People's Democratic Republic OF Algeria

Ministry of Higher Education and Scientific Research

Echahid Cheikh Larbi Tebessi University TEBESSA

Faculty of Exact Sciences and Natural and Life Sciences

Department of Computer Science

Thesis of Master

Presented for obtaining the degree of **MASTER**

**In :** Computer Science

**Speciality : Networks and IT Security**

**By :** BOUSBA Abdelsamie

**Theme**

# AI-Based Online API for Fake Speech Detection

Publicly defended, in front of the jury composed of :

| | | | | |
|---|---|---|---|---|
| Mrs. | NGHREIB Nawel | MCB | Larbi Tebessi University | President |
| Mr. | AOUINE Mohamed | MAA | Larbi Tebessi University | Examiner |
| Mr. | BOUALLEG Yaakoub | MCB | Larbi Tebessi University | Supervisor |
| Mr. | DAOUADI Kheir eddine | MCB | Larbi Tebessi University | Co-Supervisor |

Academic Year: **2023/2024**

*Dedicated to my parents.*

# Acknowledgement

I would like to express my deepest gratitude to my supervisor, **Dr. Yaakoub Bouallег** and Co-supervisor **Dr. Kheir eddine Daouadi** for their invaluable guidance and support throughout this research. I am also thankful to the jury members, **Dr. Ghreib Nawel** and **Dr. Aouine Mohamed** for accepting to evaluate my work and for their constructive feedback and recommendations.

# Abstract

Deepfake audio technology poses a growing threat to information authenticity and integrity. This thesis provides a systematic investigation of different Machine Learning (ML) methods for detecting deepfake in Arabic speech. Firstly, a novel dataset of real and synthetic Arabic audio speech was created. Then, various ML methods were evaluated for their ability to discriminate between genuine and synthesized speech. Finally, a new Arabic deepfake speech framework is proposed, including handcrafted feature extraction and classification. Feature importance analysis revealed key acoustic and prosodic cues that contribute to the detection process, where the XGBoost classifier emerged as the most effective. Experimental results demonstrated the robustness and the high accuracy of our proposed framework for Arabic deepfake speech detection compared to state-of-the-art methods. This research establishes a benchmark for Arabic deepfake audio detection and contributes to the ongoing efforts to combat the harmful effects of this technology.

**Keywords:** Deepfake Audio, Arabic Speech, Ensemble Learning, Machine Learning, Deep Learning, Generative Artificial Intelligence.

## ملخص

تشكل تقنية الصوت المزيف تهديدًا متزايدًا لمصداقية وسلامة المعلومات. تستكشف هذه الأطروحة استخدام تقنيات تعلم الالة للكشف عن الكلام العربي المزيف. تم إنشاء مجموعة بيانات جديدة من عينات الصوت العربية الحقيقية والمصطنعة، وتم تقييم نماذج تعلم الالة المختلفة لقدرتها على التمييز بين الكلام الأصلي والمولد عن طريق الذكاء الإصطناعي. برز النموذج المختار كأكثر النماذج فعالية، حيث أظهر دقة عالية ومتانة ضد تقنيات توليد الصوت المزيف المختلفة. كشف تحليل أهمية الميزات عن إشارات صوتية ولغوية رئيسية تساهم في عملية الكشف. يضع هذا البحث معيارًا للكشف عن الصوت العربي المزيف ويساهم في الجهود المستمرة لمكافحة الآثار الضارة لهذه التقنية.

**الكلمات المفتاحية**: الصوت المزيف، الكلام العربي، التعلم الآلي، التعلم العميق، الذكاء الاصطناعي التوليدي.

## Résumé

La technologie de l'audio deepfake représente une menace croissante pour l'authenticité et l'intégrité de l'information. Cette thèse fournit une investigation systématique de différentes méthodes d'apprentissage automatique (ML) pour détecter les deepfakes dans les discours arabes. Tout d'abord, un nouveau jeu de données d'audio arabe réel et synthétique a été créé. Ensuite, diverses méthodes de ML ont été évaluées pour leur capacité à discriminer entre les discours authentiques et synthétisés. Enfin, un nouveau cadre de discours deepfake arabe est proposé, incluant l'extraction de caractéristiques manuelles et la classification des caractéristiques. L'analyse de l'importance des caractéristiques a révélé des indices acoustiques et prosodiques clés qui contribuent au processus de détection, où le classificateur XGBoost s'est avéré le plus efficace. Les résultats expérimentaux ont démontré la grande précision et la robustesse de notre cadre proposé pour la détection des discours deepfake en arabe, comparé aux méthodes de pointe. Cette recherche établit une référence pour la détection de l'audio deepfake en arabe et contribue à l'effort continu pour combattre les effets néfastes de cette technologie.

