

Research Proposal

Evaluating and Enhancing the Robustness of Graph-Based Deepfake Audio Detection Models Against VoIP Compression Attacks

Diyaben Pradipbhai Patel

Student ID: 2523935

MSc in Cybersecurity with Digital Forensic Analysis

Research Methods and Professional Development

University of Wales Trinity Saint David

August 14, 2026

Table of Contents

1. Introduction.....	3
1.1 Context and Background.....	3
1.2 Research Problem	4
1.3 Research Aim and Objectives	5
1.4 Significance of the Research.....	6
2. Literature Review (Systematic Literature Review)	7
2.1 Introduction to the Literature Review	7
2.2 Literature Search Strategy.....	7
2.3 Evolution of Audio Deepfake Detection.....	9
2.4 Graph Neural Network-Based Detection Models	10
2.5 Robustness of Audio Deepfake Detection under VoIP Compression	11
2.6 Research Gap and Positioning of the Proposed Study.....	13
2.7 Literature Review Summary	14
3. Research Design and Methodology	15
3.1 Research Design.....	15
3.2.1 Research Methods.....	16
3.2.2 Research Materials.....	17
4. Expected Outcome	18
5. Timeline	19
6. Conclusion	20
7. References.....	21

1. Introduction

1.1 Context and Background

Artificial intelligence (AI) has transformed many aspects of modern society by improving automation, communication, and digital content creation. One of the most significant developments in recent years has been the advancement of generative AI, particularly in speech synthesis and voice cloning. These technologies can generate highly realistic speech from a small amount of audio data and have legitimate applications in areas such as virtual assistants, accessibility, entertainment, and customer service. However, they have also enabled cybercriminals to conduct increasingly sophisticated social engineering attacks [11], [12].

The growing accessibility of AI voice cloning tools has contributed to a rise in audio deepfakes, where synthetic speech is created to imitate a genuine speaker. Modern deepfake audio can accurately reproduce a person's tone, pronunciation, and speaking style, making it increasingly difficult for both humans and automated systems to distinguish genuine speech from manipulated audio [1], [3]. As a result, audio deepfakes have become a significant cybersecurity concern, particularly in environments that rely on voice communication for authentication or decision-making [11].

One of the most common forms of audio-enabled cybercrime is voice phishing, or **vishing**, in which criminals use AI-generated voices to impersonate trusted individuals or organisations to obtain sensitive information or authorise fraudulent transactions [11], [12]. A notable example involved the global engineering company Arup, where criminals used AI-generated deepfake technology during a video conference to impersonate senior executives, resulting in fraudulent transactions totalling approximately **US\$25 million**. This incident demonstrates that AI-enabled voice impersonation has become a genuine threat to organisations rather than a theoretical risk.

To counter this growing challenge, researchers have developed a range of automated audio deepfake detection techniques. Early approaches relied on handcrafted acoustic features combined with traditional machine learning classifiers but often struggled to generalise to increasingly sophisticated synthetic speech [10]. More recently, deep learning has significantly improved detection performance through Convolutional Neural Networks (CNNs), which automatically learn discriminative features from spectrogram representations of speech [1]. Graph Neural Network (GNN)-based architectures, particularly the **Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST)** framework,

have further advanced the field by modelling both spectral and temporal relationships within speech signals, achieving state-of-the-art performance on benchmark datasets [2].

Despite these advances, most detection models have been developed and evaluated using high-quality, uncompressed audio collected under controlled conditions. In contrast, real-world voice communication typically occurs through Voice over Internet Protocol (VoIP) platforms such as Microsoft Teams, Zoom, and WhatsApp, where speech is compressed using codecs including Opus and G.711 before transmission. This compression alters the acoustic properties of the original signal, potentially reducing the effectiveness of detection models trained on clean audio [3], [6].

This difference between laboratory evaluation and practical deployment remains an important challenge for audio deepfake detection. Although existing models perform well on benchmark datasets, their robustness under VoIP compression is still insufficiently understood. Therefore, this research investigates how a codec-aware enhancement can improve the performance of the AASIST framework when detecting deepfake audio transmitted through common VoIP communication codecs.

1.2 Research Problem

Although recent audio deepfake detection models have achieved high accuracy on benchmark datasets, a significant gap remains between laboratory evaluation and real-world deployment. State-of-the-art models, including the **Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST)** framework, have demonstrated excellent performance using high-quality, uncompressed datasets such as ASVspoof [2], [3]. However, these controlled conditions do not accurately represent the environments in which voice-based cyberattacks typically occur.

