



Addis Ababa University

College of Natural Sciences

***Afaan Oromoo Audio Deepfake Detection Model Using Hybrid Deep
Learning Techniques***

Oromia Chala Bedada

A Thesis Submitted to the Department of Computer Science in
Partial Fulfillment for the Degree of Master of Science in
Computer Science

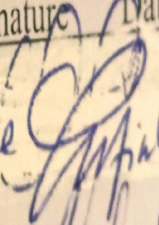
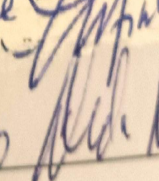
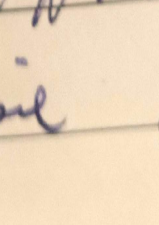
Addis Ababa, Ethiopia

May 2025

Addis Ababa University
College of Natural Sciences
Oromia Chala Bedada
Advisor: Minale Ashagrie (PhD)

This is to certify that the thesis prepared by *Oromia Chala*, titled: *Afaan Oromoo Audio Deepfake Detection Model Using Hybrid Deep Learning Techniques* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards concerning originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	Minale Ashagrie		7/14/2022
Examiner:	Dida Hilekso		07 Nov '25
Examiner:	Yaregal Assabie		26/08/2025

Abstract

Deepfake technology has quickly grown through advances in deep learning and generative adversarial networks, enabling the creation of highly realistic synthetic media. However, these technological breakthroughs have also raised serious alarms concerning the possible misappropriation of DF content for spreading misinformation, fraud, and voice impersonation. Despite the growing complexity of Deepfake generation, there exists a considerable disparity in developing effective and robust detection techniques for audio Deepfakes, particularly for languages such as Afaan Oromoo. As widely spoken as Afaan Oromoo is despite the richness of linguistic content, insufficient focused research and customized detection models exist that are addressed to its characteristic phonetic and prosodic features. The identified gap reveals that speakers and systems utilizing Afaan Oromoo are vulnerable to risks linked to synthetic audio, such as security vulnerabilities and the dissemination of misinformation. This thesis proposes a novel hybrid deep learning approach for audio Deepfake detection. A hybrid model that integrates Convolutional Neural Networks (CNNs) for spatial feature extraction with Long Short-Term Memory (LSTM) networks for capturing temporal dependencies is developed. A comprehensive dataset was created by collecting genuine audio from reliable source and generating synthetic audio using a real-time voice cloning (RTVC) system. Extensive experiments demonstrated that the proposed CNN-LSTM model successfully learned the subtle acoustic and prosodic characteristics of Afaan Oromoo speech, achieving a training accuracy of 97.40% and an overall test accuracy of 95% a figure that accurately reflects the model's robust generalization performance, as opposed to the previously noted 96% accuracy. The experimental results validated through additional metrics including precision, recall, and F1-score 0.95 for both real and fake. This research contributes a tailored, language-specific hybrid deep learning model for Deepfake detection, contribution a feasible solution to mitigate the risks associated with synthetic audio in underrepresented linguistic communities.

Keywords: Deepfake, Audio Deepfake Detection, Hybrid CNN-LSTM, Deep Learning, Afaan Oromoo, Real-Time Voice Cloning.

Dedication

I owe the greatest debt to my father, whose unconditional love, support and encouragement have enabled me to overcome a multitude of challenges.

Acknowledgments

Most of all, I would like to thank God Almighty for the numerous blessings, wisdom, and strength. He has granted me throughout my life. His blessing and wisdom were greatly essential in surmounting challenges and ensuring the success of this thesis.

I extend my genuine gratitude to my thesis advisor, Minale Ashagrie (PhD), for his persistent encouragement, constructive ideas and support during this thesis. His patience and expertise significantly influenced the direction and success of this thesis.

I would like to express heartfelt gratitude to Feda Negesse (PhD) for his outstanding mentorship and extensive experience in Afaan Oromoo linguistics. His insightful feedback, scholar guidance and rigorous instruction helped me and sharpen my thought, enhance my analysis and develop the quality of this thesis.

I would like to express heartfelt gratitude to Genenuis Galata (PhD) Department Head of Afaan Oromoo at Addis Ababa University, and the committed departmental teaching staff for their priceless assistance and significant contributions throughout of this study.

I would like to extend my sincere thanks to Addis Ababa University for providing me with comprehensive free tuition female scholarship that has enabled this academic endeavor. Their generous assistance has significantly reduced my financial loads and has simultaneously motivated me to perform well in my studies.

Finally, I extend my thanks to my parents' unconditional love, support, and sacrifices that have helped me become the person I am today, and for that I am incredibly thankful. I want to sincerely thank my husband for his patience, support, and faith in my talents. This accomplishment would not have been feasible without their love and unwavering support.

Table of Content

Lists of Figures	iii
Lists of Table	iii
Lists of Algorithms	iii
Lists of Acronyms and Abbreviations	iv
Chapter 1. Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Statement of the Problem	3
1.4 Objectives	4
1.5 Method	5
1.6 Scope and Limitations	5
1.7 Application of the Results	6
1.8 Organization of the Rest of the Thesis	6
Chapter 2. Literature Review	8
2.1 Introduction	8
2.2 Overview of Afaan Oromoo Language	12
2.3 Phonetics and Linguistics in Audio Deepfake Detection	15
2.5 Audio Deepfake Detection Approaches	15
2.4 Deepfake Detection in Low-Resource Languages	16
2.5 Datasets for Audio Deepfake Detection	17
2.6 Summary	18
Chapter 3. Related Work	20
3.1 Introduction	20
3.2 Audio Deepfake Detection Using Deep Learning	20
3.3 Summary	24
Chapter 4. Design of the Proposed System	25
4.1 Introduction	25

4.2 Dataset Creation	26
4.3 System Architecture	28
4.3.1 Components of the Architecture	29
Chapter 5. Experiment and Results	37
5.1 Introduction	37
5.2 Development Environment	37
5.3 Results and Discussion	38
5.4 Comparisons of other models over AOADS	45
5.4 Discussion	49
Chapter 6. Conclusion and Future Works	50
6.1 Conclusion	50
6.2 Contribution of This Work	51
6.3 Future Research Directions	51
References	52
Appendixes	61
Appendix A: the python code to build a model	61
Appendix B: RTVC Pipeline for Afaan Oromoo python code	64

Lists of Figures

Figure 2- 1 Deepfake Evolution [34]	10
Figure 2-2 Types of Deepfake	11
Figure 4- 1 Real Time Voice Cloning to Create Synthetic Dataset	26
Figure 4-2 Proposed Audio Deepfake Detection for Afaan Oromoo	29
Figure 5- 1 Sample of labeled data plot that shows how the data labeled as real and synthetic.	39
Figure 5-2 Graph of MFCC for sample dataset	40
Figure 5-3 Sample analysis of prosodic feature distributions	41
Figure 5-4 Prosody feature comparison plots of real and synthetic audio	42
Figure 5-5 Sample of compiling and training process of the model	43
Figure 5-6 Loss and accuracy curves during Training of the proposed model.	45
Figure 5-7 Sample of compiling and training process of the LSTM model	47
Figure 5-8 Loss and accuracy curves during Training of LSTM	47
Figure 5-9 Shows Classification Report on performance matrices of SVM model	48

Lists of Table

Table 2- 1 Afaan Oromoo Vowels [43]	13
Table 2-2 Afan Oromo consonant Alphabets [43]	13
Table 5- 1 Evaluation of the model over training and test set.	44
Table 5-2 Summary of the proposed performance matrices results	44
Table 5-3 Summary of Comparison different models over AOADS	49

Lists of Algorithms

Algorithm 4- 1 Audio Preprocessing Algorithm	30
Algorithm 4- 2 Dataset Splitting Algorithm	32
Algorithm 4- 3 Acoustic and Prosody Feature Extraction Algorithm	33

Lists of Acronyms and Abbreviations

ADFD	Audio Deepfake Detection
AOADS	Afaan Oromoo Audio Dataset
ADF	Audio Deepfake
AI	Artificial Intelligence
ANN	Artificial Neural Network
AOADFD	Afaan Oromoo Audio Deepfake Detection
CNN	Convolutional Neural Network
CNN-LSTM	Convolutional Neural Network - Long Short-Term Memory
DF	Deepfake
DL	Deep Learning
DNN	Deep Neural Network
HiFi-GAN	High Fidelity Generative Adversarial Network.
GAN	Generative Adversarial Network
GPU	Graphics processing unit
LSTM	Long Short-Term Memory
MFCC	Mel-frequency Cepstral Coefficients
MLAAD	Multi-Language Audio Anti-Spoofing Dataset
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
RTVC	Real-Time Voice-Cloning
SVM	Support Vector Machine
TSS	Text To Speech

Chapter 1. Introduction

1.1 Background

Deepfake (DF) is a content or material that is synthetically generated or manipulated using artificial intelligence (AI) methods to be passed off as real and can include audio, video, image, and text synthesis [1]. DF technology has taken on a new level of complexity, which allows the synthesized insertion of audio and visual images into social media platforms. DF makes the differentiation of real and fake information more challenging than ever before, when Deep Learning (DL), a subfield of AI, plays a key role in generating DFs [2]. Generative adversarial networks (GANs) are commonly used for this purpose, which consist of two main components: a generator and a discriminator [3]. Senga *et al.* [4] the generator attempts to create realistic multimedia content, while the discriminator evaluates the authenticity of those generated contents. Through an iterative training process, both components improve their abilities, resulting in more convincing DFs over time.

Spreading misinformation, doubting individuals, or creating fraudulent content are the results of DF [5]. To address these issues, researchers and technology companies have been developing detection methods to identify content created using DFs. These techniques involve analyzing visual artifacts, inconsistencies, or irregularities that may be indicative of manipulation. It is important to note that while DFs can be entertaining or used for harmless purposes like creating false content, they also have the potential for misuse and can be unethical actions, to mitigate the negative impact of DF technology responsible usage and awareness is important [4, 6]. The emergence of GPU (Graphics processing unit) and the availability of large datasets were key enablers of DL, and they were greatly enhanced by the development of open-source, flexible software platforms with automatic differentiation [6]. DL is a class of state-of-the-art ML techniques that has sparked incredible global interest in the last few years [7]. Due to its, ML comes with innovations to help humans with a high potential of making things easy in everyday activities within technology when we use it for good purposes [8]. One of the results of ML is DF which has led to worldwide public cyber security concerns regarding the side effects of using DF. Audio DF (ADF), recently called audio manipulation, are becoming widely accessible using mobile devices or personal PCs [9].

ADFs, technically known as logical-access voice spoofing techniques, have become an increased threat to voice interfaces due to recent breakthroughs in speech synthesis and voice conversion technology. Modern text-to-speech synthesis (TTS) is capable of realistic fakes of human voices, also known as audio DFs or spoofs [10]. Since there are many ethical applications of this technology, there is also a serious risk of malicious use, although TTS technology makes it possible to clone the voices of politicians, this presents several concerns, including spreading misinformation. Moreover, while TTS technology offers numerous ethical and beneficial applications, its ability to generate highly realistic DF audio presents potential risks. The wrong way of this technology such as cloning politicians' voices for misinformation underscores the need for robust detection mechanisms and ethical safeguards. This concern further reinforces the importance of research in DF audio detection, mainly for languages with limited resources [11].

Afaan Oromoo is one of the natural languages classified as the Cushitic languages spoken in Ethiopia, Somalia, Tanzania, and Kenya. Among more than 83 languages in Ethiopia, Afaan Oromoo is one of the widely spoken languages in the country [13, 14]. Afaan Oromoo is being used in different communication media including social media to transfer messages worldwide, express opinions or feelings regarding socioeconomic and political aspects [12]. However, the consequence of ADF for any citizen cannot be imagined if it is used for an illegal purpose. Thus, this research work aims to propose the Afaan Oromoo ADFD model using hybrid DL techniques.

1.2 Motivation

DF media manipulation altering audio and video to create false representations poses a significant threat to the integrity of public communications and trust. For example, how can we determine whether a recording of Ethiopian Prime Minister Abiy Ahmed that was made three years ago and received media attention is authentic or a fake? What if Prime Minister Abiy Ahmed spoke challenging notes that were never true in a DF audio recording that was distributed, when this kind of information aims is to misinform the public by manipulating audio and pictures to produce false representations of individuals. This might challenge self-confidence in official communications, reason public conflict, and harm the prime minister's status. Mainly for languages like Afaan Oromoo which is under resourced, the possibility of such misuse highlights the pressing need for reliable technology-based audio DF detection tools.

Similarly, DF audio could be exploited to copy trusted voices in financial scams or phishing attacks, targeting unaware individuals. According to languages in Africa, it is underrepresented in both audio DF detection and broader natural language processing (NLP) efforts. This under-representation worsens the vulnerability of its speakers to the potential misuse of DF technology. So, in the world where digital manipulation is becoming increasingly sophisticated, it is imperative to develop technologies capable of safeguarding the authenticity of audio content in Afaan Oromoo.

1.3 Statement of the Problem

ADF represents an advanced technological development while AI tools were used to replicate human voices. At the start of design, it was used for positive applications such as creating audiobooks and aiding individuals with disabilities. However, ADFs have also been misused to compromise public safety [13]. Their malicious applications include spreading misinformation, committing fraud, and executing social engineering attacks. The challenge lies in identifying genuine audio from unreal content during digital forensic investigations due to the refined nature of DF generation [14].

Despite the important progress in ADF detection, current research has primarily focused on widely spoken languages like English. Lataifeh *et al.* [15] highlights the power of DL models, particularly CNNs and RNNs, over traditional machine learning algorithms like SVM (Support vector machine) and K-Nearest Neighbor(k-NN) in Arabic audio processing. However, the study conducted is specific to the Arabic language which may not be directly applicable to the Afaan Oromoo language, which has different phonetic compositions. For Afaan Oromoo speech processing, it is therefore essential to look at the suitability and optimization of DL models to guarantee precise and effective language representation. Another research by Yi *et al.* [16] also shows the effectiveness of DL in detecting spoofed audio in Mandarin Chinese. The language-dependent data and support on advanced TTS and voice conversion capabilities may be non-generalizable to low-resource languages. Spectrogram-based models may be insensitive to language-dependent acoustic properties, which will impact detection performance. Dialectal variation, speaker variability, and noisy environments are also not considered in those works and restricted those proposed solutions robustness. High computational costs also prevent real-world deployment in under-resourced settings. Overcoming these challenges is essential to developing deployable DF detection in a variety of linguistic and computational environments.

Menglu *et al.* [17] although used the MLAAD dataset to assess the useful multilingual audio anti-spoofing dataset, the appropriateness of the proposed model for low-resource languages like Afaan Oromoo is questionable. The dataset favors common languages and lacks the unique phonetic or prosodic features and acoustic features of underrepresented languages. Even though DL models have high accuracy rates, their dependance on large and varied datasets limits their use in languages with limited data. The work also doesn't challenge things like dialectal variation, speaker variation, and noisy conditions, which are essential for field positioning. It is imperative to better such limitations to create robust and diverse DF detection schemes.

Afaan Oromoo, spoken by millions in Ethiopia and parts of Kenya, remains significantly underrepresented in ADFD research by [18]. To the best of the researchers' knowledge, no existing technological frameworks or tools are tailored for detecting audio DFs in Afaan Oromoo. This gap underscores the critical need for developing hybrid robust, language-specific detection models that accommodate the unique phonetic and linguistic characteristics of Afaan Oromoo. Addressing this issue will contribute to more inclusive and comprehensive solutions in the global fight against audio-based misinformation and digital fraud.

1.4 Objectives

General Objective

The main objective of this study is to construct an Afaan Oromoo audio DF detection model using a hybrid DL technique.

Specific Objectives

The following list of specific objectives is provided to achieve the study's overall goal:

- To review the literature to identify methods and techniques used for detecting audio DF.
- To collect and build novel audio datasets for Afaan Oromoo language with diverse speakers, environments, and noise.
- To develop a prototype of the model.
- To evaluate the prototype using accuracy, precision, recall, and F1-score metrics.

1.5 Method

This Section describes the methods that will be used for conducting this research work.

Literature Review

The literature review puts the current research in perspective and shows how it relates to what has been done before. We will conduct a review of different resources like conference proceedings, journal articles, books, etc. to understand the gaps and find solutions for the identified problem.

Data Collection and Preprocessing

In the data collection phase of the research methodology, initial data will be collected from social media YouTube and the synthetic one will be created using real-time voice cloning (RTVC) technique. Audio preprocessing like noise reduction, resampling, segmenting, and converting audio data are steps for the audio data processing function we will do.

Development Tools

Python will be used to implement the model, and DL tools like TensorFlow and Keras different python libraries will be used to build and train the DL algorithm. Libraries like Pandas and NumPy will help with data manipulation and preprocessing, and Librosa will be used for audio processing and feature extraction with Praat via python library called Parselmouth. Jupyter Notebook will be used to set up an experimental environment for interactive development.

Evaluation Metrics

The proposed DF audio detection model performance will be assessed using accuracy, precision, recall, F1-score.

