Inconsistency Detector between Low-Quality Video and Audio Using Deepfake

Muhammad Taha[1], Shakil Ahmed[1], Sana Alam[2], Wazir Ali[2], Asghar Khan[2], Abdul Kahliq[2]

[1]Sir Syed University of Engineering and Technology, Computer Engineering Department, Karachi Pakistan
[2]College of Computer Science and Information Systems, Institute of Business Management (IoBM), Karachi, Pakistan

Corresponding Author: Sana Alam sana.alam@iobm.edu.pk

Received: 11-05-2024; **Accepted:** 18-07-2024; **Published**: 30-12-2024

Abstract: This study presents the challenges and issues within low-resolution video and audio through the application of sophisticated deep learning methodologies, addressing the prevalent issue of manipulated media in recent times. By employing dual-stream convolutional architecture, we analyze the intricate relationship between auditory and visual cues, commencing with an extensive examination of existing detection approaches and their constraints when confronted with substandard video content. Utilizing the VidTimit dataset as our base, we train and test our model and check its performance in comparison to existing pre-trained models. Our evaluation framework includes accuracy metrics, confusion matrices, and F1 score to ensure efficiency. Using a variety of filters such as Edge Preserving and Gaussian Blur on video data preprocessing, we enhance the detection of disparities by optimizing the input data. Model integration is the hallmark of our innovative dual-stream convolutional architecture, where audio and visual components are perfectly integrated. In this architecture, the visual stream applies convolutional layers to capture spatial characteristics from low-quality video frames, and the audio stream applies RNNs to capture temporal patterns in audio signals. The fusion module effectively integrates these streams and makes way for synchronization analysis and anomaly detection. In this aspect, our trained network does very well in active video detection and lip-reading tasks.

Keywords: Recurrent Neural Networks, Artificial Intelligence, Deepfake

1. Introduction

In this advanced time, the issue of media manipulation in terms of fake news and false information is very evident [1]. Media manipulation in terms of generating fake and manipulated data leads to the generation of fictious information [2]. The terminology “Deepfake” is the combination of other two terms: Deep Learning and Fake and is related with the contents that appear to be realistic but that are fake [3]. In 2017, an anonymous Reddit user applied the techniques and methods of Deep Learning to generate realistic pornography video by using fake images of a person’s face [4]. Some of the deepfake applications that can very easily generate fake data are FaceApp [3], FakeApp [3], ZAO [5]. ZAO generates fake images that manipulate the original image by changing various features of the face [5]. Moreover, FaceApp and FakeApp are used to switch faces and modify different facial features. Furthermore, FaceApp also creates such images that appear to be older [4].

In short, deepfake refers to those images, audios, videos that seem to be real but they actually they are the fake contents that are generated by using AI techniques and models, particularly deep learning techniques are employed to create such seemingly realistic content [4]. As we investigate the concept of deepfake, our primary objective is twofold: to learn about the strengths and weaknesses of current deepfake methods for distinguishing fakes from real ones, as well as new methods. These strategies cover different methodologies, including, but not restricted to, profound learning models, signal handling procedures, and criminological investigations. With the increase in risks in deepfakes, it becomes vital to assess these procedures or techniques to recognize the fake contents, especially when the recordings are of low-quality. As a result, to identify the gaps in the current research environment and pave the way for more robust and dependable deepfake detection systems, it is essential to conduct a comprehensive analysis of these strategies' advantages and disadvantages.

2. Related Work

A. Traditional Approaches

This section discusses review various methods proposed for inconsistency detection in low-quality video and audio with deepfake content. This involves deep learning methods, feature-based techniques, as well as signal-processing approaches that have been extensively researched in recent times. The learning patterns of deep models in detecting subtle inconsistencies between these are capable of complex data-pattern recognition, making them apt to perform the task. Feature-based methods rely on manually selected features, whereas signal-processing approaches analyze temporal and spectral characteristics of video and audio. By looking at these methodologies, this section is intended to highlight the effectiveness and limitations of each approach in detecting inconsistencies in low-quality deepfake content.

Different methodologies have been suggested and explored for the detection of inconsistencies between low-quality video and audio in deepfake content. Each of these approaches utilizes different techniques, ranging from deep learning-based methods to traditional feature-based and signal-processing techniques. Each methodology has unique advantages and challenges, making it important to understand their characteristics.