In practice, most voice communication takes place through Voice over Internet Protocol (VoIP) platforms such as Microsoft Teams, Zoom, Skype, and WhatsApp. These platforms compress speech using codecs such as Opus and G.711 to reduce bandwidth requirements. While this compression has little impact on human listeners, it alters important acoustic characteristics that deepfake detection models rely on for classification [6]. Consequently, models trained on clean speech may experience reduced performance when analysing compressed audio transmitted through real-world communication channels.

Although the ASVspoof 2021 challenge encouraged the development of more robust anti-spoofing systems, most proposed models continued to be evaluated using benchmark datasets

rather than authentic VoIP environments [3]. In addition, recent robustness studies have shown that state-of-the-art detection systems remain vulnerable to compression and other real-world signal degradations, resulting in reduced detection accuracy [6]. These findings indicate that strong benchmark performance alone is insufficient to guarantee reliable deployment in practical cybersecurity applications.

Despite these challenges, limited research has focused on improving the robustness of graph neural network-based deepfake detection models specifically against VoIP compression. Existing studies have largely concentrated on developing new detection architectures or improving benchmark performance, while the effects of communication codecs remain comparatively underexplored [2], [5], [6]. As organisations increasingly rely on internet-based voice communication, addressing this limitation has become an important cybersecurity challenge.

Therefore, this research addresses the following research problem:

How can graph neural network-based deepfake audio detection models be enhanced to maintain reliable detection performance under realistic VoIP compression conditions?

To answer this question, the proposed research investigates the impact of common VoIP codecs and develops a codec-aware enhancement for the AASIST framework. The study aims to improve the robustness of deepfake audio detection systems while supporting their practical deployment in modern cybersecurity environments.

1.3 Research Aim and Objectives

The research problem identified in the previous section highlights the need for deepfake audio detection models that remain effective under realistic VoIP communication conditions. Accordingly, the following research aim and objectives have been developed to address this challenge.

Research Aim

The aim of this research is to develop and evaluate a codec-aware enhancement for a graph neural network-based deepfake audio detection model that improves detection robustness under Voice over Internet Protocol (VoIP) compression while maintaining high detection performance on benchmark datasets.

Research Objectives

To achieve the research aim, the following objectives have been identified:

1. To investigate the effect of common VoIP compression codecs, specifically Opus and G.711, on the performance of the AASIST deepfake audio detection model using the ASVspoof dataset.
2. To design and implement a codec-aware enhancement to the AASIST architecture that improves the model's ability to detect deepfake audio after VoIP compression.
3. To evaluate and compare the performance of the baseline and enhanced models using standard audio anti-spoofing evaluation metrics, including Equal Error Rate (EER) and minimum tandem Detection Cost Function (min-tDCF).
4. To analyse whether the proposed codec-aware enhancement improves the robustness of graph neural network-based deepfake audio detection under realistic VoIP communication conditions while maintaining competitive performance on uncompressed audio.

These objectives provide a clear framework for the proposed research and guide the development, implementation, and evaluation of the codec-aware enhancement presented in this study.

1.4 Significance of the Research

The increasing sophistication of AI-generated voice synthesis has created new cybersecurity challenges, particularly in environments that rely on voice communication for authentication, financial transactions, and organisational decision-making. As deepfake audio becomes more realistic and accessible, there is a growing need for detection systems that remain effective under real-world communication conditions. Improving the robustness of audio deepfake detection has therefore become an important research priority for both academia and industry.

This research addresses a practical limitation in current graph neural network-based deepfake detection systems. Although existing models achieve high accuracy on benchmark datasets, their performance under VoIP compression remains insufficiently explored [2], [3], [6]. By investigating the effects of common communication codecs and developing a codec-aware enhancement for the AASIST framework, this study aims to bridge the gap between benchmark performance and real-world deployment.

The findings may benefit organisations that depend on secure voice communication, including financial institutions, government agencies, healthcare providers, and large enterprises. In

addition to strengthening cybersecurity against AI-enabled voice fraud, this research may also support future studies on developing more robust and deployable audio anti-spoofing systems.

2. Literature Review (Systematic Literature Review)

2.1 Introduction to the Literature Review

The rapid advancement of artificial intelligence has accelerated the development of speech synthesis technologies, resulting in highly realistic deepfake audio capable of imitating human speech. While these technologies have valuable applications in areas such as healthcare, education, entertainment, and customer service, they have also introduced new cybersecurity risks through fraud, impersonation, and voice phishing attacks [11], [12]. Consequently, developing reliable audio deepfake detection systems has become an important research priority.