1.6 Scope and Limitations

The scope of this study is to design a model for ADFD for the Afaan Oromoo language using a hybrid DL technique. In addition to designing the model, we will also create a novel Afaan Oromoo audio dataset, which contains both real and synthetic audio data.

However, this research work has several limitations. One of the limitations may not be possible to cover all possible variations of Afaan Oromoo speech. Another limitation is

the lack of standardization in the use of Afaan Oromoo which affects the performance of DL model, for example, the use of dialects and different speaking styles could introduce variability that may affect the accuracy of the detection system.

1.7 Application of the Results

The application of ADFD is becoming increasingly important in various industries due to the rise of synthetic audio content generated using DL techniques. Some of the specific applications of the result of the study are:

- News and online social media: Afaan Oromoo is widely spoken in Ethiopia and is a major language in the country's media. ADF can manipulate news reports, leading to potential social and political consequences. The proposed model of Afaan Oromoo ADFD can be used to identify and prevent such manipulations ensuring the integrity of the news media.
- Audio authentication: Afaan Oromoo is also used in legal and business settings, where audio recordings may be used as evidence. So, the proposed Afaan Oromoo ADFD model can be used to authenticate audio recordings in Afaan Oromoo and detect any DF audio manipulations that may have been made.
- Speech recognition: Afaan Oromoo is a natural language that can be used in speech recognition systems. The proposed Afaan Oromoo ADFD model can be used to ensure the integrity of the speech recognition systems used in Afaan Oromoo by detecting and removing any DF audio inputs.
- Financial fraud: Afaan Oromoo is also used in different financial sectors. ADFs can be used to manipulate financial transactions, such as spoofing voice biometrics. The proposed Afaan Oromoo ADFD model can be used to detect and prevent such fraud in those financial sectors which are providing services by Afaan Oromoo.

Overall, ADFD for Afaan Oromoo has the potential to ensure the integrity of audio content in various applications, including news media, legal and business settings, and online social media platforms.

1.8 Organization of the Rest of the Thesis

The rest of this thesis is organized as follows: Chapter 2 explores the theoretical foundations essential for understanding ADFD, including DF technologies and the

evolution of DF, neural networks, advanced DL architecture and the unique linguistic features of the Afaan Oromoo language that influence audio processing. Chapter 3 provides a comprehensive review of existing related works, focusing on prior studies and approaches in ADFD while evaluating their applicability and significance to the challenges and opportunities within the context of Afaan Oromoo. Chapter 4 outlines the methodology adopted in this research, detailing the steps involved, such as data acquisition, preprocessing techniques tailored to Afaan Oromoo audio, and the design of a hybrid CNN-LSTM model for effective DF detection. Chapter 5 presents the experimental setup and results, offering a detailed analysis of the experiments conducted, the performance metrics used to evaluate the model, and insights gained from the outcomes. Finally, Chapter 6 concludes the thesis by summarizing the findings, highlighting the contributions of the study, discussing the limitations encountered, and providing a road map for future research directions in the ADFD Afaan Oromoo language

Chapter 2. Literature Review

2.1 Introduction

DF technology is an advanced use of AI in which generative models mostly GANs are used to produce realistic fake media in the form of video, image, and audio. Initially designed for artistic and entertainment use, but the technology has introduced severe ethical and security issues because of the potential misuse in identity deception, disinformation diffusion, and political manipulation [19] [20]. The foundation of DF technology is ML algorithms that have been trained on large sets of genuine media. GANs, the algorithm's pulse, collaborate with two neural networks: a generator to create synthetic content and a discriminator to decide whether its output is real or not. These networks, iterating continuously to hone themselves, produce results that become increasingly difficult to distinguish from genuine media [8, 21].

Initially, DF captured an abundance of headlines in the entertainment and social media field, in which manipulated videos inserted individuals into fabricated situations. A well-known example involves superimposing the face of one individual over another face in videos to give the false impression that they were present at gatherings where they were not [21]. Over time, as technology continued to improve, concerns about technology changed from mere entertainment but to the facilitation of disseminating misinformation, to facilitate fraud, to compromise personal and political security. Along with video DFs, AI-powered ADFs have also become accessible, enabling realistic imitation of a person's voice. With such capabilities come new dangers in social engineering attack landscape, voice deception, and identity spoofing of prominent individuals without their consent. Increased sophistication among audio DFs has made the detection of manipulated voice clips challenging as well, exacerbating security concerns [10, 23].

Due to its learning capabilities DL from data, where it is originated from artificial neural network (ANN), has become a hot topic in the context of computing, and is widely applied in various application areas like healthcare, visual recognition, text analytics, cybersecurity, and many more issues related to DF. Audio DFs refer to speeches that have been generated or manipulated using DL, which can be misused for impersonation. Other terms, such as synthesized speeches or synthetic speeches, also exist. However, synthesized speeches sometimes include all possible methods of synthesizing speeches, including non-deep learning-based methods. However, building an appropriate DL model

is a challenging task, due to the dynamic nature and variations in real-world problems and data [22]. Current voice cloning systems are smart enough to learn speech characteristics from a few samples and produce perceptually unrecognizable speech. These systems pose new protection and privacy risks to voice-driven interfaces. Synthetic audio has been used for malicious purposes, and it is difficult to classify what is real and fake during a digital forensic investigation [23]. DF audio classification and evaluate the current methods of DF audio detection for forensic investigation. Audio file features were extracted and visually presented using MFCC, Mel-spectrum, Chromograms, and spectrogram representations to further study the differences [24].

Deepfake Evolution

The history of DF technology has been remarkable in the past two decades, driven mainly by advancements in AI, DL and GANs. Initially designed for entertainment and research purposes, DFs now have practical applications in fields like cybersecurity, media creation, and synthetic media generation, thus raising ethical issues concerning misinformation and identity forgery [25]. DF technology, however, got an incredible push in 2014 with the development of GANs enabling realistic image and video generation through training AI models to produce incredibly likely DF content [26]. DF tools widened their scope considerably. The Face2Face project demonstrated real-time facial reenactment, allowing an actor's facial expressions to be transferred onto a video of another individual in real-time [27]. At the same time, DeepMind's WaveNet enhanced AI-based TTS synthesis, allowing AI-generated voices to sound more natural and human-like [28]. The development of voice and video DF technology gained enormous momentum when the launch of DeepVoice1, a DL based voice model, enabled AI to produce human-like vocalizations with natural prosody [29].

Masood *et al.* [30] as shown in Figure 2.1 provides a comprehensive review and detailed analysis of existing tools and ML based approaches for DF generation, and the methodologies used to detect such manipulations in both audio and video from 2018, the reality of DFs has significantly increased. Face Forensics dataset arrived as a tool to detect DFs, an indication of increasing fears over misinformation brought about by AI. From 2019 to 2021, DF technology grew more realistic. Text-Based Head Movement Editing allowed AI to generate new facial expressions from static images, enhancing DF video quality Neural Talking Heads continued to enhance AI-generated lip-syncing,

making synthetic avatars even more realistic [31]. FSGAN, a face-swapping GAN model, allowed for real-time face replacement, with Mark Zuckerberg DF causing worldwide concern over the threat of AI-generated misinformation [32]. As shown in Figure 2.1 from 2020 onward, real-time generation of DFs was a leading area of focus. By 2021, as shown by Figure 2.1 AI-powered DF methods such as Speech Portraits, Pose-Guided AI and LipSync3D enabled synthetic videos to become nearly identical to original videos [33].

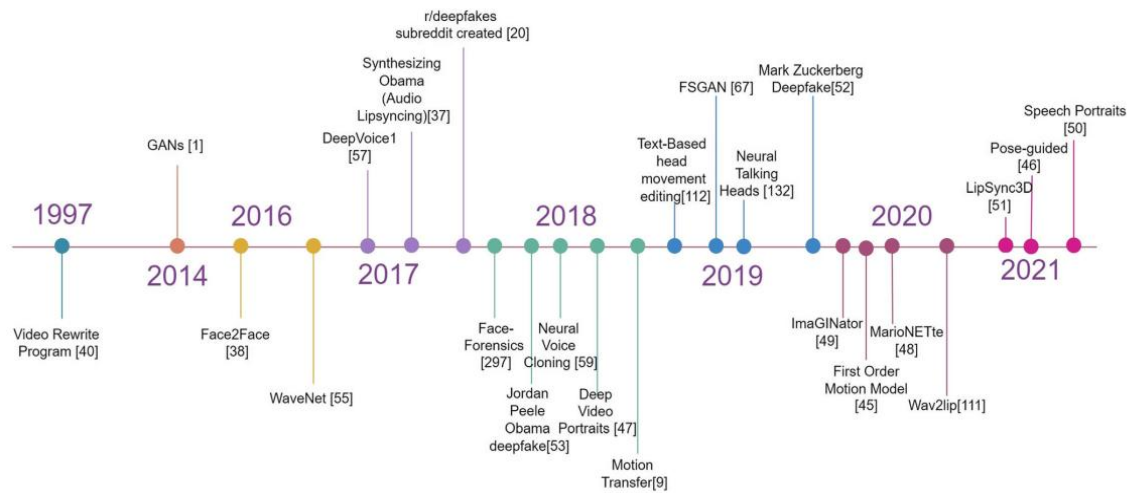


Figure 2-1 Deepfake Evolution [34]

Khan *et al.* [35], highlights how DF detection systems are developing and how difficult it may be to distinguish between real and fake content in financial settings, DF research is now aimed at the generation of hyper-realistic media, real-time DF processing, adversarial AI detection, and ethical control. DF technology is everywhere in media, entertainment, and virtual assistance but the threats of AI-based misinformation identity theft, and cyber-attacks, are always the driving forces behind DF detection and control.

Types of Deepfake

DF technology can be categorized based on the type of content it manipulates or synthesizes, including video, audio, image, and text DFs [36]. These types represent the different domains where generative models are applied to create synthetic media. As shown by Figure 2.2 DF can be classified into video, audio, image text and real time DF.

Video Deepfake: Kaur *et al.* [37] emphasizes the dominance of DL based methods in detecting DFs despite their computational efficiency and generalization limitations. Video DFs involve deploying existing video content or creating entirely synthetic video to mimic real individuals or events. With the help of techniques such as face swapping or

reform, these DFs can alter facial expressions, lip movements, or even the entire face of a subject in a video. GANs and motion transfer techniques are commonly employed in creating such DFs.

Audio Deepfake: ADFs utilize AI models, including neural TTS systems and GANs, to replicate human voices. These models are trained on real voice recordings to produce synthetic speech that imitates specific individuals in tone, pitch, and speaking style. ADFs are often used for voice cloning, virtual assistants and entertainment purposes but are also manipulated for fraud and misinformation [35,36].

Image Deepfake: Image DFs involve the generation or alteration of still images to produce realistic representations of people, objects, or scenes. Techniques such as StyleGAN and face morphing are widely used for applications like creating synthetic faces, enhancing images, or even producing fake identity photos for fraudulent purposes.

Text Deepfake: Text DFs involve the use of NLP models, such as GPT-like architectures,

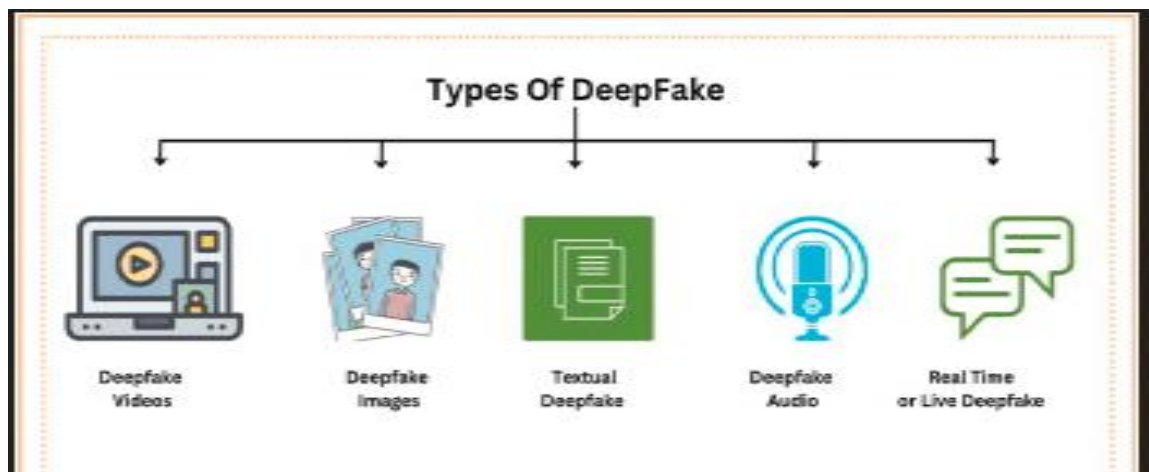


Figure 2-2 Types of Deepfake
[51]

to generate synthetic text that appears authentic. These models can replicate writing styles, generate fake news articles, or even impersonate individuals in written communication. Text DFs pose significant risks in disinformation and social engineering [38].

Real-Time or Live Deepfake: describes the use of DL methods to dynamically alter or create audio visual information. These systems process a live video or audio feed in real-time, allowing for applications in entertainment, augmented reality, and even potentially dishonest situations. For instance, they can change a person's visage or facial expressions

while they speak. Maintaining delays visual or audio reliability while keeping processing delay low enough to make the output seem instantaneous is the primary difficulty [39].

2.2 Overview of Afaan Oromoo Language

The source of Afaan Oromoo can be traced back to the early Cushitic languages that emerged in the Horn of Africa thousands of years ago [18]. The Oromoo people, supposed to have migrated to their current territories from the northeastern regions of Africa, have historically been pastoralist and farmers, which contributed to their linguistic and cultural development. Afaan Oromoo is the most widely spoken Cushitic language in Ethiopia and is also spoken in parts of Kenya, Somalia, and Tanzania [14, 52]. Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written.

Language Characteristics of Afaan Oromoo

The writing system of the language is straightforward, which is designed based on the Latin script. Unlike English or other Latin-based languages there are no skipped or unpronounced sounds/alphabets in the language. Besides, in a word where the vowels are doubled the sounds are stretched or elongated. Like in English, Afaan Oromo has vowels and consonants. Afaan Oromo vowels are represented by the five basic letters such as a, e, i, o, u. Besides, it has the typical Eastern Cushitic set of five short and five long vowels by doubling the five vowel letters: aa, ee, ii, oo, uu". Consonants, on the other hand, do not differ greatly from English, but there are few special combinations such as "ch" and "sh" (same sound as English), "dh" in Afaan Oromo is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Another Afaan Oromo consonant is "ph" made when with a smack of the lips toward the outside "ny" closely resembles the English sound of "gn". In general, Afaan Oromo has 33 letters called "Qubee". All letters in English language are also in Afaan Oromo [40]. It is widely used in commercial, religious, social, political and educational sectors, and in the mass media though there is no officially assigned dialect to be the representative standard form [41].

Afaan Oromo Phonetics

In the TTS and STT systems, a language's grapheme to phoneme correlation shows that a target word is converted from its written (grapheme) to its sound (phoneme) form. The associated

characteristics of the language, such as voicing, tongue position, tongue height, place of articulation, and method of articulation, are utilized to define the phone set [42].

Vowel System: Afaan Oromoo has five vowel phonemes: /a/, /e/, /i/, /o/, and /u/, which show vowel agreement which mean that certain vowels within a word correspond to be either front or back vowels. Vowel length is also contrastive in the language, meaning that short and long vowels can distinguish meaning in words. For example, "*bira*" (by the side) and "*biiraa*" (beer) have different meanings due to vowel length [12]. Table 2.1 shows Afaan Oromoo vowel system with their type, short and long.

Table 2- 1 Afaan Oromoo Vowels [43]

Short Vowels	Long Vowels
A	Aa
E	Ee
I	Ii
O	Oo
U	Uu

Consonant System: Afaan Oromoo has a relatively rich set of consonants, including voiced and voiceless stops, soft, high, and liquids. A notable feature of the language is the presence of implosive consonants, which are produced with an inward airflow [43]. Table 2.2 shows phonetic representation of Afaan Oromoo consonants and its International Phonetic Alphabets (IPA) equivalences

Table 2- 2 Afan Oromo consonant Alphabets [43]

Type	Phonemes (IPA)	Orthography
Voiced stops	b, d, g	b, d, g
Voiceless stops	t, k, ʔ	t, k, ' (glottal stop)
Ejective stops	p', t', k'	ph, x, q
Implosive	ɗ (retroflex implosive)	Dh

Fricatives	f, s, ʃ, h	f, s, sh, h
Affricates	dʒ (voiced), tʃ, tʃ̥ (voiceless and ejective)	j, ch, c
Nasals	m, n, ɲ	m, n, ny
Lateral	L	L
Flap	R	R
Semivowels	w, j	w, y

Acoustic Features

Duration: Vowel and consonant length are phonemically significant in Afaan Oromoo. Long vowels and consonants are noticeably longer in duration than their short counterparts. Acoustic studies have shown that vowel length can be up to twice as long as short vowels and consonantal germination can similarly affect speech duration [44].