B. Deep Learning-based Approaches

Recently, deep learning techniques, mainly including CNNs and RNNs, have achieved wonderful success in the broad aspects of computer vision and NLP tasks. In terms of inconsistency detection, video frames are more likely to have features extracted by using CNN, while audio signals can involve temporal dependencies to be learned by RNNs. Combining visual and auditory features allows deep learning models to efficiently detect inconsistencies between the video and audio parts of deep fake content. Among the notable researchers studying the complex nature of deep fake technology, Dagar and Vishwakarma [6] did a comprehensive literature review and provided very valuable insights into deepfakes, including their generation, detection, and possible applications. Zhang [4] also added something significant to the discussion with a deepfake generation and detection survey wherein the state-of-the art techniques of that domain had been thrown under the lime light. To contrast it, Ahmed et al. [7] had come up discussing the deeper levels of such false information affecting the whole modern society: it further went to its systematic impact and remedies followed so far. Kingra, Aggarwal, and Kaur [8] have provided a survey for the emergence of deepfakes and video tampering detection approaches. This was helpful in discussing the concerns of the ever-growing and evolving issue of deepfakes. Shelke and Kasana [9] has conducted an exhaustive survey about the passive technique for detecting digital video forgery and added to the knowledge corpus of detecting forged content.

C. Feature-Based Approaches

Feature-based methods depend on handcrafted features that are extracted from video and audio data for detecting inconsistencies. Such features could be texture patterns, color histograms, motion descriptors, and audio characteristics such as pitch and intensity. Feature-based techniques are usually interpretable, enabling researchers to understand the specific cues the model uses to identify inconsistencies. However, these may be insufficient in describing intricate patterns and can face the challenge of realistic variability in the deepfakes. Among the active researchers in the area of deepfake detection, Dr. Siwei Lyu from the University at Albany, SUNY, is an accomplished researcher who has explored numerous techniques under his umbrella. The works of Dr. Lyu include deep learning-based techniques, feature-based methods, and signal processing for detecting inconsistencies in deepfake content. He has been making major contributions to the development of deepfake detection algorithms, utilizing CNNs and RNNs to analyze video and audio streams in parallel. Dr. Lyu's work also explores traditional feature-based methods, such as texture analysis and motion descriptors, for signs of manipulation. He has further researched signal processing techniques for analyzing temporal and spectral characteristics that help in identifying artifacts in deepfake videos. Focusing on multimodal analysis and ensemble modeling, Dr. Siwei Lyu's research provides a very valuable insight into the strengths and limitations of different approaches, furthering the field of deepfake detection and media forensics. He received much recognition for the work done, and further research ensued to combat the increasing deepfake technology challenges.

D. Signal Processing based Approaches

Signal processing techniques analyze the time and frequency aspects of video and audio signals for the presence of anomalies. Techniques such as audio watermarking and video compression analysis identify signs of tampering. Moreover, these techniques can be used to detect misalignment between video and audio tracks, which is another major indicator of deepfake content. Wang [10] also investigated the temporal dependencies and synchronization differences. This sheds light

on the artefacts and inconsistencies that appear when generating deepfakes. These two studies both stress the use of audio as well as visual clues for all-around detection.

E. Hybrid Approaches

Some researchers have explored hybrid approaches that combine multiple techniques to improve the overall effectiveness of inconsistency detection. For example, a hybrid model might use deep learning for feature extraction and then apply signal-processing techniques for further analysis. By combining the strengths of different methodologies, hybrid approaches aim to enhance accuracy and robustness in detecting low-quality deep fake content. These researchers' work has significantly advanced the field of deepfake detection and related areas. Dagar and Vishwakarma's [6] literature review offers a holistic view of the existing research, helping researchers and practitioners navigate the vast landscape of deepfake generation, detection, and applications. Zhang's [4] survey provides a comprehensive overview of the current state of deepfake techniques, enabling researchers to stay up to date with the latest advancements in the field. Ahmed et al.'s [7] systematic review of false information sheds light on the pervasive impact of misinformation, prompting discussions on combating digital disinformation. Kingra, Aggarwal, and Kaur's [8] survey on video tampering detection approaches contributes to strengthening the ability to identify and counter deep fake content. Shelke and Kasana [9] contribute valuable insights into passive techniques for detecting digital video forgeries, adding to the toolbox of forensic tools available to combat the rise of manipulated media.

To conclude, the related work highlights potential future directions for developing more robust inconsistency detection techniques. Advancements in deep learning, multimodal learning, and data augmentation offer promising paths to overcoming existing limitations. The importance of interdisciplinary collaborations between AI, multimedia processing, and cybersecurity experts is emphasized to address the complex challenges posed by deepfake content. As deepfake technology continues to advance, ongoing development of detection systems is essential to ensure the integrity of digital media and mitigate the risks of misinformation.

