DEEP FAKE AUDIO DETECTION BASED ON DEEP LEARNING

ATHALLAH ARIQ WIJAYA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS LAPORAN

JUDUL: DEEP FAKE AUDIO DETECTION BASED ON DEEP LEARNING

SESI PENGAJIAN: 2022 / 2023

Saya: ATHALLAH ARIQ WIJAYA

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan unituk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. * Sila tandakan (✓)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan)
✓	TIDAK TERHAD	

ARIQ W

(TANDATANGAN PELAJAR)

Alamat tetap: Jalan Damai II, Ciganjur,

Jagakarsa, Jakarta Selatan, Indonesia

Tarikh: 25/9/2023

(TANDATANGAN PENYELIA)

Dr. Nur Fadzilah binti Othman

Nama Penyelia

Tarikh: 25/9/2023

CATATAN: * Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

DEEP FAKE AUDIO DETECTION BASED ON DEEP LEARNING

ATHALLAH ARIQ WIJAYA

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Security) with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I hereby declare that this project report entitled

DEEP FAKE AUDIO DETECTION BASED ON DEEP LEARNING

is written by me and is my own effort and that no part has been plagiarized

without citations.

STUDENT : _____ Date : 25/9/2023

(ATHALLAH ARIQ WIJAYA)

I hereby declare that I have read this project report and found

this project report is sufficient in term of the scope and quality for the award of

Bachelor of Computer Science (Computer Security) with Honours.

SUPERVISOR : _____ Date : 25/9/2023

(DR NUR FADZILAH BINTI OTHMAN)

DEDICATION

This research project is devoted to my parents' persistent love and support, who have been the foundation of my academic path. Their sacrifices and encouragement developed my character and contributed to my personal progress. This accomplishment is a testimonial to their unwavering dedication and the ideals they instilled in me.

I would like to express my heartfelt appreciation to my outstanding supervisor, Dr. Nur Fadzilah Binti Othman, for their tremendous mentorship and advice. Their knowledge, unwavering desire for learning, and unshakable belief in my abilities have consistently encouraged me to strive for excellence and attain new heights.

Finally, I commit my research project to myself. It represents my unwavering determination, resilience, and dedication to personal and intellectual growth. It is a daily reminder of my capacity to overcome obstacles and achieve my goals. I am extremely proud of the time and effort I have put in.

I want to thank my parents, my supervisor, and myself for your invaluable contributions to this research study. Your constant support, invaluable advice, and unshakable faith in my talents have been the driving force behind my achievements.

ACKNOWLEDGEMENTS

I am deeply grateful to Universiti Teknikal Malaysia Melaka (UTeM) for providing us, undergraduate students from the Faculty of Information and Communication Technology (FICTS), with an excellent opportunity to embark on our final year project journey.

In addition, I want to thank my supervisor, Dr. Nur Fadzilah Binti Othman, for their constant support, helpful guidance, and excellent mentorship throughout the course of this study. Their considerable knowledge, critical input, and persistent commitment to quality have all contributed significantly to the trajectory and calibre of this research endeavour. I am really grateful for their constant encouragement and unwavering faith in my ability.

In addition, I am grateful to my classmates and friends for their strong support, encouragement, and stimulating talks, which have substantially improved my understanding of the subject. Their companionship and shared experiences made this study journey even more rewarding and left indelible memories.

Finally, I want to thank my parents, Eka Wijaya Adi Endra and Diana Hidayati, for their constant love, encouragement, and understanding throughout my academic journey. Their unshakable faith in my skills, as well as the sacrifices they have made, have been the driving force behind my achievements.

ABSTRACT

With the advance of technology, especially in the field of AI, new and recent threats have emerged called deep fake audio. Instead of using a video to create a convincing, legit video, this threat only relies on audio. An example of this threat in action is using it to call for social engineering or to create fake audio propaganda by broadcasting it on the radio or other streaming services that only use audio as the bridge. Although this threat only relies on the audio, it can be hard to distinguish between the real and fake because the listener cannot see the source of the audio or speaker of the voice. Many previous studies have proposed a novel approach by using deep learning models, from simple ones like CNN to complex ones like ResNet and combinations of two or more models. In addition, with the many approaches proposed by many researchers, it is hard to choose which model can be used efficiently and easily. By using the dataset provided by ASV Spoof 2019, a dataset that has been used by many research papers to research deep fake audio, we investigated three deep learning models: CNN, ResNet, and ResNet+LSTM, and how to implement those models using the Python language and PyTorch framework. Out of the three chosen models, the ResNet model has achieved the best performance against the other two models by using macro-F1 as the measurement of performance. The CNN model has the lowest performance, and the ResNet+LSTM model performance is between the CNN model and the ResNet model. Even though the performance of the ResNet+LSTM model in terms of macro-F1 is lower than that of the ResNet model, the accuracy when the model predicts the data in the evaluation dataset is greater than that obtained by the ResNet model because the ResNet+LSTM model tends to choose the spoof data, which also means that the model needs more training to be flexible to choose the correct label of data.