Prosodic Features

Prosody is an umbrella term used to cover a variety of interconnected and interacting phenomena, namely stress, rhythm, phrasing, and intonation. The phonetic expression of prosody relies on several parameters, including duration, amplitude, and fundamental frequency [45]. As a technical term in phonological descriptions of spoken languages, intonation refers to patterned variation in voiced source pitch that serves to contrast and to organize words and larger utterances [46]. While many techniques use low-level audio features or optimization black-box model training, focusing on the features that humans use to recognize speech will likely be a more long-term robust approach to detection. Structured language is defined not just by the words we use, but also by the acoustic features of our voices, known as prosody. Prosody is a linguistic catch-all for the suprasegmental parts of speech (e.g., stress, pitch, and tone). In linguistics, prosodic analysis is generally used to capture abnormal variations in prosody, voice quality, and pronunciations to diagnose pathological speech [47].

Generally, Afaan Oromoo offers a diverse linguistic situation due to its vowel and consonant systems, characteristics, and prosodic patterns, among other phonetic and auditory characteristics. In tasks like ADFD, where phonetic and auditory signals are

employed to detect differences in speech, these qualities are essential for speech-processing applications and offer insightful information [48].

2.3 Phonetics and Linguistics in Audio Deepfake Detection

The development of ADF technology, which modifies audio recordings to create synthetic speech that sounds human, presents challenging obstacles to the identification of deceiving audio material [49]. The application of phonetic and linguistic analysis to ADF detection is discussed in this section, with a focus on the areas which are important in creating reliable detection systems.

Phonetic Features in Deepfake Detection

Teshita Kebede *et al.* [50] presented to investigate and develop a speech-based command and control system for Afan Oromo access phonetics provides a foundation for distinctive between genuine and synthetic speech. Several phonetic features can be studied to identify inconsistencies that may indicate manipulation. Acoustic analysis which analyzes spectral features such as formant, pitch, and intensity can help differentiate between natural and synthesized voices. Variations in rhythm, stress, and intonation patterns are essential components of speech. ADF often lacks the natural prosody found in authentic speech, making it detectable through analysis of these features.

Linguistic Features in DF Detection

Speech patterns the syntactic and semantic structures of sentences, provide context for understanding speech. So, in addition to phonetics, linguistic analysis can uncover anomalies in speech that are suggestive of DF technology [51]. DF models may generate grammatically correct but contextually inappropriate phrases, which can be flagged during analysis. Lexical diversity genuine speakers show a range of vocabulary and expressions that vary with context [52]. DF systems may produce repetitive or overly simplistic language, highlighting a lack of true linguistic diversity. Speaker characteristics are another linguistic clue related to dialect, accent, and individual speech patterns can also be used for detection [53].

2.5 Audio Deepfake Detection Approaches

The emergence of audio DF technology is transforming the application of speech technologies. While its widespread use has brought convenience, it has also raised

unprecedented security and ethical challenges. Using TTS or VC techniques, DF technology can mimic specific individuals' voices, such as political figures, enabling the spread of misinformation and even interfering with elections [54].

Feature-based Approaches

Feature-based analysis is one of the traditional methods for detecting ADFs, which focuses on extracting specific characteristics from audio signals, such as Mel-frequency Cepstral coefficients (MFCCs), spectrograms, and linear predictive coding (LPC). These features capture the timbre, pitch, and other attributes of speech, which can help differentiate between genuine and fake audio. Studies leveraging MFCC and LPC features have shown that these traditional speech features can successfully discriminate between natural and synthetic speech by identifying subtle inconsistencies introduced during the generation process [55].

Deep Learning-based Approaches

With models like CNNs, RNNs, and transformers being frequently used, DL has become a potent tool for DF detection. When CNNs have been extensively used for analyzing spectral features of audio, allowing them to detect minute anomalies that may indicate synthetic tampering. Another one RNNs is, particularly architectures like LSTM and GRU, are proficient in capturing temporal dependencies within audio, which can be crucial in identifying speech patterns altered by DF models. Transformers, with their ability to model long-term dependencies in audio, have shown great success in detecting highly complex synthetic audio [42, 43].

2.4 Deepfake Detection in Low-Resource Languages

DF technology's quick development has sped thoughtful doubts about the possible misuse of synthetic audio, especially in low-resource languages where there is a lack of data and detection tools [56]. Large, well-annotated datasets are hard to come by for low-resource languages, which make it difficult to train dependable and accurate audio DF detection algorithms [57].

Scarcity: One of the primary challenges in detecting ADF in low-resource languages is the lack of extensive, labeled datasets for training ML models. Many low-resource languages do not have sufficient and labeled audio datasets available for public use, making it difficult to build train DL models effectively [35, 50].

Linguistic Diversity: Many low resource languages often show high linguistic diversity, with multiple dialects and complex phonetic systems that differ significantly from more widely spoken languages [51]. When this diversity confuses the design of detection model or framework, as models trained on a small subset of data may not perform well across the full range of phonetic and prosodic variations within the language [58].

Data Augmentation: The process of data augmentation techniques during dataset creation such as pitch shifting, time stretching, and noise addition, can help create more robust datasets for low-resource languages [59]. This allows researchers to artificially increase small datasets, improving the ability of models to detect ADF in diverse linguistic contexts [58].

Multi-modal Approaches: The usage of combining audio with other modalities, such as visual or textual information, may enhance DF detection in low-resource languages. By leveraging additional cues, such as lip movement in videos or the context of spoken text, models can make more informed decisions even in the presence of limited audio data [51, 53].

Collaborative Efforts for Dataset Creation: Collaborative initiatives between researchers, linguists, and communities can help gather and create datasets for low-resource languages. Creating open-source datasets and development cross-institutional cooperation can accelerate progress in developing ADFD systems for underrepresented languages [51,54].

2.5 Datasets for Audio Deepfake Detection

Detecting ADF, which involves artificially generated or manipulated audio to mimic a person's voice, has garnered significant attention due to its implications for security, trust, and misinformation [60]. A variety of datasets and tools have emerged to address this challenge, providing researchers with resources to train and evaluate models aimed at detecting these DFs.

TheM-AILABS Speech (2018): This dataset contains 18.7 hours of German speech recordings and was first made public by the M-AILABS team. Despite its primary purpose of text-to-speech synthesis, researchers have also employed it to produce artificial DF audio to assess spoofing detectors [61].

Baidu Silicon Valley AI Lab Cloned Audio (2018): It was created by Baidu's Silicon Valley AI Lab and includes over 6 hours of synthetically generated English cloned audio. It acts as a standard by which to measure English synthetic speech detection based on deep learning [62].

H-Voice (2020): This multilingual dataset consists of 6,672 files that contain both synthetic and imitation DFs. With support for Spanish, English, Portuguese, French, and Tagalog, it facilitates the investigation of cross-lingual variances, and the difficulties associated with identifying DF audio in a variety of linguistic contexts [63].

ASVspoof 2021 Challenge (2021): As one component of an international challenge, this dataset (size not disclosed) focuses on English-language synthetic DF speech. It has been instrumental in benchmarking state-of-the-art spoofing countermeasures and driving DF detection research [64].

FakeAVCeleb (2021): With 20,490 files of simulated English speech, FakeAVCeleb is built from celebrity speech and has been used extensively in audio visual DF research. Its realistic nature and scale make it helpful for assessing multimodal DF detection systems [65].

ADD (2022): The ADD dataset contains 85 hours of synthetic Chinese speech and was developed as part of an ADF Detection Challenge. Its large size and language specificity address the need of robust spoofing detectors in non-English settings [66].

Urdu Audio Deepfake Dataset (2022): This corpus comprises 10 hours of synthetic Urdu speech, providing a useful asset for DF detection studies in low resource languages. Its availability enables the generalization of evaluation methods to beyond high resource languages like English or Chinese [67].

2.6 Summary

Here in this chapter, we present the background material necessary for the current study on ADFD. These include significant issues like a detailed introduction to DF technology, explaining the fundamental principles involved in the generation and control of artificial media. over-all principles that define ADFD were examined, with an importance on the growing mandate for audio-based misinformation detection. In addition, the methods pertinent to ADFD were presented, comprising machine learning and signal processing

methodologies. Finally, the contribution of linguistics and phonetics to the detection of disinformation was discussed, showing how linguistic properties, i.e., phonetic composition, influence detection models' performance. In addition, we investigated the unique challenges in detecting DF in low-resource languages, provided a concise explanation of the limited data and technological resources for underrepresented languages, and analyzed their phonetic and acoustic characteristics, describing how they pose unique challenges for DF detection. The chapter concluded with an overview of the principal findings and an appeal for more research to address the existing gaps, especially for low-resource languages.

Chapter 3.Related Work

3.1 Introduction

The richness of eagerly accessible tools and techniques that can produce synthetic audio has drawn a lot of interest to ADF detection in several languages recently. As a result, this Chapter provides an overview of some recent relevant works, the results they obtained using various data sets, and the techniques they employed to identify ADF.

3.2 Audio Deepfake Detection Using Deep Learning

Munir *et al.* [67] comprises a comprehensive analysis of the development of an Urdu ADF detection to address growing fears of DF attacks on Automatic Speaker Verification (ASV) systems. The datasets were selected by the authors with great attention to phonemic coverage and balance, with real speech as well as spoofed audio from VITS, TTS and Tacotron TTS. The study also showed a human evaluation experiment to test the performance of humans in detecting real and synthetic audios. ROC curve analysis provided an AUC of 0.63, indicating that human listeners could not detect DF audio, with approximately one-third of the fake samples being detected as real. The findings highlight the need for technology-based DF detection systems, mainly for Urdu, whose existing datasets are bare. The study limited by requirement on very small and regulated Urdu datasets, which might not accurately reflect the acoustic diversity and spontaneous speech situations present in real-world settings, is one of its main limitations.

Flórez *et al.* [68] about Latin America language presents a comprehensive study on the development of a voice anti-spoofing dataset specifically designed to address the challenges posed by Latin American Spanish accents. The authors created the dataset by using advanced voice conversion and TTS techniques and generating both genuine audio samples and manipulated audio that replicate the various accents found in Latin America. The dataset consists of 8,000 audio recordings, including a balanced mix of authentic speech and synthetic audio. The study shows the effectiveness of the proposed dataset

through various experiments utilizing different ML and DL models for anti-spoofing detection. Especially, the study found that DL architectures, particularly CNNs, achieved an impressive detection accuracy of 96.8% when trained on the dataset. In contrast, conventional models like SVMs and Random Forests recorded accuracies of around 83.5% and 80.2%, respectively, illustrating the superior performance of DL approaches in this domain.

The performance matrices which are precision and recall metrics for the best-performing model surpassed 94%, indicating a high level of consistency in recognizing fake audio. Furthermore, the dataset is focus on the rich variety of Latin American Spanish accents significantly improved the models' robustness and adaptability, enabling them to generalize better to unseen accents during testing. By providing a tailored dataset that reflects the linguistic diversity of Latin America, this work lays a foundation for future advancements in voice anti-spoofing research and applications, ultimately contributing to the development of more secure voice recognition systems. While this research improves voice anti-spoofing for Spanish accents in Latin America, it has drawbacks, including a narrow linguistic focus that may limit its applicability to other languages like Afaan Oromoo. And the study focuses on the traditional ML and DL algorithms CNNs rather than examining other designs hybrid DL models or Transformers models, which could result in even greater gains in efficiency and resilience.

A significant contribution to the field of audio anti-spoofing is made by Müller *et al* [69] who provide a varied dataset designed for anti-spoofing system multiple languages. Recognizing the growing threats posed by spoofing attacks, particularly in voice recognition systems. Dataset over 12,000 audio recordings in different languages, including English, Mandarin, Spanish, and Arabic, featuring both genuine speech and synthetic audio were produced through different synthesis techniques, such as TTS and voice conversion. The authors also designed the dataset to cover various spoofing situations to ensure a wide range of phonetic and acoustic characteristics, so enhancing the dataset's applicability to real world conditions. The results from the experiment held by using the created dataset confirmed the state-of-the-art of DL models, particularly CNNs and LSTM networks, significantly outperformed conventional ML approaches in detecting spoofed audio. For instance, the best-performing CNN model achieved an impressive accuracy of 97.5% in distinguishing between genuine and synthetic audio across the diverse language samples.

Furthermore, the study shown that using multi-language samples in training improved the models' generalization possible, leading to well performance in unseen languages. The performance matrices, precision and recall for the DL models consistently scored 90%, emphasizing their effectiveness in reducing false acceptance and false negative rates in anti-spoofing applications. The created dataset helped as a benchmark for future studies, aiming to improve the strength and consistency of audio recognition technologies in the face of increasingly sophisticated spoofing threats. This study limited by dataset, because the dataset mostly made up of languages such as English, Mandarin, Spanish, and Arabic. Because of this, the unique sounds and grammar of Afaan Oromoo, along with its dialects and conversational forms, are not represented.

Asif *et al.* [70] presents a comprehension between DL and the traditional one, which is conventional ML algorithms applied to Arabic audio clips, concentrating on several features of audio processing and classification. The study emphasizes the challenges posed by the Arabic language in speech processing tasks, such as dialectal variations, speech rate, and the presence of diverse phonetic sounds. The authors of the Arabic audio DF, discovered different ML techniques, including SVM, k-NN, Decision Trees, and Naive Bayes, as well as DL approaches like CNNs and RNNs. Their study included feature extraction methods such as MFCC and spectrograms, which were used as inputs for both traditional and DL algorithms. The results demonstrated that DL models, particularly CNNs and RNNs, outperformed conventional algorithms in terms of accuracy and robustness when dealing with complex audio features and noisy data. For example, CNNs achieved an accuracy of 92.5% on a dataset of Arabic spoken digits, whereas traditional methods like SVMs achieved a maximum accuracy of 78.4%. The DL models showed greater performance in capturing temporal patterns and phonetic variations when they were compared, due to their ability to automatically learn hierarchical representations of the audio data. Additionally, when spectrograms were used as inputs, the DL models' accuracy further improved by approximately 4%, compared to using MFCC features.

The computational cost is associated with DL techniques, with training times for CNNs and RNNs being significantly longer than those for conventional methods. However, the authors concluded that the improved performance justifies the increased computational requirements, especially for tasks that demand high accuracy, which is speech recognition and speaker identification in the Arabic language. One of the main

limitations of this research is its limited focus on controlled Arabic audio datasets and the computational expense of DL models limits their applicability to resource constrained.

Ma *et al.* [71] presented a comprehensive technique by utilizing the created dataset specifically designed to address the challenges of detecting fake audio, mainly ADF created through advanced synthesis techniques. The created dataset was created following by leveraging state of the art TTS systems and voice conversion methods to generate fake audio samples in Mandarin Chinese, which were then carefully labeled. The dataset includes over 10,000 audio clips, with a balanced distribution of real and synthetic samples. In terms of results, the authors showed a series of research using several detection algorithms, including traditional machine learning models such as SVMs and DL models like CNNs and RNNs.

The outcomes of the study show that DL models provocatively outperformed ML approaches, by realizing a recognition accuracy of 95.2% for fake audio samples. In contrast, ML models such as SVMs and k-NN generated accuracies around 85% and 81.5%, respectively. Also, the authors emphasized that the performance matrices measures, which were precision and recall rates for the DL models were consistently above 90%, demonstrating their effectiveness in distinguishing between real and synthetic audio. The study also examined the impact of different audio features on recognition performance, revealing that using spectrogram representations led to a 7% improvement in accuracy compared to raw audio. As indicated by the results of the study, it is a valuable resource for future research in ADFD helping the development and evaluation of more robust detection systems. The work underscores the importance of addressing DF audio challenges, particularly within the context of the Chinese language, and sets a benchmark for subsequent studies in this evolving field. A significant limitation of this study is that it only considered Mandarin Chinese, and this might limit the use of this detection technique to other languages and computational cost and data diversity challenges.

Zhang *et al.* [72] provide a detailed summary of the rapid progress in audio DF generation and detection, outlining both the advantages and the severe threats associated with these technologies. The authors detail how recent research TTS and VC technology has resulted in the generation of very natural sounding synthesized speech. While this is beneficial to applications such as virtual assistants and creative content, it also provides

potential for misuse, such as impersonation, disinformation, and security concerns in domains like smart surveillance and biometric authentication. The survey explores recent datasets, detection approaches from feature extraction, classification models, to end-to-end systems and mentions emerging research areas such as privacy, explainability, and fairness. The analysis by the authors reveals that both human auditors and state-of-the-art machine learning models find it very difficult to consistently detect audio deepfakes. Research indicates that human detection accuracy can be as low as 73%, while even machine-based systems do not consistently outperform human intuition with accuracy. The article also covers specialized subjects such as privacy-preserving detection, interpretability, and fairness, along with the quantitative comparison of many detection models. Perhaps there is no more significant finding of this research than the discovery of ongoing vulnerabilities in human and machine detection systems, particularly with increasing sophistication in methods of generating deepfakes. The paper does acknowledge a few limitations, such as a scarcity of extensive and representative datasets, limited work on real-world deployments, and a requirement for stronger, more scalable, and more generally applicable detection techniques.