3. Methodology

The structured and meticulously designed systems intended to achieve our research goals are described in detail in the methodology section of our study. Our primary objective is to analyze the capacities of a dual-stream convolutional organization to handle the different issues related to modified sight and sound substance. We employ various methodological strategies and meticulously examine several crucial factors for comprehensive and meaningful insights. The basic components of our proposed approach incorporate:

- Development of Lip-Sync Error Detection: Embedding both audio and visual elements.
- Enhancing Versatility for Video Detection: Utilizing relevant metrics such as accuracy and efficiency in audio-visual identification.
- Lip-Sync Performance Evaluation: A comprehensive examination of lip-matching precision
- Complete Assessment: top-to-bottom evaluation of different networks.

Data Collection and Preprocessing, Model Architecture, Filtering Techniques, and Limitations and Future Directions are the steps that make up this methodology and will all be discussed in depth. Each step is a basic part of our methodology, intended to address every part of our exploration targets deliberately.

A. Data Collection and Pre-processing

The VidTimit dataset was the starting point for the project's comprehensive and diverse data collection. This dataset consolidates both short clips and groupings of mouth pictures got from an extent of video settings. Unlabeled information assumes a pivotal part in preparing a joint implanting model that means to cover a large number of circumstances and visual markers, which is the reason this dataset was picked. The mouth pictures were extracted using high-level techniques to ensure precise arrangement and extraction in order to maintain the connection between the sound and visual elements. The sound bites and mouth pictures were unequivocally synchronized to lay out significant associations for the joint installing process. The flexibility of the model is revealed by standardized the data and lift the model.

B. Dataset

The substance of exhaustive assessment lies in utilizing benchmark datasets that reflect complex genuine circumstances. Standard benchmark datasets were chosen to allow for accurate comparisons and establish new benchmarks. Adjusting these datasets with the objectives of the approach guaranteed that the assessment tended to genuine difficulties, in this

manner approving the outcomes. The VidTIMIT dataset [11] involves video and sound accounts of 35 people, at first expected 43, yet a few connections were absent. These recordings feature individuals uttering short sentences and are crucial for research in areas like automatic lip-reading, multi-view face recognition, multi-modal speech recognition, and individual identification.

The dataset's recordings are detailed in Table 01, and its structure is depicted in Figure 01. The recordings are divided into three sessions, held about a week apart between the first two and six days between the last two. The sentences, sourced from the TIMIT corpus's test section, involve each participant reciting 10 sentences. The first six sentences, sorted alpha-numerically, are allocated to the first session, the next two to the second session, and the final two sentences are part of the third session.

Section ID	Sentence ID	Sentence text
Session 1	sa1	She had your dark suit in greasy wash water all year
	sa2	Don’t ask me to carry an oily rag like that
	si1398	Do they make class-biased decisions?
	si2028	He took his mask from his forehead and threw it, unexpectedly, across the deck.
	si768	Make the lid for the sugar bowl the same as the jar lids, omitting the design disk.
	sx138	The clumsy customer spilt some expensive perfume.
Session 2	sx228	The viewpoint overlooked the ocean.
	sx318	Please dig my potatoes up before frost.
Session 3	sx408	I’d ride the subway, but I haven’t enough change.
	sx48	Grandmother outgrew her upbringing in petticoats

Table 1 VidTIMIT dataset Example of the sentences used in it

The initial two sentences are identical for every participant, whereas the latter eight sentences vary and are customized according to each participant's involvement. The audio that goes with these is carefully saved in mono 16-bit WAV format, with a sampling frequency of 32 kHz. Such detailed structuring of the dataset makes it highly appropriate for various research projects.

Figure 1 Example of Subjects in VidTIMIT database [11]. The first, Second and third columns represent images taken in sessions 1, 2 and 3 respectively in Table 1.

C. Model Architecture

This section delves into the intricate design and structure of the dual-stream convolutional network used in our study. We describe the specific layers, nodes, and connections within the model, highlighting how each component contributes to processing and analyzing multimedia content. The architecture's uniqueness lies in its ability to simultaneously handle audio and visual data streams, ensuring a comprehensive analysis.

We make sense of the reasoning behind the decision of convolutional layers, actuation capabilities, and the general organization configuration, planning to clarify how these components work as one to recognize irregularities in sight and sound substance. Our procedure depends on a state-of-the-art model design created to distinguish disparities in bad-quality video and sound materials. Utilizing both convolutional and recurrent neural networks to identify intricate spatial and temporal correlations in audio and visual signals, this architecture is based on deep learning principles.

