



Automated Detection of Deepfake Audio in Real-Time VoIP Communication

D.D.C.M. Chandrasiri
(Reg. No.: MS18908848)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Mr. Kavinga Abeywardena

Approved for MSc. Research Project:



MSc in IT sp. CS Dr. Harinda Fernando, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign:

D.D.C.M. Chandrasiri

Date:

ABSTRACT

Automated Detection of Deepfake Audio in Real-Time VoIP Communication

Chamath Chandrasiri

MSc. in Information Technology

Supervisor: Mr. Kavinga Abeywardena

December 2025

With the increasing sophistication of AI-generated deepfake audio, real-time voice communication systems such as Voice over IP (VoIP) are at heightened risk of misuse through impersonation, fraud, and misinformation. Existing detection methods primarily rely on computationally expensive deep learning models trained on static data, which are impractical for live applications constrained by low latency and limited resources. This research addresses this gap by investigating the viability of a lightweight, highly efficient Random Forest (RF) classifier for real-time deepfake audio detection in VoIP environments.

The proposed system utilizes a focused methodology: raw audio is segmented into 2-second chunks and transformed into a comprehensive 800-dimension feature vector comprising Mel-Frequency Cepstral Coefficients (MFCCs), Chroma, Spectral Contrast, and Zero-Crossing Rate. Through an iterative training process using combined standard and 'in-the-wild' datasets to ensure generalization, the final RF model achieved an overall accuracy of 93.77% on an independent test set. Critically, the system demonstrated extremely low end-to-end processing latency of approximately 76 milliseconds (well below the <200ms target).

The findings prove that this computationally efficient, classical machine learning approach can achieve both high accuracy and speed. The final model successfully met the False Positive Rate objective (<5%) with a measured FPR of 2.85% on independent data, making it a viable and practical solution for enhancing the security and trustworthiness of real-time voice interactions against emerging deepfake threats.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my supervisor, Mr. Kavinga Abeywardena, for his invaluable guidance, continuous support, and insightful feedback throughout the course of this research. His expertise and encouragement have been instrumental in shaping the direction and quality of this work.

I also extend my sincere appreciation to SLIIT for providing the opportunity and resources necessary to pursue this research. The knowledge and experience gained during this journey have been truly enriching.

Finally, I am deeply thankful to my family, friends, and colleagues for their unwavering support and encouragement which have been a constant source of strength throughout this endeavor.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS.....	v
List of Figures	vii
List of Tables	viii
Chapter 1 Introduction	1
1.1 Background.....	1
1.2 Problem Statement.....	1
1.3 Research Questions.....	2
1.4 Research Objectives.....	3
1.5 Scope and Limitations.....	4
1.6 Thesis Structure	5
Chapter 2 Literature Review	6
2.1 Review of Existing Literature	6
2.1.1 The Evolving Threat of Synthetic Audio.....	6
2.1.2 A Review of Audio Synthesis Techniques.....	8
2.1.3 Feature Extraction for Deepfake Audio Detection.....	10
2.1.4 Classical Machine Learning Approaches to Detection	14
2.1.5 Deep Learning Approaches to Detection	16
2.1.6 Standard Datasets and Evaluation Metrics.....	19
2.1.7 Challenges of Real-Time Detection in VoIP Communication.....	23
2.2 Research Gap and Objectives	26
2.2.1 Identification of the Research Gap.....	26
2.2.2 Research Objectives.....	27
Chapter 3 Methodology	29
3.1 Research Design.....	29
3.2 Dataset Selection and Preparation.....	31
3.2.1 Bonafide Audio Source: Common Voice (CV) Corpus.....	31
3.2.2 Spoofed Audio Source: DEEP-VOICE Dataset.....	31
3.2.3 Data Preparation and Splitting	32
3.3 Feature Extraction	33
3.4 Detection Model Implementation	35
3.4.1 Model Selection and Justification	35
3.4.2 Model Implementation and Training	35
3.5 System Design and Evaluation	36

3.5.1 Proof-of-Concept (PoC) System Design.....	36
3.5.2 Evaluation Metrics and Strategy	37
Chapter 4 Results and Evaluation	39
4.1 Initial Model Performance (Model V1: MFCC-only Features)	39
4.1.1 Results (Experiment 1).....	39
4.1.2 Analysis (Experiment 1)	40
4.2 Iteration: Enhanced Feature Set (Model V2)	40
4.2.1 Results (Experiment 2 - Model V2 on Internal Test Set).....	40
4.2.2 Analysis (Experiment 2)	41
4.3 Evaluation of V2 Generalization and Motivation for V3.....	42
4.4 Iteration: Combined Training Data (Model V3)	43
4.4.1 Results (Model V3 on Internal Test Set).....	43
4.4.2 Analysis (Model V3 on Internal Test Set)	43
4.5 System Demonstration and Latency Measurement.....	45
4.6 Independent Evaluation of Final Model (V3)	46
4.6.1 Results (Model V3 on Independent RitW Test Set).....	46
4.6.2 Analysis (Model V3 on Independent RitW Test Set)	46
4.7 Summary of Results	47
Chapter 5 Discussion and Conclusion	48
5.1 Discussion of Key Findings	48
5.2 Limitations of the Study.....	50
5.3 Future Work	51
5.4 Conclusion	52
References.....	53

List of Figures

Figure 3.1: System Architecture and Methodology Pipeline for Real-Time Deepfake Detection.....	29
Figure 3.2: Conceptual User Interface (UI) Wireframe for the Real-Time Detection Proof of Concept....	36
Figure 4.1: Screenshot of the Proof of Concept (PoC) Demonstration Application (Model V3).....	44

List of Tables

Table 4.1: Confusion Matrix for Initial Model (V1 - MFCC-only).....	38
Table 4.2: Confusion Matrix for Enhanced Model (V2 - Internal Test Set).....	39
Table 4.3: Confusion Matrix for Final Model (V3 - Internal Test Set).....	42
Table 4.4: Confusion Matrix for Final Model (V3 - Independent RitW Test Set).....	45