3.3 Summary

Studying ADFD has grown year by year in terms of languages, approaches, and datasets, as well as pointing out important challenges and achievements. Initial research in 2018–2020 was limited to English synthetic speech detection based on DL approaches such as Bidirectional LSTMs, CNNs, ResNet-34, and Deep-Sonar on datasets such as ASVspoof and the like. However, all these models suffered from overfitting, scalability issues, and computationally expensive operations, particularly those of audio-to-image mapping. In 2020, studies diversified to other languages like Classical Arabic, Spanish, Portuguese, French, and Tagalog, with imitation-based detection methods like logistic regression and conventional classifiers. Recent work in 2022 and 2023 has been along the lines of low-resource languages and multilingual, where CNNs, LSTMs, and hybrid models have been employed with a focus on computational efficiency challenges and language adaptability issues. Notably, the publication of the 2023 Urdu DF audio dataset, founded on the AASIST-L model, brought to light the limited human ability to detect DFs as only a third of the fake samples were correctly detected. Overall, although DF detection models have been improved, scalability, computational cost, and cross-linguistic generalization remain unresolved problems. Collectively, these efforts emphasize the

importance of tailored datasets and advanced detection techniques to enhance ADFD and ensure the reliability of voice recognition systems across diverse language backgrounds, including Afaan Oromoo.

Chapter 4. Design of the Proposed System

4.1 Introduction

This Chapter presents the proposed solution for Afaan Oromoo ADFD using hybrid DL method based on the literature review, which are CNNs and LSTM networks. Section 4.2 Dataset creation. Section 4.3 provides an overview of the general architecture of the proposed ADFD system for Afaan Oromoo, and its components are thoroughly discussed.

4.2 Dataset Creation

The dataset creation process is our first attempt at developing the Afaan Oromoo ADFD system, as shown in Figure 4.1. We utilized RTVC techniques to create the synthetic dataset. The dataset should contain labelled real and synthetic data types, encompassing the collection of a dataset that contains real and synthetic audio datasets in an approach that allows the model to be able to understand the distinction between real and synthesized speech features. The genuine audio corpus is collected from social media, YouTube, and websites of OBN (Oromia Broadcasting Network), which contain interviews of different native speakers of Afaan Oromoo so that it has natural prosodic and phonetic features. The synthetic audio dataset is created through RTVC methods, where synthetic speech is modelled after real voices. Both real and synthetic speech samples are stored in a centralized repository, referred to as AOADS (Afaan Oromoo Audio Dataset), which contain 19.5 hrs.

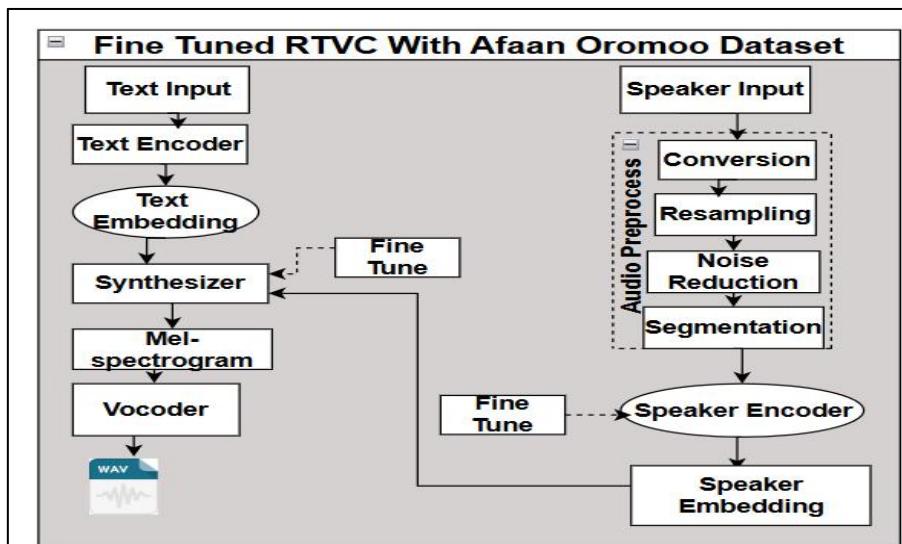


Figure 4- 1 Real Time Voice Cloning to Create Synthetic Dataset

Text Encoder: It uses a Text Input to which the text input is passed through a Text Encoder for numerical Text Embeddings conversion.

Speaker audio: sample for generating the speech imitating a speaker. Fine-Tuning Process applied to adapt it to fit in the model, which is needed for this language with unique phonetic and prosodic features of Afaan Oromoo.

Speaker encoder: The processed data is fed into a fine-tuned Speaker Encoder, which captures the overall speaker characteristics in a single representation, known as the Speaker Embedding.

Synthesizer: fine-tuned on Afaan Oromoo data and combines the Text Embedding's and the Speaker Embedding to generate a Mel Spectrogram, which is converted to an audio wave by the Vocoder.

Vocoder: Converts the Mel Spectrogram into an audio waveform. That results in an audio file that sounds very close to the natural voice and tone of the target speaker. In this study, we present HiFi-GAN (High Fidelity Generative Adversarial Network), which can successfully generate high-quality speech audio from Mel-spectrograms. It relies on a generative adversarial network (GAN) framework, which has a generator that employs mel-spectrograms as samples for generating sound waveforms, and the discriminators make sure that the produced audio is realistic. The system effectively adapts RTVC for Afaan Oromoo by fine-tuning speaker embedding. RTVC as shown by Figure 4.1, synthesizing authentic voices begins with two special encoders a text encoder and a speaker encoder that capture the essence of the input text and the speaker's unique voice characteristics. The encoder inputs to a synthesizer which accepts the aggregated data and converts it into high resolution Mel-spectrogram, which is a visual representation of the frequencies of sound that contain both linguistic and features of specific speaker. The Vocoder called HiFi-GAN develops the spectrogram by converting it into an audio waveform that closely look like natural sound.

With the RTVC model, Afaan Oromoo transcription used as text input to the model to generate synthetic audio. The generation of imperceptible synthetic audio clips was made possible by copying the voice characteristics of specific speakers within the dataset and synthesizing speech to carry out provided textual content. To obtain diversity in the synthetic dataset, uniform text inputs derived from transcribed original audio and precise aligned by using, Aeneas aligner were used to create several audio segments consisting of a variety of speaker voices, thereby encompassing various speaking styles, tones, and prosodic nuances. 7012 synthetic audio segments were created.

To fit the RTVC model to Afaan Oromo's unique phonetic and prosodic features, an end-to-end fine-tuning pipeline was established. This strategy used trendy DL frameworks, PyTorch, to facilitate model training and adaptation, alongside audio processing libraries,

Librosa, to analyze and extract features. Fine-tuning change precise hyper parameter tuning and domain adaptation techniques to capture the fine-grained speech characteristics inbuilt to Afaan Oromoo. Furthermore, transcript alignment was also important in making the synthetic speech properly capture the textual content assigned to it, thus yielding a balanced and meaningful dataset for training the DF detection model. Such precision was augmented by the model's ability to create and identify realistic synthetic speech, thus contributing to the overall efficiency of the detection system.

Transcribed and aligned audio recordings made up a comprehensive dataset from the real audio data source. Tokenizer development for Afaan Oromoo speech recognition tokenizer adapted for Afaan Oromoo speech recognition tasks using the Wav2Vec2 framework. A complete set of characters and special symbols that show the characteristic phonetic and orthographic features of Afaan Oromoo is provided. The character set consists of the complete Latin alphabet, a space symbol, and a group of additional tokens to mark language-specific diagraphs and phonemes, including "ch", "dh", "ny", "sh", "ts", and Hudhaa (" ' "). Concurrently, the speech encoder was modified using various aligned speaker samples to generate strong and unique embeddings.

4.3 System Architecture

We propose a hybrid model architecture that combines CNN and LSTM. CNN is employed to capture local patterns in the audio spectrogram, while LSTM is used to model the temporal dependencies in the sequential audio data. The model is trained on labeled datasets of authentic and fake audio samples in Afaan Oromoo. The CNN algorithm is used in this system to identify and record local patterns and characteristics in the audio data, with an emphasis on the spectrograms, which show the frequency content of the audio signal with time. Given its capacity to recognize spatial elements, the CNN is preferably suitable for picking up on minute, complex tones in the spectrogram that might point to the existence of artificial audio modifications. Figure 4.2 shows the proposed general architecture of the system.

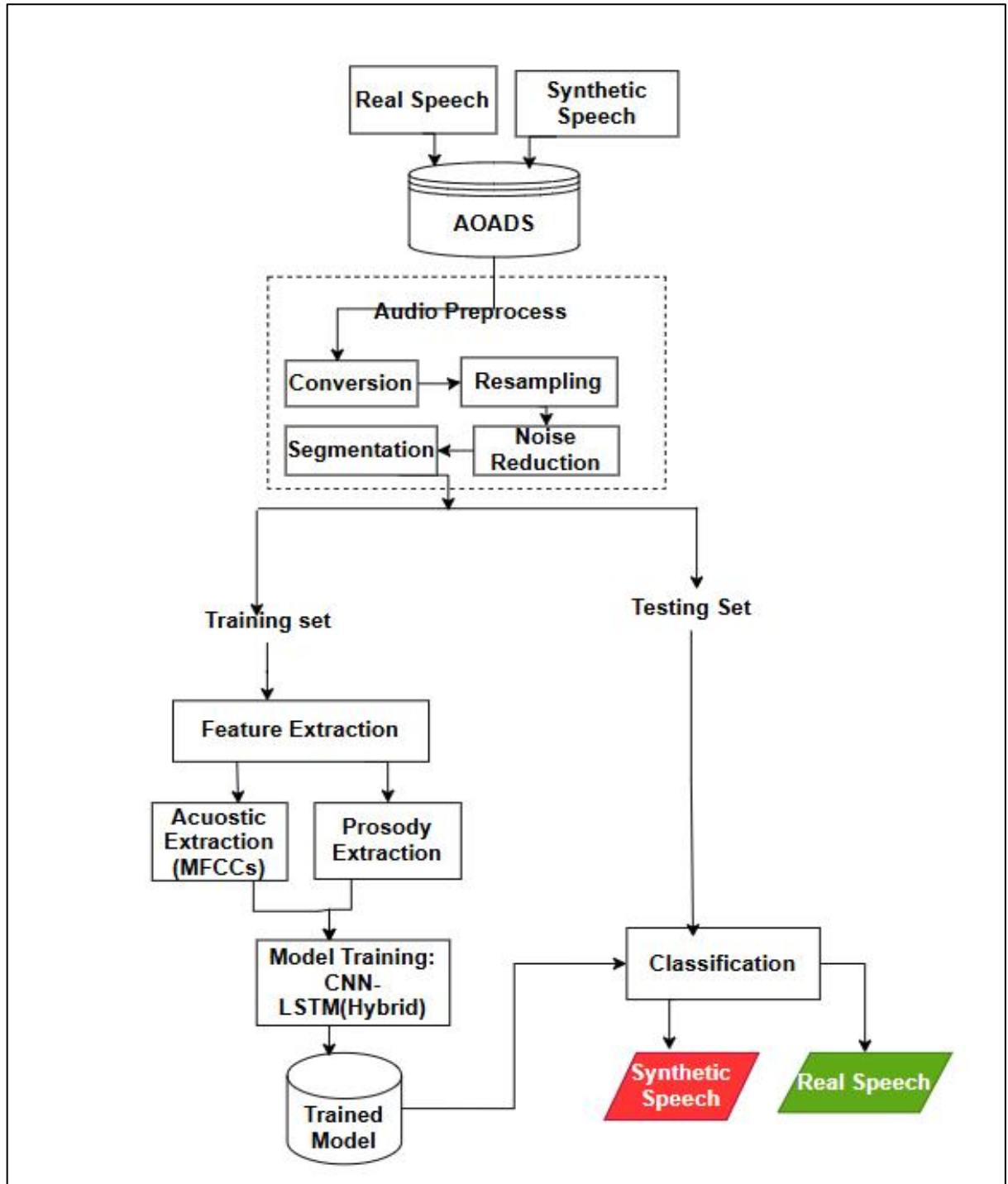


Figure 4-2 Proposed Audio Deepfake Detection for Afaan Oromoo

4.3.1 Components of the Architecture

The components of the proposed architecture are explained in the following.

4.3.1.1 Dataset Preprocessing

The preprocessing of AOADS was processed using a detailed way to guarantee high quality training input for the proposed hybrid DL model. The dataset preprocessing was applied on both the training and testing set of the datasets individually, because to protect data leak if we apply the preprocessing on the whole dataset.

Converting Audio to Standards: Audio datasets may come in various formats, sample rates, or resolutions. Preprocessing ensures standardization, making it easier to work with diverse datasets and preventing inconsistencies that can affect model performance. We convert the audio to standardized format WAV which helps for consistent feature extraction.

Resampling: for reliable processing by resampling every recording at a constant sampling rate and ensures all audio files are converted to a uniform sampling rate, which is essential for consistent input dimensions in DL models. We utilized librosa python library to resample audio to a fixed rate 16 kHz.

Noise Reduction: Audio data collected from real-world environments often contains background noise, interference, or artefacts. Preprocessing methods like filtering and denoising can help to remove unwanted noise, ensuring that the model focuses on the relevant signal. The speech signals were made clearer by using the python library librosa which is sophisticated noise reduction techniques to remove background nosiness which improve signal to noise ratio, which allows the model to focus on relevant speech features.

Segmentation: is the process of dividing a long audio recording into smaller, more manageable segments or chunks. The segmentation was based on the natural speech boundaries, which are silence, pauses, utterances, sentences, or words. The audio was divided into digestible portions of 5 seconds. Algorithm 4.1 shows the way we preprocessed AOADS.

Algorithm 4- 1 Audio Preprocessing Algorithm

Input:

```
raw_audio_files    // List of raw audio files (various
formats, sample rates, noise levels)
target_format = WAV
target_sample_rate = 16000 Hz
segment_duration = 5 seconds
```

```

silence_threshold // Threshold for silence detection
min_silence_length // Minimum silence duration for
segmentation

```

Output:

```

preprocessed_audio_segments // List of clean, standardized,
segmented audio clips ready for training/testing

```

Start

```

Initialize preprocessed_audio_segments = empty list

```

```

For each audio_file in raw_audio_files do

```

```

// Step 1: Convert audio to standardized format (WAV)

```

```

audio_signal, sample_rate = convert_to_wav(audio_file,
target_format)

```

```

// Step 2: Resample audio to fixed sample rate 16 kHz)

```

```

If sample_rate != target_sample_rate then

```

```

audio_signal = resample_audio(audio_signal, sample_rate,
target_sample_rate)

```

```

sample_rate = target_sample_rate

```

```

End If

```

```

// Step 3: Noise reduction to improve signal quality

```

```

audio_signal = reduce_noise(audio_signal, sample_rate)

```

```

// Step 4: Segment audio based on silence detection and
fixed segment duration

```

```

segments = segment_audio_by_silence(audio_signal,
sample_rate, segment_duration, silence_threshold,
min_silence_length)

```

```

// Step 5: Append all segments from this audio file to the
final list

```

```

preprocessed_audio_segments.append_all(segments)

```

```

End For

```

```

Return preprocessed_audio_segments

```

```

End

```

4.3.1.2 Dataset Splitting

Partitioning of AOADS into training, 80% and testing set, 20% sets is essential to build strong DL models. We separated our AOADS of audio samples into training and testing subsets to guarantee reliable training and model evaluation. Algorithm 4.2 shows the process of the data set split held on AOADS.

Algorithm 4- 2 Dataset Splitting Algorithm

Input: AOADS (full dataset with labels)

Output: Training Set, Testing_Set

Start

```
    shuffle (AOADS_real)
    shuffle (AOADS_synthetic)
    N_real = length(AOADS_real)
    N_synthetic = length(AOADS_synthetic)
    N_real_train = floor(0.8 * N_real)
    N_real_test = N_real - N_real_train
    N_synthetic_train = floor(0.8 * N_synthetic)
    N_synthetic_test = N_synthetic - N_synthetic_train
    train_real = AOADS_real[0 : N_real_train]
    test_real      =      AOADS_real[N_real_train      :      ]
    train_synthetic = AOADS_synthetic[0 : N_synthetic_train]
    test_synthetic = AOADS_synthetic[N_synthetic_train : ]
    Training_Set = concatenate(train_real, train_synthetic)
    Testing_Set = concatenate (test_real, test_synthetic)
    Return Training_Set, Testing_Set
```

End

4.3.1.3 Feature Extraction

After preprocessing and splitting the dataset into training and testing sets, the feature extraction process was performed to identify critical attributes of the audio signals. This module generates essential information that will be used during the training phase to build an Audio Deepfake Detection (ADFD) system for Afaan Oromoo.