D. Dual-Stream Convolutional Network Design

We have used dual-stream convolutional Network Design [12]. Through innovative joint embedding, this method connects these modalities strategically and provides a dynamic depiction of the relationship between audio and visual elements. The architecture, based on the ideas in [2], cleverly combines mouth and sound imagery by taking advantage of their inherent synchronicity. This system is shown in Figure 2.

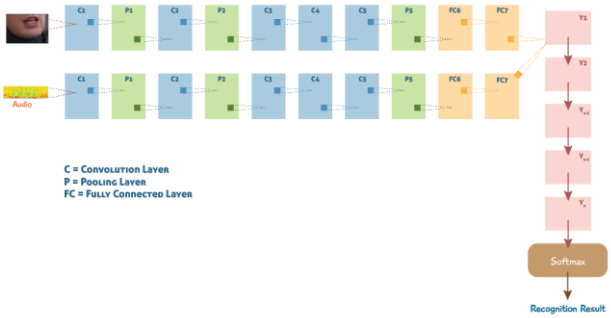

Figure 2 Model Architecture

E. Training with Unlabeled Data

The preparation routine for the dual-stream convolutional network incorporates a few phases. The network is then applied to unlabeled data that includes audio and synchronized mouth imagery. To avoid issues with gradient diminution and accelerate convergence, it begins with weights that have already been loaded. The model can recognize hidden away relationship among hearable and visual signs through an intensive course of forward and switch expansion. This stage puts significant solid areas for a performance getting, utilizing the unseen chance of information present in unlabeled data. The model secures a joint implanting cap synchronized ring in this stage, permitting it to look at and gain from the nuances of

sound and visual collaborations. This stage's model lays out the establishment for ensuing examination, upgrading the organization's ability to recognize unpretentious anomalies.

F. Synchronization Error Analysis in Audio-Visual Data

Through synchronization error the audio-video data feed to the algorithm’s expertise now includes detecting synchronization errors in audio and video content. The task is associated with identifying differences between visual and auditory timetables using the knowledge gained from joint implants. The lip-sync [13] can make mistakes in measurements got from this examination give a complex proportion of video legitimacy, filling in as basic signs of expected content control.

G. Broadening Horizons: Active Speaker and Lip-Reading Capabilities

The task is to handle the adaptability if the active speaker in detection and lip reading in addition to synchronization error detection. The essential part of this detection is to use the network and uses the audio-visual processing [14] capabilities by locating it and shifts it in the speaker activity. The lip-syncing from the lip movements while speaking is totally based on visual cues, network joint the embedding proficiency in critical.

H. Setting New Standards in Benchmark Analysis

This' organization will likely lay out new guidelines for standard datasets. The double stream convolutional organization's strength and versatility are featured by this accomplishment, exhibiting its expansive pertinence. The association's remarkable presentation in tasks like powerful speaker revelation and lip scrutinizing implies its historic potential in the media region.

I. Filters Techniques

The part covers strategies like sound decrease, signal upgrade, and information standardization, which are essential in setting up the information for input into the model. We also discuss how these methods help the model be more accurate and efficient, especially when working with low-quality audio and video. The quest for improving the presentation of the SyncNet [11][12][15] model in recognizing peculiarities in low-quality sound and video includes a basic step: applying different channels. This stage is intended to survey the effect of various separating methods on the model's accuracy and viability, helping with distinguishing the most appropriate channel for accomplishing precise outcomes.

In applying filter techniques, the research methods are applied to refine the work on the data for progress. The techniques are applied to cover the sound reduction, signal overhaul and normalization of the data because these are fundamental in setting up the data for applying in input to the model. The model is more precise and proficient by applying the strategies to assist the model while working with inferior quality sound and video. The SynNet model is basically perceiving characteristics in bad quality sound and video incorporates a fundamental stage: applying various channels. The purpose of this stage is to examine how various separating techniques affect the model's accuracy and viability, assisting in determining the best path to precise results.

J. Selection of Filters

Each handling strategy remarkably affects the video information, so a determination of channels has been painstakingly decided to cover them all. The span integrates Edge Improvement [16], Gaussian Cloudiness [17], Center Separating [18], High-Pass Filtering [19], and Low-Pass Sifting [16]. The objective of the preprocessing stage is to lessen commotion, uncover applicable elements, and set up the information for proficient SyncNet model investigation. Each channel carries special characteristics to the interaction.

K. Application of Filters and Data Preparation

These picked channels are purposely applied to the video data in the dataset. A particular channel is utilized to modify the visual components in every video, creating modified renditions of the first happy. Convolutional tasks are utilized in this sifting strategy to change the picture's pixels by the channel's attributes.