This literature review critically examines existing research on audio deepfake detection, with particular emphasis on graph neural network-based models and their performance under real-world communication conditions. The review evaluates the strengths, limitations, and contributions of current approaches to identify the research gap addressed by this study.

The review focuses primarily on peer-reviewed journal articles, conference papers, benchmark datasets, and survey studies published between **2020 and 2026**, while including a small number of earlier foundational studies. The literature was obtained from **IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar**, with priority given to recent publications because of the rapid evolution of deep learning and speech synthesis technologies.

The literature is organised into three themes: the evolution of audio deepfake detection techniques, advances in graph neural network-based detection models, and the robustness of these models under VoIP communication conditions. Together, these themes establish the foundation for identifying the research gap and positioning the proposed study.

2.2 Literature Search Strategy

A systematic literature search was conducted to identify studies related to audio deepfake detection, graph neural networks, and the impact of Voice over Internet Protocol (VoIP) compression on audio anti-spoofing systems. A structured search strategy was adopted to ensure that the reviewed literature was relevant and aligned with the research objectives.

The search was performed using **IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar**. These databases were selected because they provide

peer-reviewed journal articles, conference papers, and survey studies covering artificial intelligence, cybersecurity, speech processing, and machine learning. The search focused primarily on publications published between **2020 and 2026**, while including a small number of earlier foundational studies.

Search terms included **audio deepfake detection**, **deepfake speech**, **voice spoofing**, **audio anti-spoofing**, **graph neural networks**, **AASIST**, **RawNet2**, **VoIP compression**, **Opus**, **G.711**, **ASVspoof**, and **audio robustness**, using the Boolean operators **AND** and **OR** to refine the results.

Only peer-reviewed studies directly related to audio deepfake detection, speech anti-spoofing, or model robustness under realistic communication conditions were included. Duplicate records, non-English publications, studies focusing solely on image or video deepfakes, and papers lacking sufficient methodological detail were excluded.

The search identified **142** publications. After removing **28** duplicates, the remaining studies were screened based on their titles, abstracts, and full-text content, resulting in **18** studies being included in the final review. The literature selection process is presented in **Figure 2.1**.

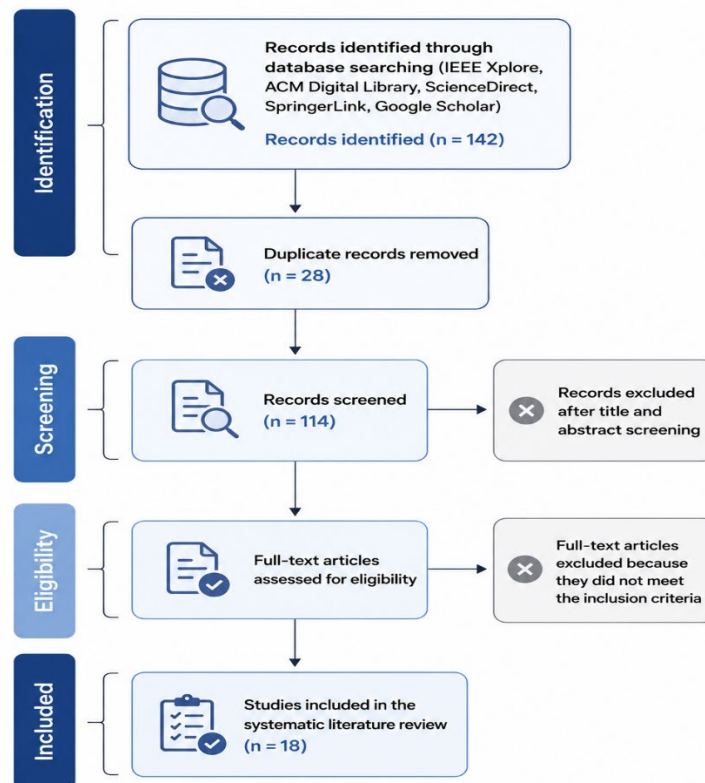


Figure 2.1. PRISMA Flow Diagram for Literature Selection.

These studies form the foundation for the thematic analysis presented in the following sections, which critically examine the evolution of detection techniques, recent graph neural network-based approaches, and the remaining challenges associated with VoIP compression.

2.3 Evolution of Audio Deepfake Detection

The rapid advancement of speech synthesis technologies has driven the evolution of audio deepfake detection methods. As synthetic speech has become increasingly realistic, researchers have moved from handcrafted acoustic features to deep learning models capable of learning complex speech representations automatically. This progression reflects the need for more robust detection techniques capable of handling increasingly sophisticated deepfake audio.