MFCCs are computed from each audio segment. MFCCs capture the short-term power spectrum of speech. Prosodic features are critical for distinguishing between natural and synthetic speech, as synthetic voices often lack natural prosodic variation. In Afaan Oromoo case pitch(F0), energy (loudness) and duration (timing) are extracted.

Acoustic Feature Extraction

MFCCs is among the most widely used features in speech and audio processing, which provide a compact representation of the spectral properties of an audio signal by capturing its short-term power spectrum on a Mel-frequency scale, which aligns with human audio opinion.

Prosody Feature Extraction

Prosody features encode the suprasegmental characteristics of speech “how the something is said,” rather than “what is said”. They comprise rhythm, speech rate, intonation, pause, pitch, and duration, which all play a role in communicating meaning, emotion, and speaker attitude. As illustrated in Figure 4.2, prosody extraction is performed in parallel with acoustic feature extraction, because in addition to MFCCs, prosody features play a vital role in capturing language-specific characteristics of Afaan Oromoo. Unlike spectral features, MFCCs, prosodic features reflect suprasegmental elements that govern the rhythm, speech rate, intonation and timing of speech. These prosodic features extracted from the datasets are critical for distinctive natural speech from synthetic audio, as DF generation often crashes to copy authentic prosodic patterns accurately. The prosody feature extraction module analyzes these temporal and phonetic patterns to complement acoustic features, thereby enhancing the robustness of the detection system. Algorithm 4.3 shows the process of feature extraction from AOADS.

Algorithm 4- 3 Acoustic and Prosody Feature Extraction Algorithm

Input: Labelled Preprocessed_Audio_Set (preprocessed audio samples)

Output: Feature_Set (feature vectors for each sample)

Start

Initialize Feature Set $\leftarrow \emptyset$

```

For each audio sample in Preprocessed_Audio_Set do

Prosody Extraction

    F0_mean, F0_median, F0_std ← Compute Pitch (audio
sample)

    energy ← Compute Energy (audio sample)

    pause    count,    pause    duration    ←    Detect
Pauses(audio_sample)

    speech_rate ← ComputeSpeechRate (audio sample)

    duration ← MeasureDuration(audio_sample)

    prosody_vector ← [F0_mean, F0_median, F0_std, energy,
pause_count, pause_duration, speech_rate, duration]

Acoustic Extraction (MFCCs)

    frames ← FrameSignal(audio_sample)

    mfcc_features ← []

    For each frame in frames do

        mfcc ← ComputeMFCC(frame)

        Append mfcc to mfcc_features

    End For

    mfcc_mean ← Mean(mfcc_features)

    mfcc_std ← StandardDeviation(mfcc_features)

    mfcc_vector ← Concatenate (mfcc_mean, mfcc_std)

End For

    Return Feature_Set

End

```

4.3.1.4 Model Training

Once the AOADS preparation is done, it moves to Training Phase. As shown in Figure 4.2, the training dataset is used to train the model whereas the testing dataset is used to evaluate the model. In the context of DF detection for AOADS, the CNN emphasizes

frequency-time patterns and variations in intensity found within the dataset. These patterns are critical for differentiating between authentic audio and synthetic or modified audio, which frequently display minor irregularities in frequency distributions or signal dynamics. The CNN module processes the AOADS in the form of spectrograms and extracts spatial features. The input to the CNN module is typically a spectrogram matrix generated during preprocessing; these inputs capture language-specific acoustic features such as tonal characteristics, vowel lengths, and phoneme distributions in Afaan Oromoo dataset. The convolutional layers can learn features specific to the language, such as, unique phoneme transitions, tonal variations in vowel-heavy words and signal characteristics of speech in Afaan Oromoo speakers. CNNs rely on pooling layers to reduce the dimensionality of feature maps while preserving the most crucial information. Pooling is essential for maintaining important audio characteristics that differentiate authentic audio from fake. It also makes the DF detection process computationally effective for AOADS. There is a CNN module for extracting spatial features with an emphasis on discovering patterns appearing in spectrograms.

To capture sequential dependencies on sequential data and audio signals, the model's LSTM module is essential. The LSTM module supports in the analysis of the temporal development of audio features, offering insight into the dynamics of speech over time, in the context of identifying DFs in the Afaan Oromoo audio dataset. LSTMs are a type of RNN designed to effectively model long-range dependencies in sequential data, which is essential for tasks like speech analysis for DF detection. Vowel lengthening, vowel duration plays a role in pronunciation and meaning. LSTM can capture how long vowels should last in a natural speech sequence and identify any abnormal variations in synthetic audio. A phoneme transition of the smooth transition from one phoneme to another is crucial for natural-sounding speech. Synthetic audio might have short transitions that are easier for an LSTM to identify, especially when trained on real data. Prosody patterns are highly important in Afaan Oromoo language, where it can affect meaning. Synthetic audio often exhibits unnatural stress patterns, and LSTM can detect these inconsistencies. Speech monitoring the rate at which speech is delivered can be an indicator of authenticity.

The LSTM can capture these anomalies by learning the typical temporal behavior of real Afaan Oromoo speech. LSTM networks are used for temporal sequence modeling, while CNN are used for spatial feature extraction. The model is a hybrid CNN-LSTM design.

Finally in this phase the hybrid ADFD model is trained as shown by Figure 4.1 of the proposed Audio DF detection for Afaan Oromoo. The validation process is also applied on this phase of the model. We made validations to the built model by using the validation dataset we split the dataset.

4.3.1.5 Classification

As shown in Figure 4.2, the detection phase is used to analyze audio samples to see if they are real or not. Input audio is processed through a preprocessing stage where noise is eliminated and standardized. The cleaned audio is then processed through feature analysis, where it is subjected to the trained CNN-LSTM model. The trained model is loaded to the detection phase to analyze spatial and temporal characteristics and compares them to acquired patterns for real and synthetic speech. We used the testing dataset in this phase to classify the audio as real or synthetic depending on the match. The output is a binary choice of whether the audio sample is real or synthetic. Real speech indicates that the audio is genuine, and Synthetic speech indicates that the audio is a DF or synthetic in our case generated by RTVC process.

Example 1: The Sound "Qa": The consonant "Qa" is a gluttonized sound common in Afaan Oromoo but uncommon in many other languages. Real Audio: In phrases like "*Qabsoo*" (struggle), the sound "Qa" is pronounced with a distinct glottal stop and harmonic overtones in the spectrogram. Fake Audio: Fake audio might change the glottalization, causing the spectrogram to show weak or absent harmonics for the "Qa" sound.

Example 2: The Sound "C" The consonant "C" (as in "*Ci*" or "*Co*") in Afaan Oromoo is pronounced as a palatalized "ch" sound. Real Audio: In words like "*Caalaa*" (better/more), the "C" sound exhibits a specific high-frequency energy pattern in the spectrogram. Synthetic Audio: synthetic audio might fail to replicate the quickness and precision of this sound, leading to unclear or inconsistent frequency components.

This systematic method guarantees the Afaan Oromoo DF detection system to be robust and reliable, solving the problems of synthetic speech detection in low-resource languages.

Chapter 5. Experiment and Results

5.1 Introduction

This chapter deals with the experimental analysis of our system, Afaan Oromoo Audio Deepfake Detection. Section 5.2 presents the development environment. Section 5.3 Result and Discussion. Section 5.4 Model Comparison. Section 5.5 Discussion.

5.2 Development Environment

The development environment for this thesis was specifically designed to provide a solid and reliable basis for building and testing a hybrid CNN-LSTM DL model for ADFD.

- The operating system is Ubuntu 22.04 LTS, a popular Linux distribution that is famous for its stability, security, and greater compatibility with newer ML libraries.
- The computing system is prepared with an Intel Core i7 processor 8th gen, 32 GB RAM, and a 1 TB hard disk PC.
- Python programming language for coding purposes. It was used as the main programming language due to its simplicity, large ecosystem, and wide community support for data science and ML.
- In audio processing and feature extraction, Librosa and Parselmouth is a Python library that provides a direct interface to Praat used to extract MFCCs and prosody features of audio Parselmouth provides direct access to PRAAT's functions within. TensorFlow, through its Keras API as the high-level API, is the foundation upon which the CNN-LSTM model is built and trained.
- Sci-kit-learn is used for data splitting, normalization, and assessment using straight performance metrics.
- NumPy and Pandas are supported to enable easy data manipulation and processing of huge datasets.

- Matplotlib and Seaborn enable comprehensive visualization of the model training process.
- Jupyter Notebook integrated into anaconda was used for interactive development and interactive experimentation.
- Draw.io drawing different diagram design activities.
- Microsoft Office 2019 for documentation purposes.

5.3 Results and Discussion

This Section presents a comprehensive performance breakdown, which includes a detailed study of experimental results with respect to several systems of measurement like accuracy, precision, recall, and F1-score of the developed prototype.

MFCC Spectrogram Analysis

Load audio files from the specified data path, extract MFCC features, and return the features along with their labels, where we labeled our dataset real by 0(zero) and synthesized by 1(one). In Figure 5.1 which shows the analysis of MFCC spectrogram, the x-axis represents time, while the y-axis represents MFCC coefficients, highlighting distinct patterns in real and fake speech. MFCC are used to extract key spectral features from audio signals. MFCC spectrograms help differentiate real and fake audio samples based on frequency content. As shown in Figure 5-1, the real audio spectrogram contains a more natural, smoother distribution of MFCC features and larger dynamic range.

The real audio spectrogram also contains a more natural, smoother distribution of MFCC features and a larger dynamic range. The color intensity appears blockier and less evenly distributed, with certain regions showing sudden changes. The synthetic audio spectrogram contains less details and smaller dynamic.

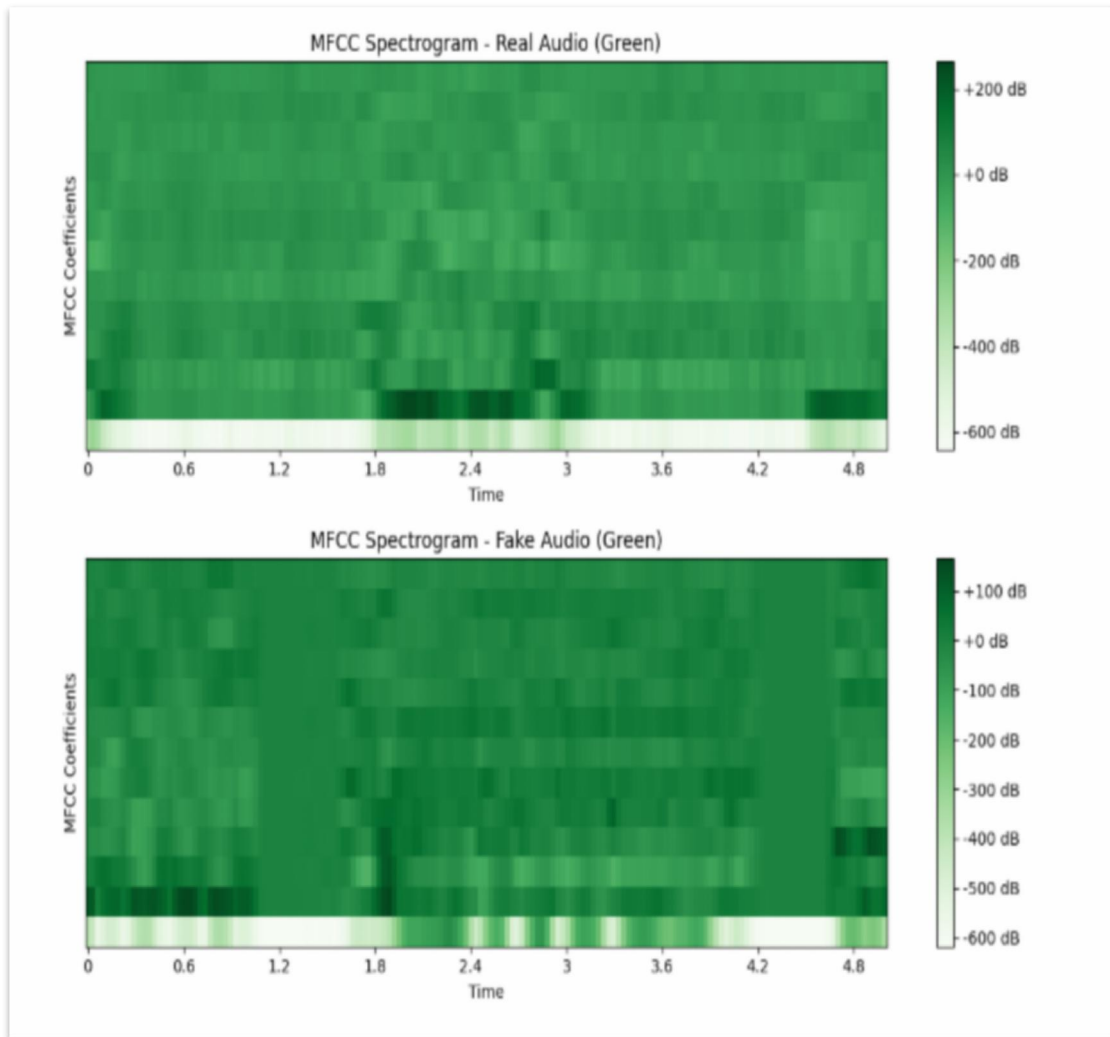


Figure 5-1 Sample of labeled data plot that shows how the data labeled as real and synthetic.

MFCCs were displayed in Figure 5-1 plot. The graphs of MFCC spectrograms with different audio samples that were labelled as real or fake illustrate that the dataset includes both synthetic and real audio recordings.

Principal components of analysis (PCA) Figure 5-2 of MFCC coefficients which reduce the MFCC feature dimensionality, visualizing data in two-dimensional space, the blue dots represent the real audio sample, while the red dots represent synthetic audio sample.

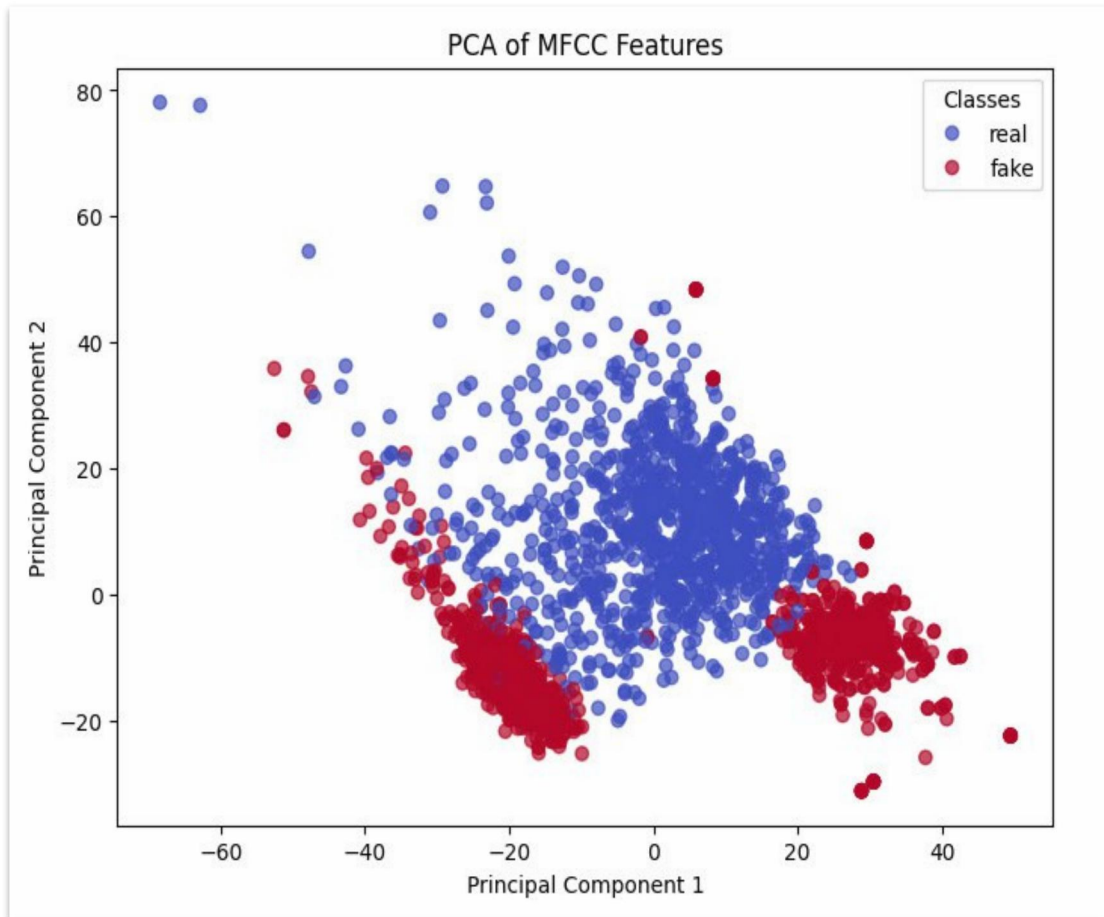


Figure 5-2 Graph of MFCC for sample dataset

The blue board in Figure 5-3 displays the distribution of mean pitch values for speech samples examined. Pitch is fundamental prosodic property that describes the perceptual frequency of the sample voice. The x-axis shows the mean pitch values for specific speeches, whereas the y-axis signifies the quantity of speeches contained in each of the specified pitch ranges. The red board in Figure 5-3 shows the distribution of speech rate values for the same collection of speech. Speech rate, typically expressed in units' syllables per second or words, which is a measure of the velocity of speech. The x-axis is the range of observed speech rates, and the y-axis is the number of utterances for each of the given rate ranges.

