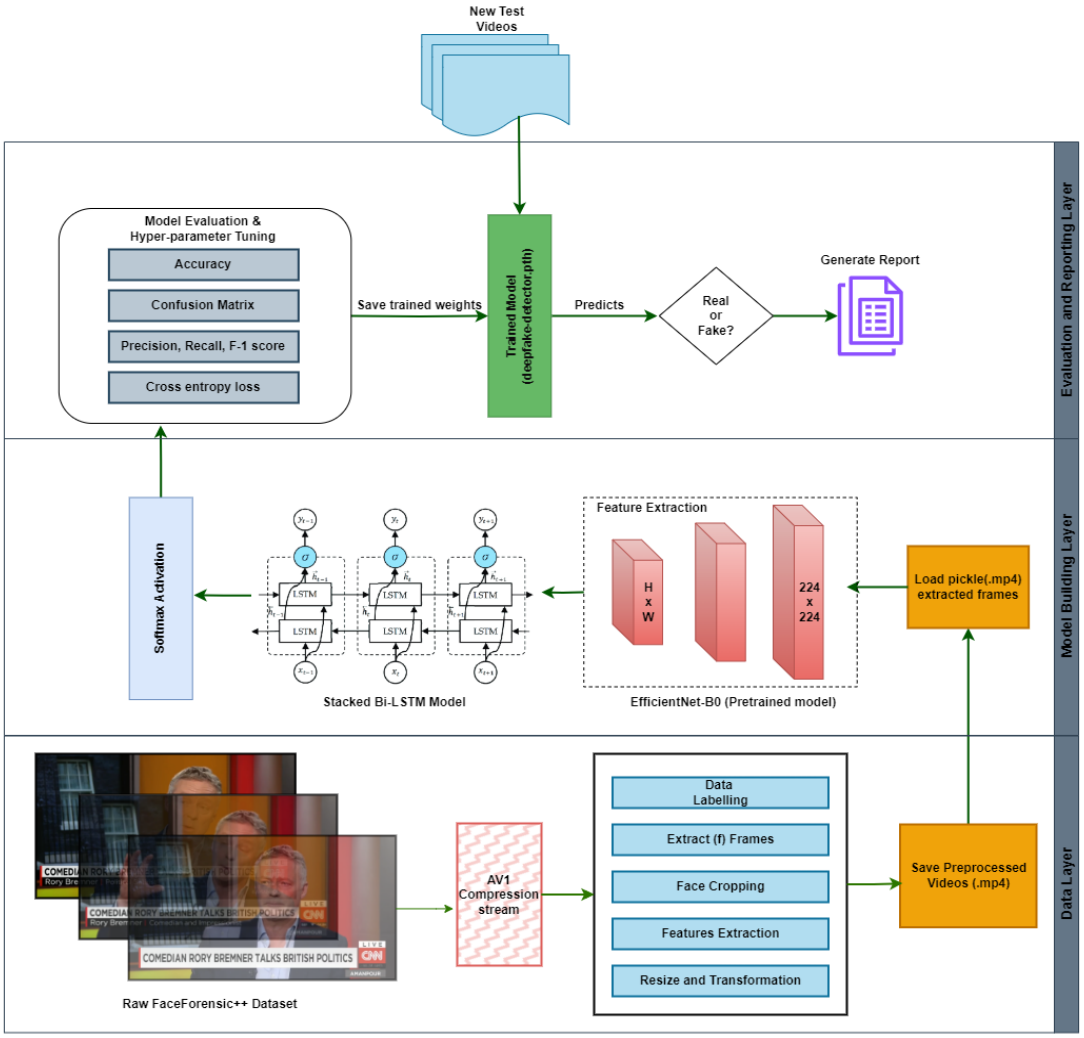

Figure 7: System Architecture

This section will cover the system architecture followed for building the deepfake video detection model throughout the project. The system architecture is divided into three categories—the data layer, the model building layer, and the evaluation and reporting layer—as shown in Figure 7 above. The first layer, known as the data layer, carried out methods for data collection, pre-processing, resizing, and transformation that resulted in storing the face-only videos that were successfully processed and transferred to the next layer for the model-building stage. The model building layer uses these processed videos as the input for the EfficientNet-B0 model (pre-trained on ImageNet), where a single RGB frame is passed at a time to extract the essential spatial features. The output of the EfficientNet-B0 is passed to the 3-layered bidirectional long-short-term memory (LSTM) model with a 1280-dimensional hidden layer and 0.5 percent dropout used as sequential inputs to detect temporal inconsistencies. The final neural network layer consists of a SoftMax activation function that helps in the conversion of the binary classification into probabilities. The final layer, the evaluation and reporting layer, helps in measuring the model performance based on the evaluation matrices. The final fine-tuned model is saved

for predicting whether new videos are real or fake with the help of pre-trained weights highlighting the facial heat map. After successfully predicting the new videos, a report is auto generated that consists of all the features extracted along with the predicted results. The next section will help us to understand about the experimental setup and parameters used to build the final fine-tuned model.