Early audio anti-spoofing systems relied on handcrafted acoustic features such as **Mel-frequency cepstral coefficients (MFCCs)**, **constant-Q cepstral coefficients (CQCCs)**, and **linear frequency cepstral coefficients (LFCCs)**, which were classified using traditional machine learning algorithms including **Gaussian Mixture Models (GMMs)** and **Support Vector Machines (SVMs)** [10]. Although these approaches established the foundation for audio spoofing detection, their reliance on manually engineered features limited their ability to generalise to modern speech synthesis techniques.

To overcome these limitations, researchers adopted deep learning models capable of learning feature representations directly from audio data. Convolutional Neural Networks (CNNs) significantly improved detection by automatically extracting discriminative patterns from spectrograms. Shafi and Mouti [1] demonstrated that CNNs using Mel-spectrogram analysis achieved higher detection accuracy than traditional feature-based approaches. However, CNNs primarily capture local spatial information and may not fully model long-range temporal and spectral relationships within speech.

Subsequent research introduced end-to-end architectures such as **RawNet2**, which learns directly from raw speech waveforms without handcrafted feature extraction [7]. Similarly, **Light Convolutional Neural Networks (LCNNs)** improved computational efficiency while maintaining competitive detection performance [8], [9]. Despite these advances, both approaches remain sensitive to speech conditions that differ from those encountered during training, highlighting continuing challenges in model robustness and generalisation.

Overall, the evolution of audio deepfake detection demonstrates consistent improvements in detection accuracy through increasingly sophisticated deep learning architectures. However, maintaining reliable performance under diverse real-world communication conditions remains an unresolved challenge, motivating the development of graph neural network-based approaches discussed in the following section.

2.4 Graph Neural Network-Based Detection Models

As audio deepfake generation techniques have become more sophisticated, researchers have recognised the limitations of conventional deep learning architectures in capturing the complex relationships within speech signals. Although Convolutional Neural Networks (CNNs) and end-to-end models such as RawNet2 have achieved strong performance, they primarily learn local feature representations and may overlook broader temporal and spectral relationships [1], [7]. This limitation has driven the development of graph neural network (GNN)-based approaches, which represent speech as interconnected structures rather than independent feature vectors.

Graph neural networks improve feature representation by modelling speech features as nodes within a graph, enabling the relationships between different acoustic components to be learned more effectively. This approach captures interactions between spectral and temporal characteristics, making GNNs particularly suitable for audio deepfake detection, where synthetic speech often contains subtle inconsistencies that are difficult to identify using conventional architectures [2].

A major advancement in this area is the **Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST)** framework proposed by Jung et al. [2]. Unlike earlier architectures that process spectral and temporal information separately, AASIST combines both within a unified graph attention network. By modelling dependencies across different regions of the speech signal, the framework achieved state-of-the-art performance on the ASVspoof benchmark and has become a widely adopted baseline for audio anti-spoofing research [2]. Its ability to learn complex feature relationships makes it well suited to detecting sophisticated AI-generated speech.

Building on this work, researchers have proposed alternative architectures to further improve feature representation. For example, Tan et al. [5] introduced a symmetrical dual-branch detection framework that combines complementary feature representations to improve discrimination between genuine and synthetic speech. Similarly, attention-enhanced LightCNN

models improve feature selection and classification performance through attention mechanisms [9]. These studies demonstrate continued efforts to improve both detection accuracy and model generalisation.

Despite these advances, most graph neural network-based models are still evaluated primarily on benchmark datasets collected under controlled conditions [2], [3]. Recent studies have shown that compression and other signal degradations can reduce detection performance, raising concerns about their reliability in real-world communication environments [6]. Therefore, although AASIST represents a significant advancement in audio anti-spoofing research, improving its robustness under VoIP compression remains an important research challenge.

Given its state-of-the-art performance, AASIST has been selected as the baseline model for this research. Rather than proposing a new detection architecture, this study seeks to enhance the robustness of the existing framework through a codec-aware approach designed for realistic VoIP communication environments.

2.5 Robustness of Audio Deepfake Detection under VoIP Compression

Although recent deep learning architectures have achieved high detection accuracy on benchmark datasets, maintaining performance under real-world communication conditions remains a major challenge. Most internet-based communication platforms, including Microsoft Teams, Zoom, WhatsApp, and Skype, compress speech using codecs such as **Opus** and **G.711** to reduce bandwidth requirements. While this compression has little effect on human perception, it alters acoustic characteristics that deep learning models use to distinguish genuine speech from synthetic audio [6].

