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is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources).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

List of Figures

List of Tables