5 Implementation

In this section, the author will discuss the tools considered whilst setting the environment for this research project. Later, the parameters selected while performing hyper-parameter tune of the proposed model to achieve the best results are also discussed in detail.

5.1 Experimental settings

For implementing the proposed model, the author adopted the TPU (tensor processing unit) runtime available through Google Colab[4], which provides additional computational power. The entire code component was implemented using the Python programming language. The use of the Handbrake application for compressing video under the AV1 codec is performed and stored in Google Drive for further pre-processing steps. The extraction of frames from the videos is carried out by ffmpeg, and later, with the use of the face_recognition library, face detection and cropping are performed. The PyTorch deep learning framework is used for building and training the proposed model, and with the help of Torchvison, the pre-trained model is loaded. Lastly, libraries such as Pandas and Matplotlib are used for handling the CSV files and visualising the model evaluation, respectively.

5.2 Parameters

After successfully executing all the pre-processed steps discussed, the processed data is passed through the defined model. This research project used a combination of two models—the EfficientNet-B0 and the stacked bidirectional LSTM—where the EfficientNet-B0 network used the pre-trained weights of the ImageNet for re-training and the stacked bi-directional LSTM with fully connected layers of 1280 x 1280 x 3 stacked layers and a 0.5 percent dropout layer. The model used the ADAM optimizer with a weight decay of 1e-3 and an initial learning rate of 1e-5. The batch size is set to 6, and the model was trained on 25 epochs, where the early stopping was set to stop the training if no improvement was observed after 5 consecutive epochs. The cross-entropy loss function is used for the binary classification and later the SoftMax activation function for the conversion of binary classification into probabilities while predicting new videos.

6 Evaluation

In this last phase of this research project, the model evaluation is carried out to examine the overall performance of the proposed ensemble model. The models are evaluated on three major categories—the raw videos, the videos compressed under AV1 at low

[4]https://colab.google/

bitrate (250 kbps), and the videos compressed under AV1 at high bitrate (1024 kbps)—for performing the comparison study to fulfil the objective of the research question stated above. The evaluation matrices such as accuracy, precision, sensitivity, F1-score, cross-entropy loss, and confusion matrix (which include false positive (FP), false negative (FN), true positive (TP), and true negative (TN)) are discussed in Section 3.5 above are considered for evaluating the model performances. The later part of this section will discuss major findings by performing a comparison of all three experiments with a detailed discussion of the pros and cons of these research studies.

6.1 Experiment 1 (Raw videos)

The first experiment was performed on the selected raw videos. In this instance, the parameters discussed in Section 5.2 are used to achieve effective model predictions. The aim of the experimental analysis was to perform a baseline comparison with compressed video under AV1 at low and high bitrates. Table 1 demonstrates the model performance measured under various matrices such as accuracy, precision, recall, and F1-score, where all matrices are close or above 90 percent, which indicates that the proposed model of efficientNet-B0 and 3-layered bidirectional LSTM is able to distinguish between real and fake raw videos.

Matrices	**Values**
Accuracy	90.178%
Precision	0.9545
Recall	0.8235
F1-score	0.8842

Table 1: Model performance on raw video

Actual	**Predicted**	
	Fake	**Real**
Fake	58	3
Real	9	42

Table 2: Confusion matrix of raw video

Table 2 depicts the confusion matrix on raw videos with a low number of false positive (FP) and false negative (FN) rates.

6.2 Experiment 2 (Low Bitrate videos: 250 kbps)

Considering experiment 1 as one of the baselines for the comparison of models, this experiment helps the author answer the research question of comparison of model performances at various bitrates. In this experiment, the AV1 compressed videos at 250 kbps bitrate

are used for model training and testing. Table 3 below shows that model performance at low bitrate under several matrices outperformed experiment 1 results. Table 4 helps in comparing the confusion matrix of both experiments, where it was observed that the raw video model had a lesser number of false positives (FP) and false negatives (FN) than compressed videos at low bitrates. This finding helps in understanding the effect of compressed videos on the deepfake model.