The importance of evaluating detection systems under realistic conditions has been recognised by the **ASVspoof** initiative, which provides benchmark datasets for audio anti-spoofing research [3]. Although the ASVspoof 2021 challenge introduced more diverse spoofing attacks, most proposed models continued to be evaluated using controlled benchmark datasets rather than compressed VoIP audio [3]. Consequently, strong benchmark performance does not necessarily indicate reliable deployment in operational environments.

Li et al. [6] further demonstrated that state-of-the-art detection models experience reduced accuracy when speech is affected by compression, background noise, and transmission artefacts. These findings show that current models remain vulnerable when audio differs from

the conditions observed during training, highlighting the need for improved robustness under practical communication environments.

Despite growing awareness of this issue, limited research has specifically addressed the impact of VoIP codec compression on graph neural network-based detection models. Most studies focus on improving model architectures or benchmark performance rather than adapting existing systems for real-world communication channels [2], [5], [6]. This limitation represents an important obstacle to the practical deployment of audio deepfake detection systems.

Overall, the reviewed literature identifies a clear gap between benchmark evaluation and real-world application. Although graph neural network-based models such as **AASIST** achieve state-of-the-art performance, their robustness under VoIP compression remains insufficiently explored. Addressing this limitation forms the basis of the proposed research.

Table 2.1 Summary of Key Literature on Audio Deepfake Detection

Study	Method / Model	Key Contribution	Identified Limitation
Tak et al. [10]	Handcrafted acoustic features with traditional classifiers	Established the foundation for early audio anti-spoofing research.	Limited generalisation to modern AI-generated speech due to dependence on manually engineered features.
Shafi and Mouti [1]	CNN with Mel-spectrogram analysis	Demonstrated that CNNs can automatically learn discriminative spectral features for deepfake detection.	Performance primarily evaluated using controlled datasets; robustness under communication distortions was not investigated.
Tak et al. [7]	RawNet2	Proposed an end-to-end framework that learns directly from raw audio waveforms without handcrafted features.	Performance may decrease when evaluated under unseen real-world conditions.
Wang et al. [8]	LightCNN	Developed a lightweight architecture capable of efficient deepfake detection	Limited ability to model long-range temporal

		with lower computational complexity.	relationships within speech signals.
Ma et al. [9]	Improved LightCNN with Attention	Enhanced LightCNN by incorporating attention mechanisms to improve feature representation.	Still relies on benchmark datasets with limited evaluation under VoIP compression.
Jung et al. [2]	AASIST	Introduced an integrated spectro-temporal graph attention network that achieved state-of-the-art performance on ASVspoof benchmarks.	Performance under codec-compressed VoIP communication remains insufficiently explored.
Tan et al. [5]	Symmetrical Dual-Branch Detection Framework	Improved deepfake detection by combining complementary feature representations through a dual-branch architecture.	Does not specifically address robustness against communication codec compression.
Liu et al. [3]	ASVspoof 2021 Benchmark	Provided a comprehensive benchmark for evaluating modern audio anti-spoofing systems.	Benchmark evaluation does not fully replicate practical VoIP communication environments.
Li et al. [6]	Robustness Benchmark	Demonstrated that state-of-the-art detection models experience performance degradation under real-world audio corruptions.	Further research is required to improve robustness against codec compression used in VoIP systems.

2.6 Research Gap and Positioning of the Proposed Study

The reviewed literature demonstrates significant progress in audio deepfake detection, particularly through deep learning and graph neural network-based models. Architectures such as **AASIST** have achieved state-of-the-art performance on benchmark datasets by effectively

modelling spectral and temporal speech characteristics [2]. Similarly, recent studies have improved detection accuracy through advanced feature learning and attention mechanisms [5], [9]. However, most research has focused on benchmark performance rather than real-world deployment.

A key limitation identified throughout the literature is the limited evaluation of detection models under VoIP communication conditions. Although existing studies have shown that compression and other signal degradations reduce detection performance [6], few have specifically investigated graph neural network-based models that remain robust after codec compression. This limitation is increasingly important as VoIP platforms have become the primary means of voice communication in many organisations.

This research addresses the identified gap by investigating the impact of VoIP compression on the AASIST framework and developing a codec-aware enhancement to improve its robustness. Rather than proposing a new detection architecture, the study builds upon an established state-of-the-art model and evaluates its performance under realistic communication conditions. In doing so, it aims to support the practical deployment of reliable deepfake audio detection systems for enterprise cybersecurity applications.