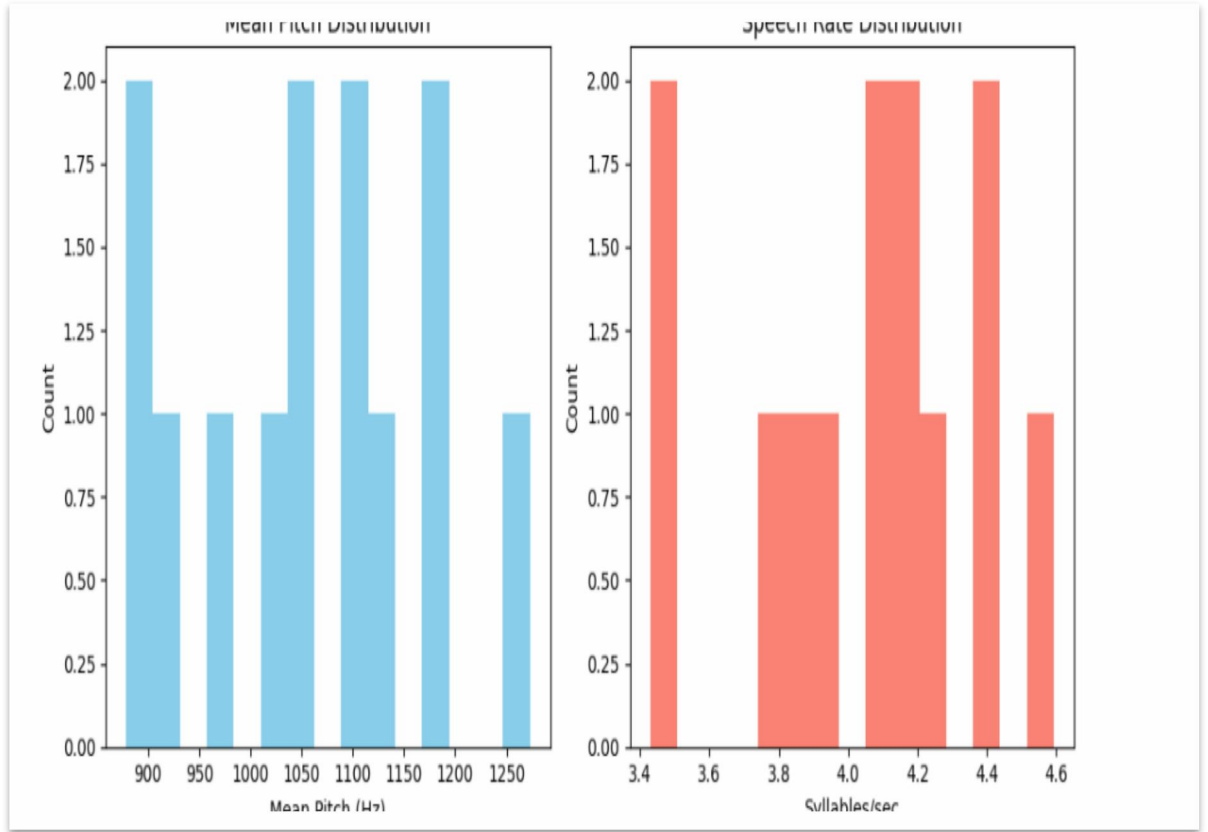


Figure 5- 3 Sample analysis of prosodic feature distributions

Mean pitch distribution as shown by Figure 5-3 tells us about the typical voice pitch in our dataset, which is important for tasks like speaker identification or synthesis. Speech rate distribution gives insight into how fast speakers are talking, which can affect intelligibility and naturalness in speech recognition and synthesis systems.

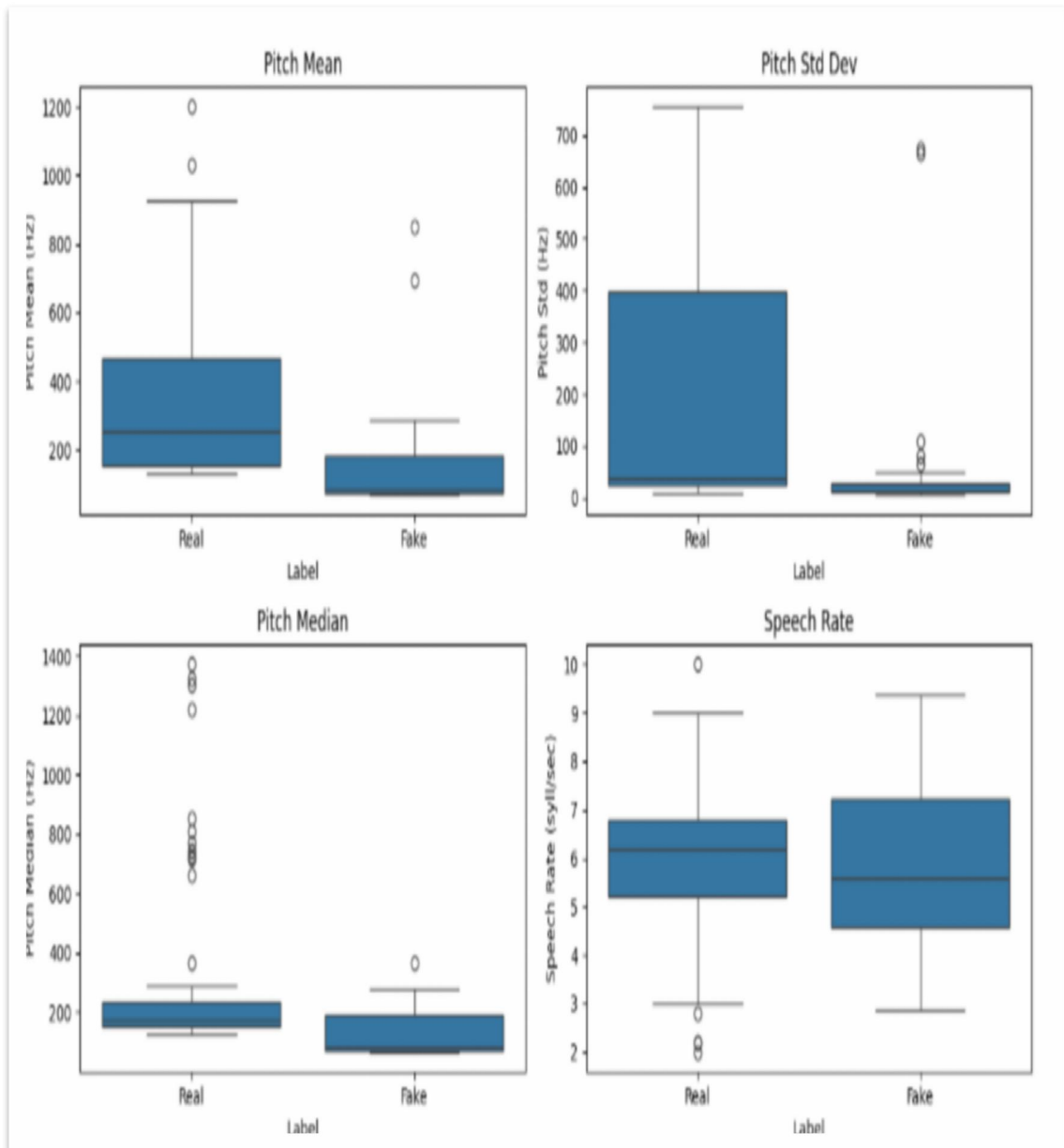


Figure 5- 4 Prosody feature comparison plots of real and synthetic audio

The boxplot illustrates that the real audio expresses a higher and more variable mean and median pitch, along with a much higher pitch standard deviation, which shows us the natural expressiveness and dynamic range that is typical of human speech. In contrast the synthetic audio is characterized by lower and less variable pitch, along with a much flatter pitch standard deviation, resulting in it being more monotonal and synthetic in nature.



Figure 5-5 Sample of compiling and training process of the model

As shown in Figure 5.5 the process of training suggests the model is learning the training data rather than learning patterns that generalize to unseen samples. To solve this issue, we apply early to stop employed to halt training once validation metrics stop improving, preventing the model from overfitting further.

Evaluate the Model on the Test Set

As we can see from Figure 5-6, the performance of the built model AOADFD model is assessed by the testing dataset to identify DF audio in Afaan Oromoo. The evaluation was done using a test dataset, which is important performance indicators such as accuracy and loss were noted.

As can be seen from Table 5.1, high accuracy in training (97.40%) implies that the model successfully learned patterns in training data. The low training loss (0.1946) also confirms the fact that the model successfully adjusted its parameters during training. The test accuracy (95%) confirms the high predictive capability of the model on new data predictions.

Table 5- 1 Evaluation of the model over training and test set.

No	Performance Metrics	Result %
1	Training Accuracy	97.40%
2	Training Loss	0.1946
3	Test Loss	0.25
4	Test Accuracy	95%

Performance Evaluation Results

This report gives a measure of the performance of the model in classifying real and synthetic audio based on precision, recall, F1-score, and accuracy.

Table 5- 2 Summary of the proposed performance matrices results

Matrices	Real	Fake
Precision	0.98 (98% correct when predicted "Real")	0.92 (92% correct when predicted synthetic)
Recall	0.92 (Identifies 92% of real samples)	0.98 (Identifies 98% of synthetic samples)
F1-score	0.95	0.95

These results indicate that when evaluated on the test set, the model achieves a 95% accuracy (i.e., it correctly classifies about 95% of the test samples) and has a test loss of approximately 0.25. A high accuracy and relatively low loss suggest that the model has learned to make accurate predictions on unseen data.

Figure 5-8 shows the accuracy and loss, which indicates the possibility of overfitting, because training accuracy is high, but validation accuracy fluctuates, and validation loss is stable. To combat this, we employed early stopping, which monitors validation loss and halts training when it stops improving, preventing the model from overfitting to the training data. This approach promotes improved generalization to new data and prevents overfitting.

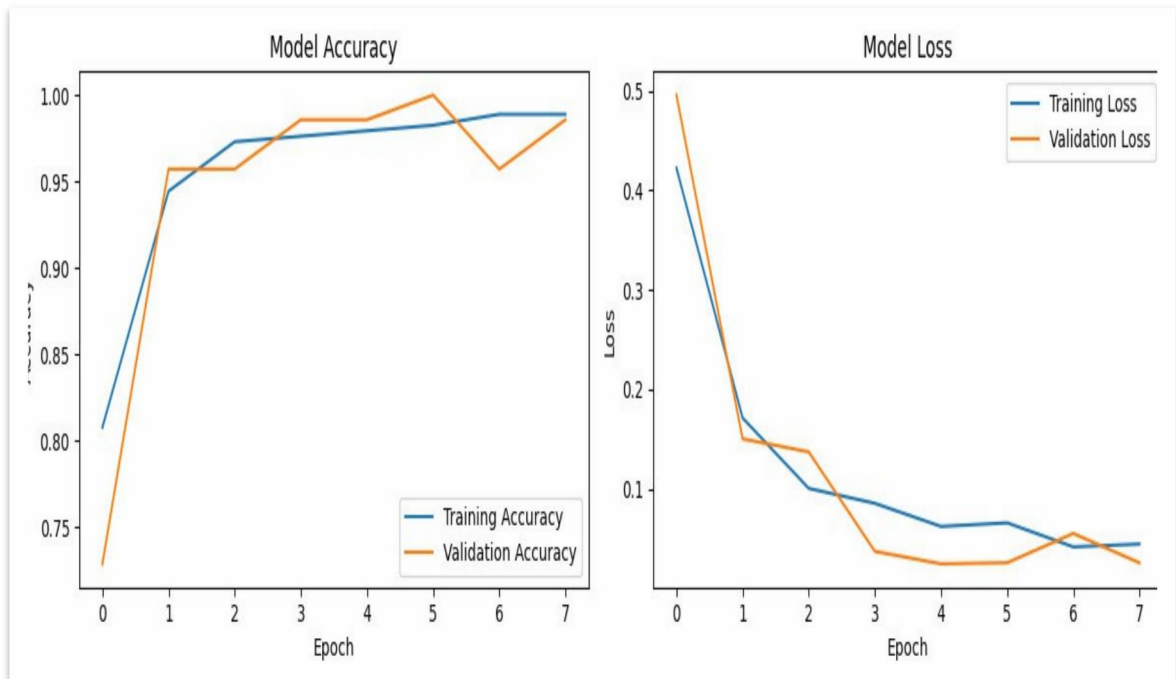


Figure 5-6 Loss and accuracy curves during Training of the proposed model.

5.4 Comparisons of other models over AOADS

CNN model to detect DF audio with AOADS

This report discusses the testing of a CNN model to detect DF audio for Afaan Oromoo over AOADS to compare with our proposed model. The model is trained with a dataset consisting of real and artificially created audio samples, and the model is tested with key testing parameters like accuracy, loss, precision, recall, F1-score, and confusion matrix analysis. The CNN model scores 78% accurate, which means the model is correct in

identifying real and synthetic audio with an accuracy of 78% on the test set. The loss of value, though, suggests that there could be some scope for improvement, possibly by hybrid with other DL algorithms the model performance can be improved. The accuracy of the single CNN-based DF detection model for Afaan Oromoo over AOADS, with a test accuracy of 78%, is considerably less than that of the CNN-LSTM model, with a test accuracy of 95%.

LSTM model to detect DF audio over AOADS

LSTM over AOADS is discussed in this topic to compare with the proposed hybrid model for audio DF detection over the prepared AOADS for Afaan Oromoo audio DF detection. LSTM was trained with AOADS consisting of real and synthetically created audio samples, and the model is tested with key testing parameters like accuracy, loss, precision, recall, F1-score, and confusion matrix analysis. The LSTM model scores 90% accurate, which means the model is correct in identifying real and synthetic audio with an accuracy of 78% on the test set. The loss of value, though, suggests that there could be some scope for improvement, possibly by hybrid with other DL algorithms the model performance can be improved. The accuracy of the single model LSTM based DF detection model for Afaan Oromoo over AOADS, with a test accuracy of 90%, is considerably less than that of the CNN-LSTM model, with a test accuracy of 95%.

As shown in Figure 5-7 the single LSTM model, the model trained well, had high accuracy on the validation and test sets, but the accuracy of the model is less than the proposed hybrid proposed CNN-LSTM model for Afaan Oromoo language.

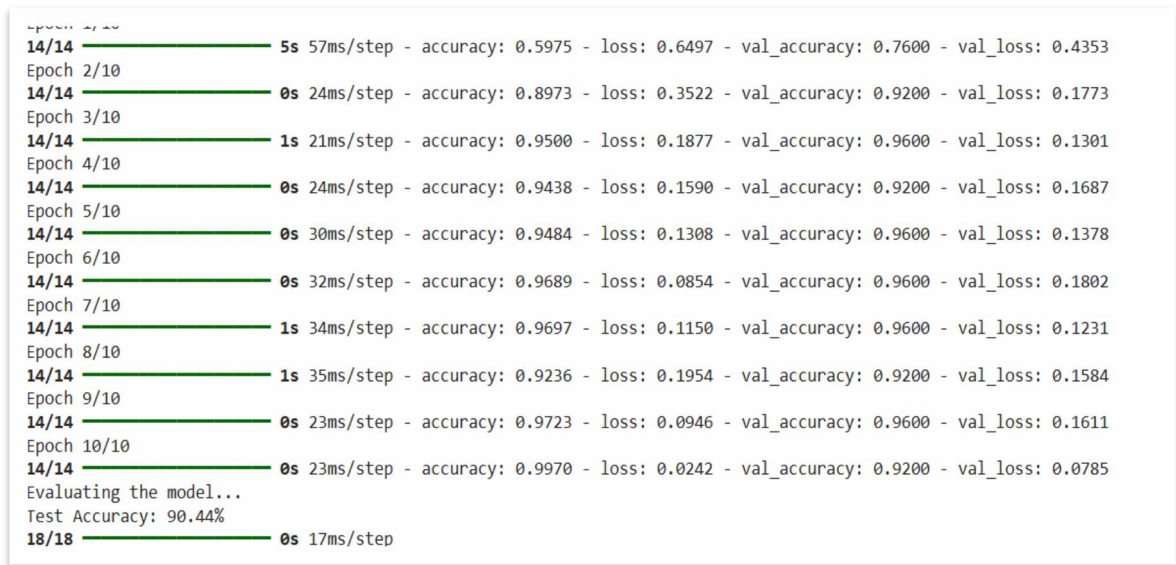


Figure 5- 7 Sample of compiling and training process of the LSTM model

As shown in Figure 5-8 both the training and validation accuracy rising quickly and then the training accuracy being slightly higher than validation accuracy. Model loss is presented in the right plot, which drops for both the training and validation sets, showing that the model is learning properly. The validation loss and accuracy curves closely track the training curves. The model's performance is strong and stable throughout the training process, but the overall performance of LSTM model shows 90% which shows the accuracy of this model is less than the CNN-LSTM, the hybrid proposed model.