Matrices	**Values**
Accuracy	91.964%
Precision	0.9615
Recall	0.8772
F1-score	0.9174

Table 3: Model performance on compressed video at low bitrate (250 kbps)

Actual	**Predicted**	
	Fake	**Real**
Fake	52	03
Real	12	47

Table 4: Confusion matrix of compressed video at low bitrate (250 kbps)

6.3 Experiment 3 (High Bitrate videos: 1024 kbps)

The findings from raw and compressed videos at low bitrate are compared with the model performance of compressed videos at high bitrate (1024 kbps) through this experiment. Table 5 shows that the accuracy, F1-score, and precision observed in high-bitrate videos are comparatively lesser than in low-bitrate videos but higher than in raw videos. The recall is observed to be higher than both models, indicating the ability to predict real and fake videos. On the other hand, Table 6, representing the confusion matrix, shows that the false positives (FP) and false negatives (FN) are less than in the low-bitrate model.

Matrices	**Values**
Accuracy	91.07%
Precision	0.8852
Recall	0.9474
F1-score	0.9153

Table 5: Model performance on compressed video at high bitrate (1024 kbps)

Actual	Predicted	
	Fake	**Real**
Fake	49	06
Real	06	51

Table 6: Confusion matrix of compressed video at high bitrate (1024 kbps)

6.4 Discussion: Comparison study

The above sections of model evaluation discussed the proposed model's performance under statistical measures such as accuracy, precision, recall, F1-score, and confusion matrix. All of these parameters are significant to evaluate but can misguide the model evaluation without comparing the model under the cross-entropy loss. The evaluation using cross-entropy loss leverages the comparison on a more in-depth scale by using training and validation loss and accuracy. In this section, the author will discuss the cross-entropy loss to investigate whether the model is a good fit, over-fitted, or under-fitted.

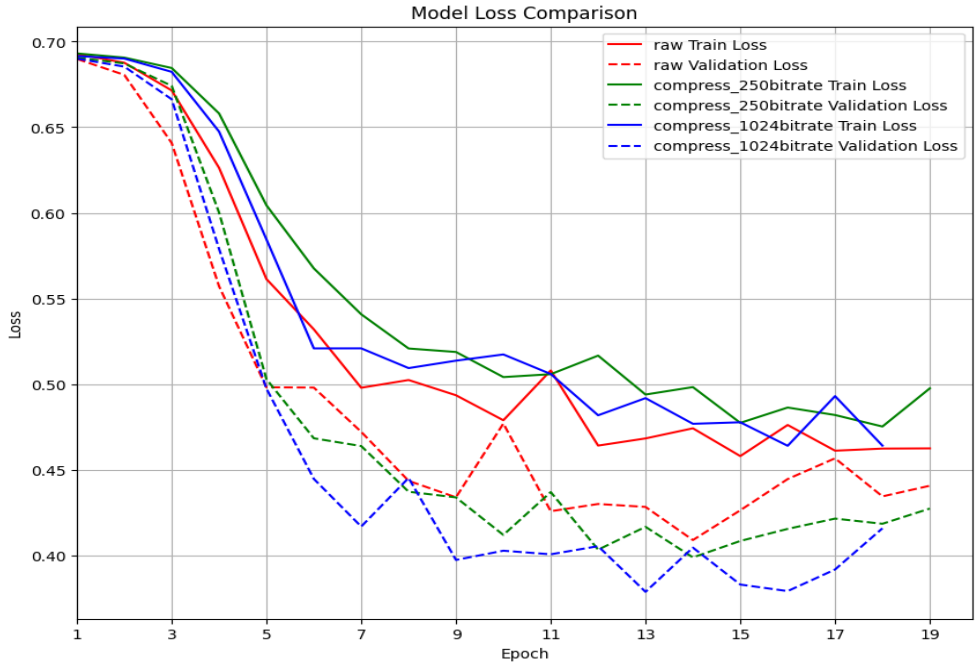

Figure 8: Comparison study of raw, low high bitrate and high bitrate videos on the training and validation loss

Figure 8 above illustrates the comparison of models on the training and validation losses with respect to the increase in the number of epochs. The lower value of loss indicates the model's ability to identify inconsistencies within the frames and distinguish between real and fake videos while training or testing. It was observed that the training loss on raw videos was less than the other two models, and the validation loss of high-bitrate video models was less than the low-bitrate model. These findings help in understanding the increase in complexity of training models with compressed videos at

low bitrates. Additionally, by observing Figure 8, we can further conclude that it was consistently decreasing in training and validation loss by all three models till the epoch 8, which later achieved a plateau of consolidation. The results obtained were until the early stopping was triggered, where no improvement was observed after 5 consecutive epochs.