2.7 Literature Review Summary

This literature review examined the evolution of audio deepfake detection, advances in graph neural network-based detection models, and the challenges of deploying these systems in real-world communication environments. The reviewed studies show that deep learning architectures, particularly AASIST, have significantly improved audio anti-spoofing performance [2]. However, they also demonstrate that existing models remain vulnerable to VoIP compression and other communication-related distortions [3], [6].

The review identifies a clear research gap in improving the robustness of graph neural network-based deepfake detection under realistic communication conditions. This gap provides the foundation for the proposed research, which seeks to enhance the AASIST framework through a codec-aware approach. The next chapter presents the research design and methodology adopted to achieve this aim.

3. Research Design and Methodology

3.1 Research Design

The primary objective of this research is to evaluate and improve the robustness of graph neural network-based deepfake audio detection under realistic Voice over Internet Protocol (VoIP) compression conditions. To achieve this objective, the study adopts a **Design Science Research (DSR)** approach combined with an **experimental research strategy**. This approach is appropriate because the research involves developing, implementing, and evaluating a codec-aware enhancement to an existing deepfake detection model rather than investigating human behaviour or perceptions.

Design Science Research focuses on creating and evaluating artefacts that solve identified problems. In this study, the proposed artefact is a codec-aware enhancement to the **Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST)** framework. The enhancement will be designed to improve the model's robustness when processing audio affected by VoIP compression while maintaining competitive performance on uncompressed speech.

The study follows a **positivist research philosophy**, where conclusions are drawn from objective experimental observations and quantitative performance measurements. Rather than relying on subjective interpretation, the proposed enhancement will be evaluated using standard audio anti-spoofing metrics and compared directly with the baseline AASIST model. This enables the effectiveness of the proposed approach to be measured using reproducible experimental procedures.

An experimental strategy has been selected because it allows the proposed enhancement to be evaluated under controlled conditions while systematically measuring the impact of different VoIP codecs on detection performance. The overall research workflow is presented in **Figure 3.1**, beginning with dataset preparation and VoIP compression simulation, followed by baseline model implementation, codec-aware enhancement, and performance evaluation.

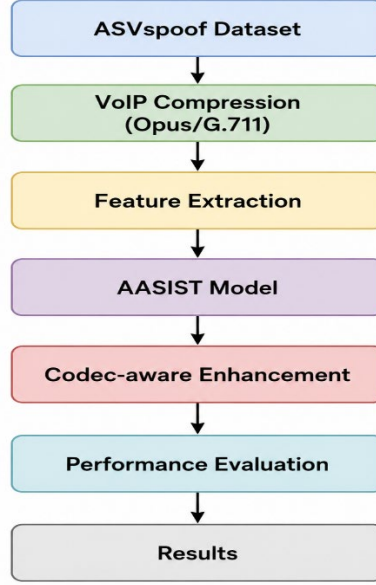


Figure 3.1. Overall Research Workflow.

The proposed experimental architecture is illustrated in **Figure 3.2**, which outlines the complete processing pipeline from the ASVspoof dataset through VoIP compression, feature extraction, graph neural network-based detection, codec-aware enhancement, and final performance evaluation.

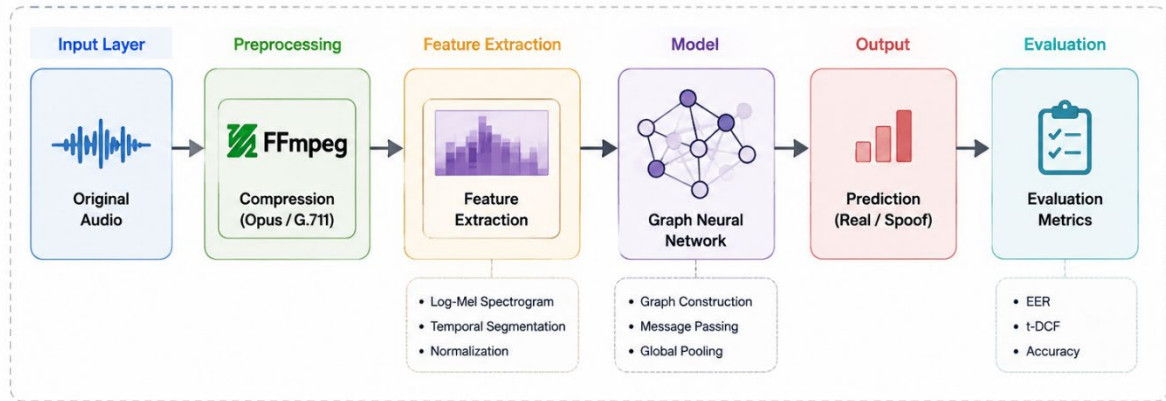


Figure 3.2. Proposed Experimental Architecture.