A significant period of information planning follows the separating. This step guarantees that the sifted recordings meet the SyncNet model's feedback prerequisites. Endeavors, for instance, resizing, normalization, and course of action are performed to ensure consistency and closeness with the model's designing. The coordinated accounts then, go about as commitments to the following examination, considering a comprehensive evaluation of the channel's impact on the model's show.

L. Acknowledging Limitations and Future Prospects

To completely understand the procedure, recognizing its limitations is important. The methodology's suitability depends on having access to high-quality data, so it may have trouble in situations where there isn't a lot of data available. While setting new standards is admirable, it is essential to conduct additional research into the model's applicability across a variety of settings and types.

M. Consideration of Ethical Implications

In the closing segment, we recognize the constraints of our ebb and flow research and recommend future headings. Information quality, model adaptability, and computational assets are among the limitations we talk about. This part in like manner considers the potential inclinations in the dataset and how might affect the results.

Due to the potential dangers posed by manipulated content, strict ethical standards were upheld throughout the research. Issues associated with insurance, consent, and incidental effects were meticulously considered, ensuring the assessment adhered to reliable practices by the propelling moral scene. Information assortment, creative design, top to bottom assessment, and moral care are undeniably illustrated in this review. This strategy is a brilliant illustration of state-of-the-art research, reinforcing the believability and realness of media content as the computerized world keeps on evolving.

4. Results and Discussion

The results of our research are presented in this chapter, the essential point was to look at the exhibition of our model by directing a thorough examination with different laid out pre-prepared models, highlighting its viability intending to the difficulties presented by bad quality edges.

The VidTimit dataset [11], a large collection of video content with various speech patterns and quality levels, was used in our investigation. We carried out preprocessing procedures like alignment, resizing, and normalization to bring the dataset into line with the goals of our investigation. This preliminary stage was critical in guaranteeing the information adjusted to the model's prerequisites and laying out a uniform norm for assessment.

A. Analysis of Comparative Models

The main part of this study is a thorough comparison to find out how well our two-stream convolutional architecture for detecting differences in low-quality audio and video works. This expected an exhaustive correlation with existing models and procedures in the field. This examination was vital in distinguishing the model's assets and likely regions for development. We benchmarked our methodology against different difficulties in controlled sight and sound substance to evaluate its true relevance. As shown in Figure 3, comparing our lip-sync model to a variety of pre-trained models was an important part of our study. This empowered us to recognize the unmistakable benefits and limits of our model contrasted with others. Using a standard dataset for this analysis ensured an equitable evaluation and strengthened the foundation for significant insights.

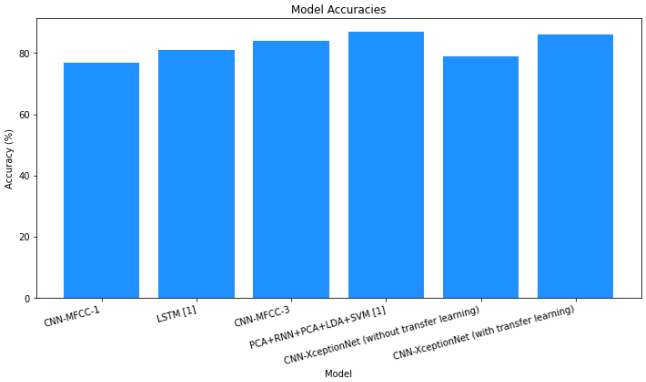

Figure 3 Comparison of Model Accuracies

S. No	Model	Accuracy
1	CNN-MFCC-1	77%
2	LSTM	81%
3	CNN-MFCC-3	84%
4	PCA+RNN+PCA+LDA+SVM	87%
5	CNN-XCEPTIONNET: WITHOUT TRANSFER LEARNING	79%
6	CNN-XCEPTIONNET: WITH TRANSFER LEARNING	86%

Table 2: Analysis of Model Performances

Table 2 showcases the presentation of a few calculations; each surveyed for their capacity to distinguish irregularities in bad-quality video and sound. The CNN-MFCC-1 calculation showed a 77% precision, demonstrating its viability in recognizing explicit information designs. The LSTM model's ability to recognize temporal data dependencies was demonstrated by its 81% higher accuracy. With an accuracy of 84%, the CNN-MFCC-3 algorithm demonstrated the feature detection advantages of a deeper convolutional network.