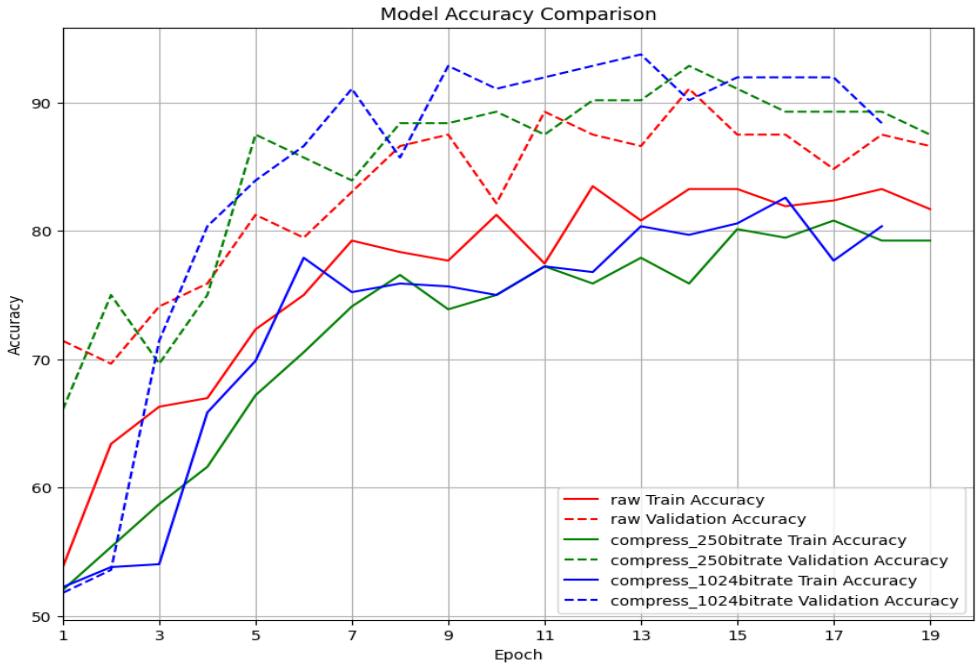

Figure 9: Comparison study of raw, low high bitrate and high bitrate videos on the training and validation accuracy

Similarly, Figure 9, illustrates the changes in training and validation accuracy with the increase in the number of epochs. It was observed that training accuracy on raw videos was higher than the high-bitrate and low-bitrate videos. On the other hand, the validation accuracy of high-bitrate videos outperformed low-bitrate and raw video models. It was observed that the model's loss and accuracy of both training and validation sets were simultaneously decreasing and increasing respectively, which indicated that the models are good fitted for the selected sample dataset. These findings are based on the randomly sampled data which cannot be directly compared with the previous state-of-the-art results mentioned above in the related work Section 2. However, by considering the same size and type of datasets, a later model can be trained and compared with previous models Kuang et al. (2022); Chen et al. (2022); Guan et al. (2023); Wu et al. (2023). The combination of EfficientNet-B0 and 3-layered bidirectional LSTM models increases the strength of model by extracting features in the initial layer and identification of temporal inconsistencies within frames with the help of sequential mechanisms of LSTM layer. The models had to be trained with a restricted quantity of data and processing power, which raised challenges for this study. This resulted in unexpected spikes and drops in loss and accuracy while training and testing the model, as seen in Figures 8 and 9. This research study emphasized accurately training the model by following the prominent methods discussed in the methodology and addressing the effect of compression on deepfake detection models.

7 Conclusion and Future Work

In this paper, a new ensemble model is proposed that uses EfficientNet-B0 for extracting the essential features from the frames and a 3-layered bidirectional LSTM layer for identifying temporal inconsistencies within each frame. The aim of this paper is to showcase the effect of lossy compression techniques—AV1 codecs—on videos, as their purpose is to remove unnoticeable features from the videos to reduce the size of the data. Another major effect of using AV1 compression techniques is to add more unwanted artifacts to the videos or blur the videos, which will increase the model complexity to train the model. These techniques can confuse the model to distinguish between real and fake videos.