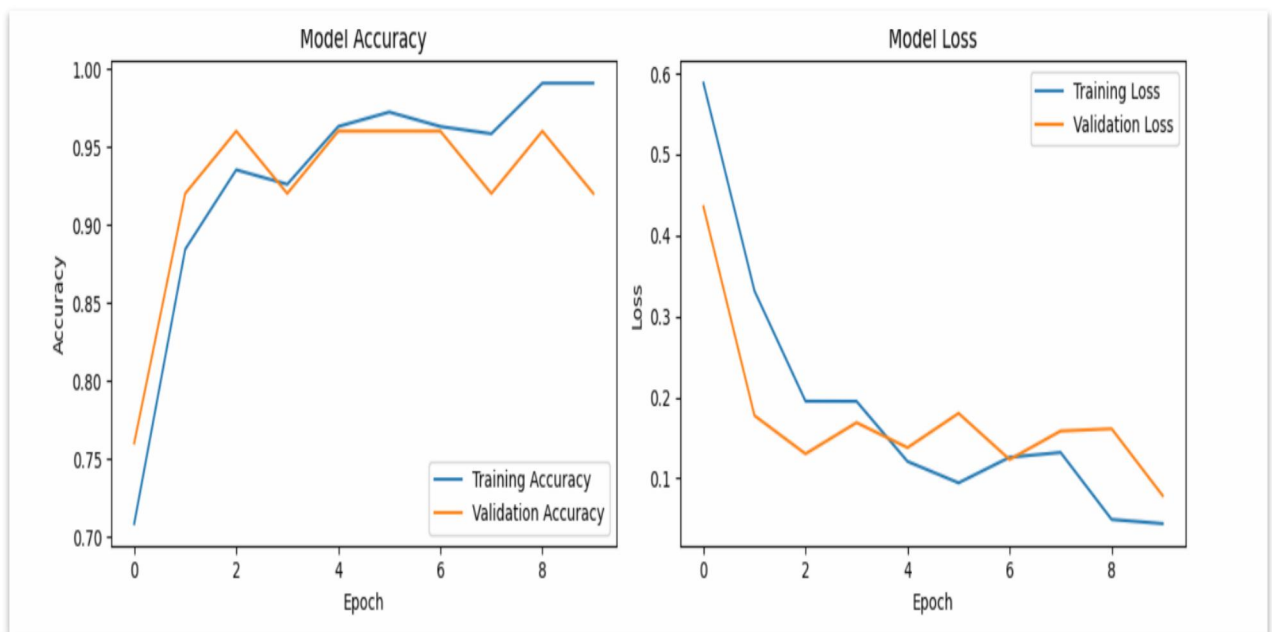


Figure 5- 8 Loss and accuracy curves during Training of LSTM

SVM model to detect DF audio over AOADS

A Support Vector Machine (SVM) DF audio detection model based on the AOADS dataset is given in this subject to compare with the suggested hybrid model for audio DF detection on the prepared dataset for Afaan Oromoo audio DF detection. SVM is trained with a dataset of real and synthetically generated audio samples, and the model's performance is evaluated based on significant testing metrics such as accuracy, loss, precision, recall, F1-score, and confusion matrix analysis. As shown in Figure 5-8 the model has perfect recall 1.0 of the real class, that means it detects all real instances without failing to detect any, and its precision and F1-score are close to perfect but slightly lower. Figure 5-8 shows the F1-score for fake is between these two measures, indicating a balance between precision and recall. But the overall performance of the SVM model is overall 90%.

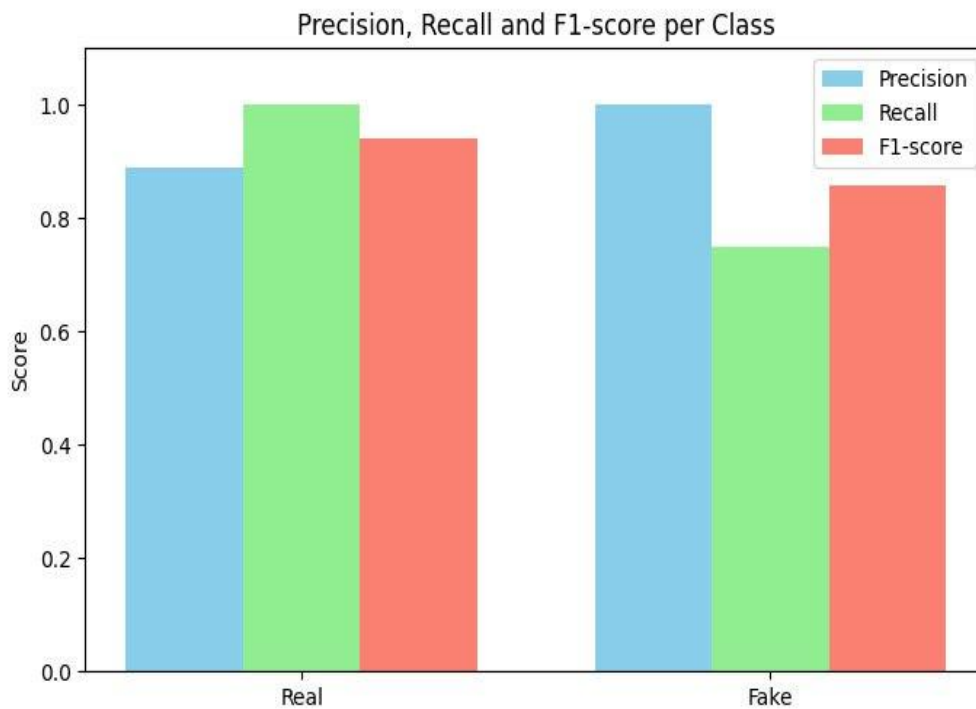


Figure 5- 9 Shows Classification Report on performance matrices of SVM model

The proposed hybrid CNN-LSTM model for AOADFD approach effectively fit in spatial feature extraction performed by CNN layers with temporal feature learning performed by LSTM layers, enabling it to successfully capture both spectral and sequential

dependencies present in audio data. The built CNN-LSTM model also had higher precision 0.98 for real and 0.92 for synthetic, recall 0.92 for real and 0.98 for synthetic, and F1-score (0.95 for both classes), whereas the CNN model had high precision for real audio (1.00) but low recall 0.56), which may lead to high misclassification of real audio as synthetic. Also, the hybrid CNN-LSTM built model based on AOADS had a reduced test loss of 0.25, reflecting larger overview over the test loss of 0.87 achieved by the CNN model. These results imply that the addition of serial modeling using with others like LSTMs greatly improves the model's capacity to distinguish between authentic and DF audio, so explanation CNN-LSTM is a superior method of DF detection in Afaan Oromoo.

Table 5- 3 Summary of Comparison different models over AOADS

Metrics	CNN-LSTM (hybrid)	CNN	SVM	LSTM
Test Accuracy	95%	78%	90%	90%
Stability	Good, low overfitting	High loss, unstable	Moderate	Good, stable
Overall	Best overall; balanced, robust	High precision, low recall for real	Good recall for real, lower for fake	Good, but less than hybrid

5.4 Discussion

The proposed Afaan Oromoo DF audio detection model, hybrid CNN-LSTM approach to identifying DFs in Afaan Oromoo is very effective in several points of testing. The carefully developed dataset consists of real audio obtained from open sources and synthetic audio clips created using RTVC guarantees linguistic inconsistency and natural differences between speakers. After preprocessing and MFCC and prosody feature extraction takes place the CNN module extracts spatial features from prosody and MFCC spectrograms, while LSTM layers gather temporal relationships, allowing the detection

of minute differences typical of synthetic audio. The training processes, optimized with Adam and monitored through a validation split, illustrate the model's ability to learn effectively 97.40% training accuracy and generalize to new samples 95% test accuracy. Performance evaluations validate the model's good classification performance with an overall accuracy of 95%, high precision and well-balanced F1-scores for both real and fake classes. Although a small amount of real audio was classified as false, none of the fake samples were classified as real, with excellent precision being evidence by the model. Furthermore, the loss and accuracy plots exhibit initial training fluctuations that converge to stability in subsequent epochs, indicating better generalization. The results show that the hybrid CNN-LSTM model, when combined with extensive data preprocessing and comparing it with CNN model, is highly effective at identifying DF audio in the Afaan Oromoo language.

Chapter 6. Conclusion and Future Works

6.1 Conclusion

In recent years, DF audio identification has gained more prominence as synthetic technology in audio progress. The development of DL approaches and the emergence of GAN has significantly encouraged the development of DF technology, making it possible to create very realistic synthetic media. These developments have, though, heightened concerns over the potential for the use of deepfake content in fraud, voice impersonation, and the spread of misinformation. Despite the increasing sophistication in DF creation, there is still a large gap in coming up with effective and robust detection methods for audio DFs, particularly for Afaan Oromoo language. Although Afaan Oromoo is widely spoken, there is limited specific research and customized detection systems designed to fit its phonetic and prosodic nature.

This research suggests and develops a hybrid CNN-LSTM model for DF audio identification in Afaan Oromoo, with limited research for effective DF detection. The carefully designed AOADS contains real audio from open sources and synthesized audio clips produced by RTVC to ensure linguistic diversity and real speaker variations. Preprocessing methods like audio file conversion to mono, resampling, segmentation and

prosody and MFCC feature extraction effectively capture the spectral and temporal characteristics necessary for DF detection.

The proposed model of Afaan Oromoo DF detection CNN block learns spatial features from MFCC spectrograms, and LSTM blocks gather temporal relationships, allowing the model to identify subtle differences characteristic of synthetic audio. The training, optimized using the Adam optimizer and monitored using a validation split, shows the model learning well with a training accuracy of 97.40%, and generalizing to new samples with a test accuracy of 95%. Even though there was a very low misclassification rate for real audio samples, none of the fake samples were misclassified as real, proving the model's high precision. Besides, the graphical representations of loss and accuracy both exhibit early fluctuations in the training process that later attain stabilization in the later epochs, indicating an improved capacity to generalize. The results demonstrate that the CNN-LSTM model, when coupled with accurate data preprocessing and meticulous fine-tuning, achieves remarkable efficacy in identifying DF audio in the Afaan Oromoo language.

6.2 Contribution of This Work

This study makes important contributions to the DF audio detection landscape, especially for Afaan Oromoo. We release here the very first model to detect DFs in audio, tuned to the Afaan Oromoo language. The proposed hybrid CNN-LSTM model leverages the good representation of spatial features by CNN layers and the representation of temporal dependencies by the LSTM layers in spoken language. This new architecture assists the capability of the model to practice addressing the complexities associated with detecting DF audio, particularly in languages characterized by a variety of dialects and accents.

6.3 Future Research Directions

Even though the suggested model has proven to have great results, there are numerous areas of improvement and further research within the subject area of DF audio detection of Afaan Oromoo:

- **More Widen Data Augmentation and More Variety Dataset**

Although it exists, it is very diverse, which makes it possible to augment with real-world records capturing other accents, dialects, and environmental sounds. Moreover, the dataset can be further augmented with additional fabricated audio samples produced by

more advanced models for synthetic speech, in such a way that it becomes more robust against more advanced DFs.

- **Advanced model architecture**

The existing hybrid model can be improved with the use of either attention mechanisms or transformers, which have been shown to increase the ability to focus on significant audio features. This would further refine the capability of identifying subtle differences in DF audio

References

- [1] S. H. Al-Khazraji, H. H. Saleh, A. I. Khalid, and I. A. Mishkhal,, " Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications,," *Eurasia Proc. Sci. Technol. Eng. Math*, vol. 23, 2023.
- [2] Ramdurai, Balagopal, and Prasanna Adhithya, ". The impact, advancements and applications of generative AI," *International Journal of Computer Science and Engineering*, vol. 10, pp. 1-8., 2023.
- [3] Gaur, Loveleen, Saurav Mallik, and Noor Zaman Jhanjhi, "Introduction to deepfake technologies," *CRC Press*, pp. 1-8, 2022.
- [4] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, " Generative artificial intelligence: a systematic review and applications,," *Multimed. Tools Appl*, Aug. 2024.
- [5] Tenove, Chris, and Heidi JS Tworek., "Online disinformation and harmful speech: Dangers for democratic participation and possible policy responses," *Journal of Parliamentary & Political Law*, vol. 13, pp. 215-232, 2019.
- [6] Mohammadi, Mehdi, Ala Al-Fuqaha, Sameh Sorour, and Mohsen Guizani, "Deep

- learning for IoT big data and streaming analytics: A survey," *IEEE Communications Surveys & Tutorials* 2, vol. 4, pp. 2923-2960, 2018.
- [7] Alom, Md Zahangir, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul AS Awwal, and Vijayan K. Asari, "A state-of-the-art survey on deep learning theory and architectures," *electronics*, vol. 8, no. 3, p. 292, 2019.
- [8] A. Shukla, " Leveraging AI and ML for Advance Cyber Security," *Journal of Artificial Intelligence & Cloud Computing* , vol. , 01 February 2024..
- [9] Almutairi, Zaynab, and Hebah Elgibreen, "A review of modern audio deepfake detection methods: challenges and future directions," *Algorithms* , vol. 15, no. 5, p. 155, 2022.
- [10] N. M. Müller, K. Pizzi, and J. Williams,s, " Human Perception of Audio Deepfake," *Association for Computing Machinery*, vol. 1, no. 1, 2019.
- [11] Hery, Andrew, Oluwaseyi Joseph, Olaoye Femi, and Hivez Luz, ".*Audio deepfakes: threats to voice assistants and voice-activated systems*, 2024.
- [12] S. K. Abdella and W. J. Sori, "Detection of Emotions in Afan Oromo Social Media Texts Using Deep Learning Method," *Ethiop. J. Sci. Sustain. De*, pp. 70-84, 2024.
- [13] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," *Algorithms*, vol. 15, no. 5, May 2022.
- [14] Dixit, Abhishek, Nirmal Kaur, and Staffy Kingra, "Review of audio deepfake detection techniques: Issues and prospects," *Expert Systems*, vol. 40, no. 8, 2023.
- [15] M. Lataifeh, A. Elnagar, I. Shahin, and A. B. Nassif, "Arabic audio clips: Identification and discrimination of authentic Cantillations from imitations," *Neurocomputing*, vol. 418, pp. 162-177, 2020.
- [16] . Gu, X. Yi, and X. Zhao, "FMFCC-A, Z. Zhang, Y., "A challenging Mandarin dataset for synthetic speech detection," *Proc. 20th Int. Digital Forensics and Watermarking*, pp. 117-131, 2021.
- [17] Li, Menglu, Yasaman Ahmadiadli, and Xiao-Ping Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, no. Cryptography and Security (cs.CR); Multimedia (cs.MM); Audio and Speech Processing (eess.AS), 2024.

- [18] G. D. Kaba, S. A. Olalekan, and G. Asra, "Afaan Oromo Language Fake News Detection in Social Media Using Convolutional Neural Network and Long Short Term Memor," *Journal of Electrical and Electronics Engineering*, vol. 5, 2022.
- [19] Gaurang Bansal, Aditya Nawal, Vinay Chamola, and Norbert Herencsar, "Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation," *Association for Computing Machinery*, vol. 11, pp. 1551-6857, 2024.
- [20] Pawelec, Maria, Deepfakes: "Manipulation on Demand? Evolution of Deepfake Technology, Societal Impact, and the Path Forward", 2025.
- [21] S. Karnouskos, "Artificial Intelligence in Digital Media: The Era of Deepfakes," *EEE Trans. Technol. Soc*, vol. 1, pp. 138-147, 2020.
- [22] I. Sarker, " Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," <https://doi.org/10.1007/s42979-021-00815-1>, vol. 2, p. 420, 2021.
- [23] Schafer, Karla and Choi, Jeong-Eun and Zmudzinski, Sascha, "Explore the World of Audio Deepfakes: A Guide to Detection Techniques for Non-Experts," in *Association for Computing Machinery*, New York, NY, USA, 2024.
- [24] Hamza, Ameer and Javed, Abdul Rehman and Iqbal, Farkhud and Kryvinska, Natalia and Almadhor, Ahmad and Jalil, Zunera and Borghol, Rouba, "Deepfake Audio Detection via MFCC Features Using Machine Learning," *IEEE Access*, pp. 1-1, 2022.
- [25] Dr. A.Shaji George and A. S.Hovan George, "Deepfakes: The Evolution of Hyper realistic Media Manipulation," *doi: 10.5281/ZENOD*, Nov. 2023.
- [26] Bai, Yu, Yu Zhao, Yajing Shao, Xinrong Zhang, and Xuefeng Yuan, "Deep learning in different remote sensing image categories and applications: status and prospects," *International Journal of Remote Sensing*, vol. 43, no. 5, pp. 1800-1847, 2022.
- [27] Thies, Justus, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. "Face2face: Real-time face capture and reenactment of rgb videos."n, pp. , pp. 2387-2395, 2016.
- [28] Bengesi, S., El-Sayed, H., Sarker, M.K., Houkpati, Y., Irungu, J. and Oladunni, T, "Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformer," *IEEe Access.*, 2024.