The consolidated methodology of PCA+RNN+PCA+LDA+SVM accomplished 87% exactness, showing the collaboration of different insightful strategies. The CNN-XceptionNet model, with and without move learning, exhibited accuracies of 79% and 86%, individually, featuring the effect of applying pre-prepared model information to improve execution. To detect inconsistencies in multimedia content, this evaluation provides a nuanced perspective on each algorithm's strengths and areas for improvement.

B. Metrics for Evaluation

After the videos had been preprocessed, we looked at how well the SyncNet model did with each filter. Using these filtered videos as a starting point, the model was trained and tested against the unfiltered baseline model. We used measurements like exactness, accuracy, review, and F1 score for a quantitative evaluation, as displayed in Figure 4.

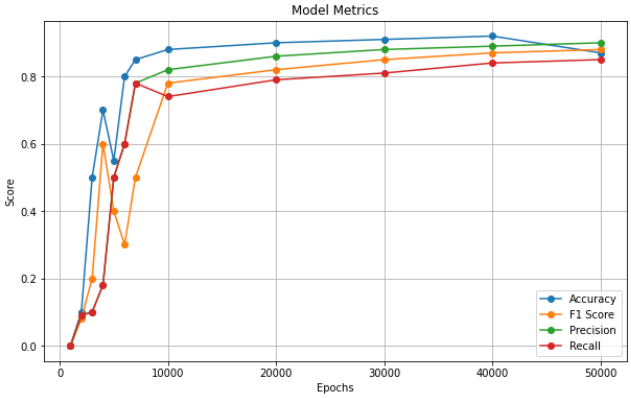

Figure 4 Performance of Model

The efficiency of each filter in improving the model's ability to distinguish between low-quality video and audio is revealed through this in-depth process of filter application, preprocessing, and model evaluation. The near examination perceives the commitment of explicit channels towards precision improvement, sound decrease, or their immaterial effect on execution.

The utilization of shifted channels presents a fundamental part of revelation and advancement inside our approach. We discovered how these filters affect the performance of the SyncNet model by exposing it to various preprocessing methods. This helped us achieve our overarching objective of robustly detecting inconsistencies in multimedia content. Diverse quantitative metrics were used to evaluate lip-sync error detection, active speaker detection, and lip reading. Mean Absolute Error (MAE) metrics and frame-level accuracy were utilized for lip-sync error detection. Dynamic speaker location was evaluated utilizing accuracy, review, and F1-score, while lip perusing's exactness was estimated through a word-level arrangement with the expressed substance.

C. Investigating Current Architectural Models

A basic part of our review included a careful examination of existing structures relevant to lip-sync demonstration. This investigation gave important experiences into their primary standards and adequacy across different situations, illuminating our methodology and laying the basis for future upgrades.

D. Refining the Model Techniques

A focal objective of our review was to refine the current model through imaginative video pre-handling methodologies. We were able to improve the model's performance, particularly when working with low-quality frames, thanks to the lessons learned from our examination of existing architectures and our specialized mouth detection method.

E. Confusion Matrix

We can get metrics like accuracy, precision, recall, and the F1 score using the confusion matrix in Figure 5. It gives a comprehensive view of how well each algorithm does at identifying true positives, true negatives, and false positives. This examination distinguishes the qualities and regions for the development of every calculation.

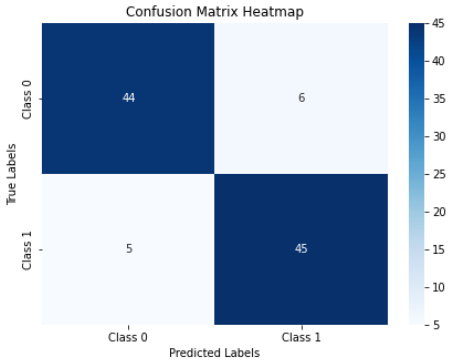

Figure 5 Confusion Matrix

V. APPLICATION AREAS

There are numerous potential applications for the developed deepfake detection framework:

A. Surveillance and Security

Low-quality video detection can help to identify anomalies or suspicious activities. Analyzing video quality in surveillance systems [20] can improve the detection of unusual or potentially threatening behavior. Moreover, low-quality video or audio detection can help trigger the alerts in security systems. The quality detection in security systems can enhance the reliability [21, 22] of alarm mechanisms. Identification of the low-quality media can be used to investigate crimes or gather evidence. The research by [23] indicates that analyzing and improving media quality plays a vital role for accurate forensic investigations.