3.2.1 Research Methods

This study adopts a **quantitative research method** because its primary objective is to measure and compare the performance of audio deepfake detection models under different Voice over Internet Protocol (VoIP) compression conditions. Quantitative methods are appropriate for this

research as they enable objective evaluation using numerical performance metrics rather than subjective interpretation. The effectiveness of the proposed codec-aware enhancement will therefore be assessed through measurable improvements in deepfake detection performance.

An **experimental research strategy** will be employed to evaluate the proposed enhancement. The baseline **Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST)** model will first be implemented using the original ASVspooft dataset. The same dataset will then be processed using **FFmpeg** to simulate VoIP transmission by applying two widely used communication codecs, **Opus** and **G.711**. This approach enables a controlled comparison between model performance on clean audio and codec-compressed audio while ensuring that the only experimental variable is the applied compression.

The proposed codec-aware enhancement will then be incorporated into the baseline AASIST framework to improve its robustness against codec-related distortions. Performance will be compared with the original model to determine whether the enhancement improves detection reliability under realistic communication conditions.

Model performance will be evaluated using standard audio anti-spoofing metrics, including **Equal Error Rate (EER)** and the **minimum tandem Detection Cost Function (min-tDCF)**, which are widely used within the ASVspooft evaluation framework [3]. **Classification accuracy** will also be reported to provide an additional measure of overall detection performance. Lower EER and min-tDCF values will indicate improved robustness and more reliable detection of synthetic speech.

3.2.2 Research Materials

The research will utilise the ASVspooft 2021 dataset, which is one of the most widely recognised benchmark datasets for audio anti-spoofing research [3]. The dataset contains both genuine and AI-generated speech samples produced using a variety of speech synthesis and voice conversion techniques, making it suitable for evaluating the robustness of deepfake audio detection models. The dataset will be divided into training, validation, and evaluation sets according to the official ASVspooft protocol to ensure consistency with previous studies.

To simulate realistic communication environments, the original audio samples will be processed using **FFmpeg**, an open-source multimedia framework. FFmpeg will apply the **Opus** and **G.711** codecs to reproduce the compression commonly introduced by Voice over Internet Protocol (VoIP) platforms. Using controlled codec simulation allows the effect of

compression on detection performance to be measured while maintaining a consistent experimental environment.

The proposed model will be implemented using **Python**, with **PyTorch** serving as the primary deep learning framework. Supporting libraries, including **Librosa**, **NumPy**, and **Pandas**, will be used for audio processing, feature extraction, numerical computation, and data management. These tools are widely adopted in speech processing research and provide the flexibility required to implement, train, and evaluate the proposed codec-aware enhancement.

Model training and evaluation will be performed on a workstation equipped with an **NVIDIA GPU** to accelerate deep learning computations. Where local computing resources are insufficient, cloud-based platforms such as **Google Colab** may be used to support model training. The hardware selected is appropriate for training graph neural network models and processing large audio datasets within the available project timeframe.

The research uses publicly available benchmark datasets and does not involve human participants or the collection of personal data. Consequently, the study presents minimal ethical risk. All datasets will be used in accordance with their published licensing conditions, and the research will comply with the University's ethical guidelines, ensuring transparency, reproducibility, and responsible handling of research data.

4. Expected Outcome

As this study is a research proposal, the outcomes presented in this section are anticipated rather than confirmed. Based on the reviewed literature and the proposed methodology, it is expected that the baseline **AASIST** model will demonstrate reduced detection performance when evaluated on audio compressed using the **Opus** and **G.711** codecs. This reduction is expected because VoIP compression alters acoustic characteristics that graph neural network-based detection models use to distinguish genuine and synthetic speech [6].

It is anticipated that the proposed codec-aware enhancement will improve the robustness of the **AASIST** framework by reducing the impact of codec compression on detection performance. Compared with the baseline model, the enhanced system is expected to achieve lower **Equal Error Rate (EER)** and **minimum tandem Detection Cost Function (min-tDCF)** values while maintaining competitive accuracy on uncompressed benchmark data. Such findings would demonstrate that graph neural network-based deepfake detection models can be adapted for more reliable operation in realistic communication environments.