- [29] G. Genelza, " A systematic literature review on AI voice cloning generator: A game-changer or a threat?., " *Journal of Emerging Technologies*,, vol. 4, no. 2, pp. 54-61, 2024.
- [30] Masood, Momina, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, vol. 53, no. 4, pp. 3974-4026, 2023.
- [31] Yao, Xinwei, Ohad Fried, Kayvon Fatahalian, and Maneesh Agrawala, "Iterative text-based editing of talking-heads using neural retargeting," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 3, pp. 1-14, 2021.
- [32] Rehaan, M., Kaur, N. and Kingra, S, " Face manipulated deepfake generation and recognition approaches: A survey," *Smart Science*, vol. 12, no. 1, pp. 53-73, 2024.
- [33] Lahiri, A., Kwatra, V., Frueh, C., Lewis, J. and Bregler, C, "Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," *In Proceedings of the IEEE/CVF conference on computer vision and patt*, 2021.
- [34] Masood, Momina and Nawaz, Marriam and Malik, Khalid and Javed, Ali and Irtaza, Aun and Malik, Hafiz, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, pp. 1-53, 2022.
- [35] Khan, R., Taqi, M. and Afzal, A., " Deepfakes in Finance: Unraveling the Threat Landscape and Detection Challenge," *IGI Global*, pp. 91-120, 2024.
- [36] Croitoru, F.A., Hiji, A.I., Hondru, V., Ristea, N.C., Irofti, P., Popescu, M., Rusu, C., Ionescu, R.T., Khan, F.S. and Shah, M, "Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook," *arXiv: arXiv:2411.19537*, no. F.-A
- [37] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S. and Xia, F, "Deepfake video detection: challenges and opportunities," *Artificial Intelligence Review*, vol. 57, no. 6, p. 159, 2024.
- [38] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C. and Tao, D, " Deepfake generation and detection: A benchmark and survey," 2024.
- [39] Javed, M., Zhang, Z., Dahri, F.H. and Laghari, A.A, " Real-time deepfake video

- detection using eye movement analysis with a hybrid deep learning approach," *Electronics*, p. 2947, 2024.
- [40] F. T. Megersa, "Hierarchical Afaan Oromoo news text classification," *New Media and Mass Communication*, vol. 88, pp. 1-11, 2020.
- [41] Rundasa, & Diriba, Gichile & Gichile Rundasa, Diriba "Text to Speech Synthesizer for Afaan Oromo Using Deep Neural Networ," vol 22 ,p 1947-1958, 2025.
- [42] N. NEGESE, "Speech-based Question and Question Answering Classification for Afaan Oromo Language," *.Diss. St. Mary's University,*, 2023.
- [43] Chala, Tamrat Delessa, Ashenafi Chalchisa Guta, and Muluken Hussen Asebel., "Design and development of a text-to-speech synthesizer for Afan Oromo," *SN Computer Science*, vol. 3, 2022.
- [44] F. Negesse, "An acoustic analysis of Oromo vowels of the northern dialect," *Journal of the International Phonetic Association* , vol. 53, pp. 1033-1048, 2023.
- [45] A. Arvaniti, "The Phonetics of Prosody," *Oxford Research Encyclopedia of Linguistics*, 2020.
- [46] Beckman, Mary E and Venditti, Jennifer , "Intonation," *The handbook of phonological theory*, pp. 485--532, 2011.
- [47] Warren, Kevin and Olszewski, Daniel and Layton, Seth and Butler, Kevin and Gates, Carrie and Traynor, Patrick, "Pitch Imperfect: Detecting Audio Deepfakes Through Acoustic Prosodic Analysis," *arXiv preprint arXiv:2502.14726*, 2025.
- [48] Anagaw, A, "Lexical and Phonological Variation among Affan Oromo Language Speakers of Kemisie Society," *Abyssinia Journal of Business and Social Sciences*, pp. 17-24, 2017.
- [49] O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," *IEEE Access*, vol. 11, p. 132652–132682, 2023.
- [0] Teshite, Kebede, Getachew Mamo, and Kris Calpotura, "Research Article Afaan Oromo Speech-Based Computer Command and Control: An Evaluation with Selected Commands," *Journal:Advances in Human-Computer Interaction*, vol.2023,p. 9959015, 2023.
- [51] Negese, Tariku, "Phoneme Distribution and the Sonority Nature of Consonant Clusters in Afaan Oromo," *International Journal online of Humanities*vol. 7, p. 18,

- 2021.
- [52] Naveen, Palanichamy, and Pavel Trojovský, "Overview and challenges of machine translation for contextually appropriate translation," *Isience* , vol. 27, 2024.
 - [53] Kathiresan, Thayabaran, "Understanding the Effects of Intrinsic Speaker Variability to Enhance ASR Performance," University of Zurich PhD diss, 2019.
 - [54] Zhang, Bowen and Cui, Hui and Nguyen, Van and Whitty, Monica, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, p. 1989, 2025.
 - [55] F. Putri, " Study on noise suppression based on the spectro-temporal modulation," 2023.
 - [56] Das, S., Dutta, A., Roy, K., Mondal, A. and Mukhopadhyay, A, " A Survey on Automatic Online Hate Speech Detection in Low-Resource Languages," *arXiv preprint arXiv:241*, 2024.
 - [57] Z. Byambadorj, R. Nishimura, A. Ayush, K. Ohta, and N. Kitaoka, "Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation," *EURASIP J. Audio Speech Music Proces*, vol. 2021, p. 42, 2021.
 - [58] A. Przewdzia, "Optimizing Speech Recognition for Low-Resource Languages," *Northern Sotho*, 2024.
 - [59] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," *Electronics*, vol. 11, p. 3795.
 - [60] Sunil, Reshma, Parita Mer, Anjali Diwan, Rajesh Mahadeva, and Anuj Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," *Heliyon*, vol. 11, num. 3, 2025.
 - [61] "M- AILABS, "TheM- AILABS Speech Dataset," 2018".
 - [62] "Baidu Silicon Valley AI Lab, "Cloned Audio Dataset," 2018.".
 - [63] Ballesteros, D.M., Rodriguez, Y. and Renza, "A dataset of histograms of original and fake voice recordings (H-Voice)," vol. 29, no. Data in brief, p. 105331, 2020.
 - [64] Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A. and Lee, K.A, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," no. IEEE/ACM Transactions on

- Audio, pp. 2507-2522, 2023.
- [65] Khalid, Hasam and Tariq, Shahroz and Kim, Minha and Woo, Simon S, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.
- [66] Li, Menglu, Yasaman Ahmadiadli, and Xiao-Ping Zhang, "A Survey on Speech Deepfake Detection," *ACM Computing Surveys*, vol. 57, p. 1-38, 2025.
- [67] Munir, Sheza, Wassay Sajjad, Mukeet Raza, Emaan Abbas, Abdul Hameed Azeem, "Deepfake defense: Constructing and evaluating a specialized Urdu deepfake audio dataset," vol. 202, no. In Findings of the Association for Computational Li, pp. . 14470-14480, 2024.
- [68] Flórez, PA Tamayo, Rubén Manrique, and B. Pereira Nunes, "HABLA: A dataset of Latin American Spanish accents for voice anti-spoofing," vol. 2023, no. In Proc. INTERSPEECH, pp. 1963-1967, 2023.
- [69] Müller, Nicolas M., Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger, "Mlaad: The multi-language audio anti-spoofing dataset.," in *International Joint Conference on Neural Networks*, 2024.
- [70] Asif, Amna, Hamid Mukhtar, Fatimah Alqadheeb, Hafiz Farooq Ahmad, and Abdulaziz Alhumam, "An approach for pronunciation classification of classical Arabic phonemes using deep learning," *Applied Sciences* , vol. 1, p. 238, 2021.
- [71] Ma, H., Yi, J., Wang, C., Yan, X., Tao, J., Wang, T., Wang, S. and Fu, R, "CFAD: A Chinese dataset for fake audio detection. Speech Communication," vol. 164, p. 103122, 2024.
- [72] Zhang, Bowen and Cui, Hui and Nguyen, Van and Whitty, Monica, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, p. 1989, 2025.
- [73] T. Becher and S. Torka,, "Exploring AI-Enabled Cybersecurity Frameworks: Deep-Learning Techniques, GPU Support, and Future Enhancement," *arXiv: arXiv:2412.12648*, 2024.
- [74] N. M. Müller, K. Pizzi, and J. William, "Human Perception of Audio," *Association for Computing Machinery*, vol. 1, no. 1, 2019.

- [75] Tune, Kula Kekeba., "Development of Cross-Language Information Retrieval for Resource-Scarce African Language," *International Institute of Information Technology*, pp. 80-92., 2015.
- [76] G. D. Kaba, S. A. Olalekan, and G. Asra, "Afaan Oromo Language Fake News Detection in Social Media Using Convolutional Neural Network and Long Short Term Memo," *Journal of Electrical and Electronics Engineerin*, vol. 5, 2022.
- [77] Sajeeda, A. and Hossain, B.M, "Exploring generative adversarial networks and adversarial training," *International Journal of Cognitive Computing in Engineering*, vol. 3, pp. 78-89, 2022.
- [78] Khan, R., Taqi, M. and Afzal, A, "Deepfakes in Finance: Unraveling the Threat Landscape and Detection Challenges. In Navigating the World of Deepfake Technology," *In Navigating the World of Deepfake Technology*, pp. 91-120, 2024.
- [79] Z. Khanjani, G. Watson, and V. P. Janeja, "Audio deepfakes: A survey," *Front. Big Dat*, vol. 5, p. 1001063, 2023.
- [80] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Syst. Appl*, vol. 252, p. 124260, 2024.
- [81] Sunil, R., Mer, P., Diwan, A., Mahadeva, R. and Sharma, A, " Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," *Heliyon*, 2025.
- [82] McGettigan, C., Bloch, S., Bowles, C., Dinkar, T., Lavan, N., Reus, J. and Rosi, "Voice cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities.," 2024.
- [83] Trujillo-Guerrero, M.F., Román-Niemes, S., Jaén-Vargas, M., Cadiz, A., Fonseca, R. and Serrano-Olmedo, J.J, "Accuracy comparison of CNN, LSTM, and transformer for activity recognition using IMU and visual markers," *IEEE Access*, vol. 11, pp. 106650-106669, 2023.
- [84] Alomar, K., Aysel, H.I. and Cai, X, "RNNs, CNNs and Transformers in Human Action Recognition: A Survey and A Hybrid Model," *arXiv preprint arXiv:2407.06162*, no. Alomar, K., Aysel, H.I. and Cai, X., 2024. RNNs, CNNs and Transformers in Human Action Recognition: A Survey and A Hybrid Model. arXiv preprint arXiv:2407.06162., 2024.

- [85] Elpeltagy, Marwa, Aya Ismail, Mervat S. Zaki, and Kamal Eldahshan, "A novel smart deepfake video detection system.," *International Journal of Advanced Computer Science and Applications*, 2023.
- [86] Hashmi, A., Shahzad, S.A., Lin, C.W., Tsao, Y. and Wang, H.M, "Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights," *arXiv preprint arXiv:2411.07650*., 2024.
- [87] Babirye, C., Nakatumba-Nabende, J., Katumba, A., Ogwang, R., Francis, J.T., Mukiibi, J., Ssentanda, M., Wanzare, L.D. and David, D, "Building text and speech datasets for low resourced languages: A case of languages in East Africa," 2022.
- [88] W. Tegegne, "The development of written Afan Oromo and the appropriateness of Qubee, Latin Script, for Afan Oromo writing," *Historical Research Letter*, vol. 28, 2016.
- [89] Kumar, Y. and Singh, N., "A comprehensive view of automatic speech recognition system-a systematic literature review," *EEE*, no. international conference on automation, computational and technology management (ICACTM), pp. 168-173, 2019.
- [90] G. O. Ganfure, "Comparative analysis of deep learning based Afaan Oromo hate speech detectio," *Journal of Big Data*, p. 76, 2022.
- [91] Almutairi Z, Elgibreen H, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," *Algorithms* , vol. 5, p. 155, 2022.
- [92] Schafer, Karla and Choi, Jeong-Eun and Zmudzinski, Sascha, "Explore the World of Audio Deepfakes: A Guide to Detection Techniques for Non-Experts," in *Association for Computing Machinery*, New York, NY, USA, 2024.

Appendixes

Appendix A: the python code to build a model

```
import tensorflow as tf

from tensorflow.keras import layers, models, regularizers

from tensorflow.keras.callbacks import EarlyStopping, ModelCheckpoint


# -----

# 3. Build the CNN-LSTM Model (ADFD)

# -----

def create_model(input_shape):
    """
    Create and return a CNN-LSTM model.

    The model is named ADFD (Audio DeepFake Detection).
    """

    model = models.Sequential(name="ADFD")

    # Convolutional Block 1

    model.add(layers.Conv2D(32, (3, 3), activation='relu', input_shape=input_shape))

    model.add(layers.MaxPooling2D((2, 2)))

    model.add(layers.BatchNormalization())

    # Convolutional Block 2

    model.add(layers.Conv2D(64, (3, 3), activation='relu'))

    model.add(layers.MaxPooling2D((2, 2)))

    model.add(layers.BatchNormalization())
```

```

# Convolutional Block 3

model.add(layers.Conv2D(128, (3, 3), activation='relu'))

model.add(layers.MaxPooling2D((2, 2)))

model.add(layers.BatchNormalization())

# Determine output shape before reshaping

model.add(layers.Flatten()) # Flatten CNN output for LSTM compatibility

# LSTM layers for temporal processing

model.add(layers.Reshape((-1, 128))) # Ensure correct shape for LSTM layers

model.add(layers.LSTM(64, return_sequences=True))

model.add(layers.LSTM(32))

# Fully Connected (Dense) layers

model.add(layers.Dense(64, activation='relu', kernel_regularizer=regularizers.l2(0.01)))

model.add(layers.Dropout(0.5))

model.add(layers.Dense(1, activation='sigmoid')) # Binary classification (DeepFake vs.
Real)

return model

# Load dataset

# X_train is a 4D array (num_samples, time_steps, features, channels)

# X_train.shape -> (num_samples, 13420, 1)

# X_train is already preprocessed

input_shape = (X_train.shape[1], X_train.shape[2], X_train.shape[3]) # (30, 128, 1)

adfd = create_model(input_shape)

# Compile the model

adfd.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

```

```

# Define callbacks

callbacks = [

    EarlyStopping(monitor='val_loss', patience=2, restore_best_weights=True),

    ModelCheckpoint(filepath="ADFD_model.keras", monitor='val_loss',
save_best_only=True)

]


# Train the model

history = adfd.fit(X_train, y_train,

    epochs=20,

    batch_size=32,

    validation_split=0.1,

    callbacks=callbacks)


# Save the trained model

adfd.save("ADFD_model_final.keras")

```

Appendix B: RTVC Pipeline for Afaan Oromoo python code

```
from synthesizer.inference import Synthesizer

from encoder import inference as encoder

from vocoder import inference as vocoder


# Load the fine-tuned Afaan Oromoo models

encoder.load_model('encoder/saved_models/pretrained.pt')

synthesizer = Synthesizer('synthesizer/saved_models/afaan_omoo/pretrained/pretrained.p
t')

vocoder.load_model('vocoder/saved_models/afaan_omoo/pretrained/pretrained.
pt')

# Reference audio (for speaker embedding)

ref_wav = 'audio_preprocessed/Ab1.wav'

preprocessed_wav = encoder.preprocess_wav(ref_wav)

embed = encoder.embed_utterance(preprocessed_wav)


# Synthesize

text = "Galatoomi! Afaan Oromoo warren Afaan Oromoo dubbachuu barbaaddan
hundi keessaniyyuu."

specs = synthesizer.synthesize_spectrograms([text], [embed])

generated_wav = vocoder.infer_waveform(specs[0])

# Save output

Import soundfile as sf

sf.write("output_afaan_omoo.wav", generated_wav)
```

Declaration

I, declare that this thesis is my original work and has not been presented for a degree in any other university. All materials utilized for the thesis have been properly acknowledged.

Declared by: Name: _____

Signature: _____

Date: _____

Confirmed Advisor:

Name_____

Signature: _____

Date: _____