B. Social Media

Social media channels broadcast multimodal data which provide a ground for rapid propagation of news [25] as well as other misleading information. Such large amount of fake news available online have the potential to trigger serious problems [26] at an individual level as well as in the society. Moreover, identifying low-quality media can help detect fabricated or manipulated content. News as well as social media organizations struggle for the identification and combating visual disinformation presented to their audiences. Such processes are complicated due to the enormous number of media items being produced, how quickly media items spread, and the most of the times deceptive edits remain invisible to the naked eye. Furthermore, media firms can utilize the instruments in order to confirm the credibility or the quality of video content before spread. One of the main reason behind the generation of such content are freely available softwares on the web allow any individual, without special skills [27], to create very realistic fake videos. Such shared data can be used to manipulate public opinion in many daily life situations. Hence, there is a need for such tools with the capability to detect false multimedia content in order to avoid the spread of false information.

C. Quality Enhancement

Online platforms can improve their content moderation systems by incorporating the framework for identifying and removing harmful Deepfake content, thereby safeguarding users and preserving community standards. The quality assessing of video and audio content is help to identify and address issues. A number of approaches are available for the quality assessment to pinpoint and rectify multimedia issues [28]. Moreover, detection of the low-quality regions is helpful to improve the quality of media files. Identification and correction of low-quality areas is important for restoration and enhancement of multimedia content [29].

D. Digital Forensics

Online entertainment organizations can carry out the structure to screen and banner controlled content, defending clients against deepfake-based tricks and falsehood. Video as well as audio tampering detection can reveal in case of editing or alteration. Various techniques have been proposed for detection of tampering [30]. Low-quality media may also contain fabricated or manipulated content. The detection of such inconsistencies in media quality is crucial for identifying fake news [31], as these inconsistencies can be useful to find thee alterations or fabrication. Moreover, low-quality deepfakes can be detected to prevent their spread and misuse. Low-resolution in videos can be key indicators in the detection of deepfake content [32]. Therefore, detecting and analysis of low-quality video as well audio files can be used as an evidence.

E. Cyber Security

Deepfakes can effectively be employed to boost phishing attacks by producing believable yet fabricated video or audio content. Identifying these deepfakes is essential to safeguard both individuals and organizations from cyber fraud [33]. Additionally, deepfakes can be employed to mimic individuals for harmful intentions. Technologies for detection are fundamental in preventing identity theft and unauthorized access to confidential information [34].

5. Limitation and Future work

Our method has a few drawbacks, despite the promising outcomes. First, the quantity and quality of the training data significantly impact the model's performance. The model's generalizability may be compromised if the dataset does not adequately represent the variety of real-world scenarios. Second, the current framework primarily targets the detection of deepfakes in low-quality video environments, which may restrict its applicability to high-quality deepfakes employing more advanced methods. Thirdly, our model could fail with videos that naturally have audio-visual mismatches, such as poor lip synchronization that are not of malicious intent but due to technical flaws. In resource-poor settings, the model's training and deployment would also call for an unnecessary number of computational resources.

These limitations can be overcome in future research by exploring multiple options. First, this model will be more robust and generalizable with the wider variety of video qualities and contexts included in the dataset. Higher quality deepfakes detection capability can be improved with advanced deep learning architectures like transformer models. By coming up with ways to discriminate between such adversarial mismatches and natural mismatches, the model's accuracy is expected to increase even more. To further make the model available to real-time applications, particularly when resources are limited, optimization to be deployed on edge devices is possible. Ultimately, our framework could help build a more holistic strategy to address digital misinformation if it can be integrated with existing infrastructures for cybersecurity.

6. Conclusion

A novel deepfake detection framework was created to spot discrepancies between low-quality audio and video content in this study. We demonstrated that deepfake manipulations can be detected using cutting-edge machine learning methods even when video quality is compromised during compression or transmission over low-bandwidth networks. The fact that our model was able to tell the difference between genuine and manipulated media with such high precision demonstrates its robustness and effectiveness. The findings support our hypothesis that audio-visual discrepancies, especially in synchronization and semantic coherence, can be reliable indicators of deep-fake content.

References

[1] Borges, L., Martins, B. and Calado, P., 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. Journal of Data and Information Quality (JDIQ), 11(3), pp.1-26.

[2] Aldwairi, M. and Alwahedi, A., 2018. Detecting fake news in social media networks. Procedia Computer Science, 141, pp.215-222.

[3] Rana, M.S., Nobi, M.N., Murali, B. and Sung, A.H., 2022. Deepfake detection: A systematic literature review. IEEE access, 10, pp.25494-25513.

[4] Zhang, T., 2022. Deepfake generation and detection, a survey. Multimedia Tools and Applications, 81(5), pp.6259-6276.

[5] de Seta, G., 2021. Huanlian, or changing faces: Deepfakes on Chinese digital media platforms. Convergence, 27(4), pp.935-953.