The research is expected to contribute to both academia and industry by providing insights into the effect of VoIP compression on audio deepfake detection and by demonstrating a practical approach for improving model robustness. These findings may support the development of more reliable anti-spoofing systems for enterprise cybersecurity, digital forensics, and secure voice communication.

Several limitations are acknowledged. The study will evaluate only the **Opus** and **G.711** codecs, meaning the findings may not be directly applicable to all communication platforms or compression methods. In addition, the experiments will be conducted in a controlled environment using benchmark datasets, which cannot fully replicate real-world network conditions such as packet loss, latency, or unstable internet connections. These limitations provide opportunities for future research.

5. Timeline

The proposed research is expected to be completed over a period of **six months**, with each phase aligned to the research objectives and project milestones. The project begins with the literature review and ethics approval, followed by dataset preparation and VoIP compression simulation. The subsequent stages involve implementing the baseline AASIST model, developing the proposed codec-aware enhancement, and evaluating model performance using standard audio anti-spoofing metrics. The final phase focuses on analysing the findings, completing the dissertation, and preparing the project for submission. The proposed project schedule is presented in **Figure 5.1**.

Activity	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
Literature Review						
Ethics Approval						
Dataset Preparation						
VoIP Compression Simulation						
Baseline AASIST Implementation						
Codec-Aware Model Enhancement						
Performance Evaluation						
Results Analysis						
Report Writing						
Final Submission						

Figure 5.1. Proposed Project Timeline.

6. Conclusion

The proposed research aims to improve the robustness of graph neural network-based audio deepfake detection under realistic Voice over Internet Protocol (VoIP) communication conditions. By investigating the effects of codec compression and developing a codec-aware enhancement for the AASIST framework, the study seeks to bridge the gap between benchmark performance and practical deployment. The proposed methodology provides a systematic approach for designing, implementing, and evaluating the enhancement using recognised datasets and evaluation metrics. It is anticipated that the findings will contribute to the development of more reliable audio anti-spoofing systems and support future research into secure AI-enabled voice communication within enterprise cybersecurity environments.

7. References

- [1] A. M. Shafi and S. Mouti, "Deepfake Audio Detection Using Convolutional Neural Networks and Mel-Spectrogram Analysis," in *2026 6th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Fez, Morocco, 2026, doi: 10.1109/IRASET68627.2026.11538779.
- [2] J.-W. Jung *et al.*, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371, doi: 10.1109/ICASSP43922.2022.9747272.
- [3] Y. Liu *et al.*, "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2722–2737, 2023, doi: 10.1109/TASLP.2023.3285283.
- [4] X. Wang *et al.*, "ASVspoof 5: Crowdsourced Audio Deepfake Dataset," *arXiv preprint*, arXiv:2408.08739, 2024.
- [5] X. Tan, Y. Wang, and Z. Zhang, "A Symmetrical Dual-Branch Detection Framework for Audio Deepfake Detection," *Symmetry*, vol. 17, no. 5, p. 641, 2025, doi: 10.3390/sym17050641.
- [6] H. Li, Y. Chen, and X. Wei, "Robustness of Audio Deepfake Detection Models Against Real-World Corruptions," *arXiv preprint*, arXiv:2503.17577, 2025.
- [7] H. Tak *et al.*, "End-to-End Anti-Spoofing with RawNet2," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373, doi: 10.1109/ICASSP39728.2021.9414234.
- [8] Z. Wang *et al.*, "Light Convolutional Neural Network with Feature Genuinization," *arXiv preprint*, arXiv:2009.09637, 2020.
- [9] X. Ma *et al.*, "Improved LightCNN with Attention Modules for ASV Spoofing Detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2021.
- [10] H. Tak *et al.*, "An End-to-End Anti-Spoofing System Based on Raw Waveform and Residual Networks," in *Proceedings of Interspeech*, 2018.

- [11] P. J. Wall, S. Furnell, G. Schaefer, and K. Mayes, "Vishing: Detecting Social Engineering in Spoken Communication—A First Survey & Urgent Roadmap to Address an Emerging Societal Challenge," *Computer Standards & Interfaces*, vol. 95, Art. no. 104015, 2025.
- [12] G. B. Gañac, M. P. A. Lucero, and E. A. De Guia, "Emerging Threat: The Use of AI Voice Cloning Software and Services to Deceive Victims Through Phone Conversations and Its Potential Effects on the Filipino Population," in *Proceedings of the 2024 International Conference on Computing, Communication and Artificial Intelligence*, ACM, 2024, doi: 10.1145/3702138.3702145.