ABSTRAK

Dengan kemajuan teknologi, terutamanya dalam bidang Kecerdasan Buatan (AI), ancaman baru dan terkini telah muncul yang dikenali sebagai audio palsu mendalam. Berbeza daripada menggunakan video untuk mencipta video yang meyakinkan dan sah, ancaman ini hanya bergantung kepada audio. Contoh ancaman ini dalam tindakan adalah dengan menggunakannya untuk tujuan kejuruteraan sosial atau mencipta propaganda audio palsu dengan menyiarkannya di radio atau perkhidmatan streaming lain yang hanya menggunakan audio sebagai penghubung. Walaupun ancaman ini hanya bergantung kepada audio, ia boleh sukar untuk membezakan antara yang sebenar dan palsu kerana pendengar tidak dapat melihat sumber audio atau pembicara suara. Banyak kajian terdahulu telah mencadangkan pendekatan baharu dengan menggunakan model pembelajaran mendalam, dari yang ringkas seperti CNN hingga yang kompleks seperti ResNet dan gabungan dua atau lebih model. Selain itu, dengan banyak pendekatan yang dicadangkan oleh ramai penyelidik, sukar untuk memilih model yang boleh digunakan secara cekap dan dengan mudah. Dengan menggunakan dataset yang disediakan oleh ASV Spoof 2019, dataset yang telah digunakan oleh banyak makalah penyelidikan untuk mengkaji audio palsu mendalam, kami mengkaji tiga model pembelajaran mendalam: CNN, ResNet, dan ResNet+LSTM, serta bagaimana melaksanakan model-model tersebut dengan menggunakan bahasa Python dan rangka kerja PyTorch. Daripada tiga model yang dipilih, model ResNet telah mencapai prestasi terbaik berbanding dua model lain dengan menggunakan nilai macro-F1 sebagai ukuran prestasi. Model CNN mempunyai prestasi terendah, dan prestasi model ResNet+LSTM berada di antara model CNN dan model ResNet. Walaupun prestasi model ResNet+LSTM dalam hal macro-F1 lebih rendah daripada model ResNet, ketepatan apabila model meramalkan data dalam dataset penilaian adalah lebih tinggi daripada yang diperoleh oleh model ResNet kerana model ResNet+LSTM cenderung memilih data palsu, yang juga bermakna bahawa model memerlukan latihan lebih banyak untuk menjadi fleksibel dalam memilih label data yang betul.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

LIST OF ABBREVIATIONS

FYP	**-**	**Final Year Project**
CNN	**-**	**Convolutional Neural Networks**
ResNet	**-**	**Residual Network**
LSTM	**-**	**Long Short-Term Memory**
ResNet+LSTM	**-**	**Residual Network + Long Short-Term Memory**
CPU	**-**	**Computational Processing Unit**
GPU	**-**	**Graphical Processing Unit**

LIST OF ATTACHMENTS

CHAPTER 1: INTRODUCTION

1.1 Introduction

With the advancement of technology, people were introduced to many novel technologies that can help people with everything from a small task like planning a trip to a big task like helping humans go to outer space. Of course, there are many more amazing technologies that can help humans or even be used as a tool to do negative things such as share fake information or misinformation.

The spread of false information, which is made possible by the creation of sophisticated algorithms and tools, is one important issue that has surfaced in recent years. This misinformation includes the development of bogus audio, image, and video content in addition to fake news. In order to produce a convincing video of the target person saying or doing things they never actually did, these AI-generated works of art, known as deepfakes (Camacho et al., 2021).

Deepfakes are merely one type of synthetic media that can be produced in text, image, video, and audio formats. Concerns about the abuse of AI technology have been highlighted because of how easily manipulated or fraudulent multimedia content may be produced (Bartusiak & Delp, 2021). In several events involving the dissemination of false information and the fabrication of fabricated content, deepfakes in particular have emerged as a major strategy (Lim et al., 2022b).