The experiments performed in this research project used videos compressed under AV1 at low bitrate (250 kbps) and high bitrate (1024 kbps) along with raw videos of the FaceForensics++ dataset by performing a random sampling technique for selecting 280 real and 280 fake videos for building a model simulation version. Later, only 10 frames from each of these videos were used for training the model. The proposed model achieved over 90% accuracy on all three types of data, where the high-bitrate model outperformed the other two models in terms of accuracy and recall, along with a lesser number of false positives (FP) and false negatives (FN). On the other hand, the low-bitrate video model achieved a higher number of false positives (FP) and false negatives (FN). These help to answer the research question of whether using AV1 compression at low bitrates can increase model complexity and decrease model performance. The strength of this proposed model is that it is adapted to AV1 compressed techniques at low and high bitrates. Although this research has a few limitations, the following could be considered in any future studies.

- Increasing the size of the dataset by using the other datasets mentioned above in Section 2.1 as supplementary data can increase the model's strength while training. Additionally, increasing the number of frames from each can be beneficial. By using the same size of dataset used by the state-of-the-art models, this proposed model would be significant and reliable for performing comparisons among them.
- This model was restricted to videos without audio, which lacks audio-related inconsistencies that would help in enhancing the model's performance.
- The use of the latest powerful deep learning models for building and deploying a real-time model with the help of appropriate cloud resources.

References

Akyazi, P. and Ebrahimi, T. (2018). Comparison of compression efficiency between hevc/h.265, vp9 and av1 based on subjective quality assessments, *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6.

Chen, P., Xu, M. and Wang, X. (2022). Detecting Compressed Deepfake Images Using Two-Branch Convolutional Networks with Similarity and Classifier., *Symmetry (20738994)* **14**(12): 2691. Publisher: MDPI.

Guan, L., Liu, F., Zhang, R., Liu, J. and Tang, Y. (2023). Mcw: A generalizable deepfake detection method for few-shot learning, *Sensors* **23**(21).
URL: *https://www.mdpi.com/1424-8220/23/21/8763*

Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z. and Borghol, R. (2022). Deepfake audio detection via mfcc features using machine learning, *IEEE Access* **10**: 134018–134028.

Hassan, N., Miah, A. S. M. and Shin, J. (2024). A deep bidirectional lstm model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition, *Applied Sciences* **14**(2).
URL: *https://www.mdpi.com/2076-3417/14/2/603*

Hussain, S., Neekhara, P., Dolhansky, B., Bitton, J., Ferrer, C. C., McAuley, J. and Koushanfar, F. (2022). Exposing vulnerabilities of deepfake detection systems with robust attacks, *Digital Threats* **3**(3).
URL: *https://doi.org/10.1145/3464307*

K, V., Ramesh, P., Viknesh, H. and Devanand, S. (2023). Compressed deepfake detection using spatio-temporal approach with model pruning, *Procedia Computer Science* **230**: 436–444. 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023).
URL: *https://www.sciencedirect.com/science/article/pii/S187705092302104X*

Khedkar, A., Peshkar, A., Nagdive, A., Gaikwad, M. and Baudha, S. (2022). Exploiting spatiotemporal inconsistencies to detect deepfake videos in the wild, *2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22)*, pp. 1–6.

Kuang, L., Wang, Y., Hang, T., Chen, B. and Zhao, G. (2022). A dual-branch neural network for deepfake video detection by detecting spatial and temporal inconsistencies, *Multimedia Tools and Applications* **81**(29): 42591–42606. Copyright - © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021; Last updated - 2023-11-29.
URL: *https://www.proquest.com/scholarly-journals/dual-branch-neural-network-deepfake-video/docview/2740204834/se-2*

Rana, M., Murali, B. and Sung, A. (2021). Deepfake detection using machine learning algorithms.

Raveena, Punyani, P. and Chhikara, R. (2023). Comparison of different machine learning algorithms for deep fake detection, pp. 58–63.

Suratkar, S. and Kazi, F. (2023). Deep Fake Video Detection Using Transfer Learning Approach, *Arabian Journal for Science and Engineering* **48**(8): 9727–9737.
URL: *https://doi.org/10.1007/s13369-022-07321-3*

Uhrina, M., Sevcik, L., Bienik, J. and Smatanova, L. (2024). Performance comparison of vvc, av1, hevc, and avc for high resolutions, *Electronics* **13**(5).
URL: *https://www.mdpi.com/2079-9292/13/5/953*

Wu, B., Su, L., Chen, D. and Cheng, Y. (2023). Fpc-net: Learning to detect face forgery by adaptive feature fusion of patch correlation with cg-loss, *IET Computer Vision* **17**(3): 330–340.
URL: *https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.12169*