[6] Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: generation, detection, and applications. International journal of multimedia information retrieval, 11(3), 219-289.

[7] Ahmed, M., Bachmann, S., Martin, C., Walker, T., Rooyen, J., & Barkat, A. (2022). False Information as a Threat to Modern Society: A Systematic Review of False Information, Its Impact on Society, and Current Remedies. Journal of Information Warfare, 21(2), 105-120.

[8] Kingra, S., Aggarwal, N., & Kaur, N. (2023). Emergence of deepfakes and video tampering detection approaches: A survey. Multimedia Tools and Applications, 82(7), 10165-10209.

[9] Shelke, N. A., & Kasana, S. S. (2021). A comprehensive survey on passive techniques for digital video forgery detection. Multimedia Tools and Applications, 80, 6247-6310.

[10] Wang, J., Li, Z., Zhang, C., Chen, J., Wu, Z., Davis, L. S., & Jiang, Y. G. (2022). Fighting Malicious Media Data: A Survey on Tampering Detection and Deepfake Detection. arXiv preprint arXiv:2212.05667.

[11] Sanderson, C. The vidtimit database (No. REP_WORK). IDIAP. 2002.

[12] Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. Id-reveal: Identity-aware deepfake video detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 15108-15117.

[13] Khalil, S. S., Youssef, S. M., & Saleh, S. N. iCaps-Dfake: An integrated capsule-based model for deepfake image and video detection. Future Internet. 2021. 13(4), 93.

[14] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. An Improved Dense CNN Architecture for Deepfake Image Detection. IEEE Access. 2023. 11, 22081-22095.

[15] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. An Improved Dense CNN Architecture for Deepfake Image Detection. IEEE Access, 2023. 11, 22081-22095.

[16] Shahzad, S. A., Hashmi, A., Khan, S., Peng, Y. T., Tsao, Y., & Wang, H. M. Lip Sync Matters: A Novel Multimodal Forgery Detector. In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2022. (pp. 1885-1892). IEEE.

[17] Korshunov, P., & Marcel, S. Vulnerability evaluation and detection of deepfake videos. In 2019 International Conference on Biometrics (ICB). 2019. (pp. 1-6). IEEE.

[18] Chung, J. S., & Zisserman, A. Out of time: automated lip sync in the wild. In Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 251-263). Springer International Publishing.

[19] Korshunov, P., & Marcel, S. Speaker inconsistency detection in tampered video. In 2018 26th European signal processing conference (EUSIPCO). 2018. (pp. 2375-2379). IEEE.

[20] Bhakat, S., & Ramakrishnan, G. (2019, January). Anomaly detection in surveillance videos. In Proceedings of the ACM India joint international conference on data science and management of data (pp. 252-255).

[21] Hoh, B., Gruteser, M., Xiong, H., & Alrabady, A. (2006). Enhancing security and privacy in traffic-monitoring systems. IEEE Pervasive Computing, 5(4), 38-46

[22] Chyan, P. (2019, October). Design of intelligent camera-based security system with image enhancement support. In Journal of Physics: Conference Series (Vol. 1341, No. 4, p. 042009). IOP Publishing.

[23] Pedapudi, S. M., & Vadlamani, N. (2023). Digital forensics approach for handling audio and video files. Measurement: Sensors, 29, 100860.

[24] Hangloo, S., & Arora, B. (2022). Combating multimodal fake news on social media: methods, datasets, and future perspective. Multimedia systems, 28(6), 2391-2422.

[25] Thomson, T. J., Angus, D., Dootson, P., Hurcombe, E., & Smith, A. (2022). Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities. Journalism Practice, 16(5), 938-962.

[26] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science, 31(7), 770-780.

[27] Verdoliva, L. (2020). Media forensics and deepfakes: an overview. IEEE journal of selected topics in signal processing, 14(5), 910-932.

[28] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[29] Kaprykowsky, H., Liu, M., & Ndjiki-Nya, P. (2009, November). Restoration of digitized video sequences: an efficient drop-out detection and removal framework. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 85-88). IEEE.

[30] Tyagi, S., & Yadav, D. (2023). A detailed analysis of image and video forgery detection techniques. The Visual Computer, 39(3), 813-833.

[31] Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022). Deep learning for fake news detection: A comprehensive survey. AI open, 3, 133-155.

[32] Akhtar, Z. (2023). Deepfakes generation and detection: a short survey. Journal of Imaging, 9(1), 18.

[33] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. Journal of Business Research, 154, 113368.

[34] Romero Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. International Review of Law, Computers & Technology, 1-30.