We have barely begun to explore the potential repercussions of synthetic media. Deepfakes and other types of fabricated information have yet to be fully understood for the genuine harm they can cause. At the individual, societal, and global levels, this ticking time bomb has the potential to cause devastation (Bartusiak

& Delp, 2021). In addition, a deep fake audio can be used to commit fraud for example, A-K Energy's CEO was instructed to transfer 220,000 Euros to a Hungarian supplier when he believed he was speaking with his German supervisor (Khanjani et al., 2023).

1.2 Problem Statement

With the rise of deepfake technology, criminals are using it to create fake audio that can be used to spread false information, such as conduct phishing scams through phone calls. The challenge lies in developing effective methods to detect and differentiate between genuine and manipulated audio recordings.

The main problem is that these fake audio recordings can sound incredibly realistic, making it difficult for people to determine whether they are listening to a genuine voice or a manipulated one. This poses a significant risk to individuals' privacy and security, as well as the trustworthiness of audio content.

The goal is to develop robust and reliable techniques that can accurately identify deepfake audio. By doing so, the goal can minimize people from falling victim to fake information, fraudulent phone calls, and other malicious activities that exploit manipulated audio. This will help ensure the authenticity and reliability of audio content, providing individuals with the necessary tools to make informed decisions and maintain their trust in the audio they encounter in their daily lives. In addition, many deep learning algorithms has been proposed by many researchers around the world, this study also need to find which model has the better accuracy.

1.3 Research Question

Based on the aforementioned issue on problem statement, a few research questions have been developed, as shown in Table 1.1.

Table 1.1: Table of Research Question

No	Research Question
1	What type of deep learning algorithm used to detect deepfake audio?
2	Which deep learning algorithm is effective to detect deepfake audio?
3	How is the performance and the accuracy of the model?

1.4 Research Objective

Based on the aforementioned issue on problem statement, a few research questions have been developed, as shown in Table 1.2.

Table 1.2 Table of Research Objective

No	Research Objective
1	To investigate the different types of deep learning models.
2	To compare deep learning algorithms in detecting deepfake audio.
3	To propose the deep learning algorithm that has better accuracy in detecting deepfake audio.

1.5 Research Summary Matrix

Table 1.3: Table of Research Summary Matrix

No	Research question	Research objective
1	What type of deep learning algorithm used to detect deepfake audio?	To investigate the different types of deep learning models.
2	Which deep learning algorithm is effective to detect deepfake audio?	To compare deep learning algorithms in detecting deepfake audio.
3	How is the performance and the accuracy of the model?	To propose the deep learning algorithm that has better accuracy in detecting deepfake audio.

1.6 Scope of Research

Three deep learning algorithm methods will be used in this research to detect deepfake audio. Although they are popular and good in terms of performance and accuracy, the fundamental reason uses three different deep learning model such as CNN, ResNet50, and resnet50 combine with LSTM is to know how to implement it

and to know which model is the best among the three. These three models are closely related to each other especially CNN and ResNet50.

Convolutional Neural Networks (CNNs) are highly popular and frequently applied deep learning algorithms. They have shown effectiveness in a variety of computer vision applications, such as face detection, facial recognition, object detection, and others. In the other hand, Residual Network (ResNet) is also popular and well known as one of the go-to-go solution for image related problem.

ResNet is similar to Convolution Neural Network (CNN) but with additional special step or layer that make it better than traditional Convolutional Neural Network. This layer or special is enable researcher to solve a problem when employing deep layer in CNN which is vanishing gradient. The name of ResNet50 comes from the name of the model itself and the number comes from how many layers that the model has. LSTM or long short-term memory is an enhanced version of Recurrent Neural Network (RNN) algorithm. LSTM itself has a capability to learn dependencies in the data and make prediction based on the dependencies in the data. LSTM itself usually utilize for stock prediction, NLP (Natural Language Processing), language translation and more.

Python is another important tool in this study because it is the core component to implement those three models. Python is and high-level programming language where human can understand by studying the syntax and how they read. Python is a programming language that based on object-oriented programming (OOP) language. Meaning python supposedly use by creating a class and an object of the class, one advantage of OOP is the code can be modular, which is fixing and troubleshooting the code is easier because it the code consist of small chuck of code that user can manage piece by piece.

Another tool used in this study is Pytorch. Pytorch is an open-source deep learning framework that can help researcher in making deep learning model. Pytorch is an OOP like approach to make the model. Thus, making it easy to maintain and experiment.

The last tool used in this study is Kaggle kernel. Kaggle kernel is a cloud service provided by Kaggle targeted for people who want to try and build a machine learning or deep learning model. They offer a ready to go environment that people