**Mots-clés :** Audio Deepfake, Discours Arabe, Apprentissage Ensemble, Apprentissage Automatique, Apprentissage Profond, Intelligence Artificielle Générative.

# Contents

**Table des figures** i

**Liste des tableaux** ii

**1 General Introduction to Deepfake Audio Detection** 1

1.1 Context description . . . 1

1.1.1 DeepFake Definition . . . 1

1.1.2 DeepFake Types . . . 2

1.1.3 DeepFake Audio Types . . . 3

1.1.4 The Rise of DeepFake . . . 4

1.1.5 Deepfakes as a Cybersecurity Threat . . . 5

1.1.6 Deepfake Fraud Examples . . . 6

1.2 Problem Statement . . . 9

1.3 Motivation . . . 10

1.4 Objectives . . . 10

1.5 Outline . . . 11

**2 Artificial Intelligence Background** 13

2.1 Definition of Artificial Intelligence . . . 13

2.1.1 Types of Artificial Intelligence . . . 14

2.1.2 Subfields of Artificial Intelligence . . . 14

2.2 Machine Learning . . . 14

2.2.1 Supervised learning . . . 15

2.2.2 Unsupervised learning . . . 17

2.2.3 Semi-supervised learning . . . 18

2.2.4 Reinforcement machine learning . . . 18

2.3 Deep Learning . . . 19

2.4 Generative AI: The Engine Behind Deepfakes . . . 23

2.4.1 Applications of Generative Artificial Intelligence . . . 26

2.4.2 Benefits of generative AI . . . . . . . . . . . . . . . . . . . . . . . . . 28
2.4.3 Challenges, limitations and risks . . . . . . . . . . . . . . . . . . . . 29

**3 Audio Fundamentals 31**
3.1 Properties of Sound . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
3.2 Digital Audio Basics . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
3.2.1 Analog vs. Digital Audio . . . . . . . . . . . . . . . . . . . . . . . 33
3.2.2 Sampling Rate and Bit Depth . . . . . . . . . . . . . . . . . . . . . 33
3.2.3 Audio File Formats . . . . . . . . . . . . . . . . . . . . . . . . . . 34
3.3 Audio Signal Processing . . . . . . . . . . . . . . . . . . . . . . . . . . 34
3.3.1 Fundamental Concepts . . . . . . . . . . . . . . . . . . . . . . . . 34
3.3.2 Filtering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
3.3.3 Feature Extraction . . . . . . . . . . . . . . . . . . . . . . . . . . 37
3.4 Speech Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
3.4.1 Phonetics and Phonology . . . . . . . . . . . . . . . . . . . . . . . 39
3.4.2 Speech Recognition . . . . . . . . . . . . . . . . . . . . . . . . . . 39
3.4.3 Text-to-Speech (TTS) Synthesis . . . . . . . . . . . . . . . . . . . 40
3.4.4 Voice Biometrics . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
3.5 Challenges in Audio Analysis . . . . . . . . . . . . . . . . . . . . . . . 42

**4 State of the Art in Deepfake Audio Detection 45**
4.1 Traditional Audio Forensics Techniques . . . . . . . . . . . . . . . . . 45
4.2 Current State-of-the-Art Detection Methods . . . . . . . . . . . . . . . 47
4.2.1 Machine Learning-Based Methods . . . . . . . . . . . . . . . . . . 47
4.2.2 Deep Learning-Based Methods . . . . . . . . . . . . . . . . . . . . 47

**5 Contributions and Results 51**
5.1 Proposed Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . 51
5.1.1 Contribution 1: Dataset Creation - Arabic Audio Deepfake . . . . . 51
5.1.2 Contribution 2: Proposed Framework For Deepfake Audio Detection 55
5.2 Experimental study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 57
5.2.1 Training and Evaluation . . . . . . . . . . . . . . . . . . . . . . . . 57
5.3 Experimental Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
5.3.1 Experimental Setup . . . . . . . . . . . . . . . . . . . . . . . . . . 59
5.3.2 Performance of Different Models . . . . . . . . . . . . . . . . . . . 60

**General Conclusion** 63

**Bibliographie** 66

# Table des figures

2.1 Machine learning Types . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
2.2 ANN architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
2.3 CNN architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
2.4 RNN architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
2.5 LSTM architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
2.6 Transformers architecture [2]. . . . . . . . . . . . . . . . . . . . . . . . 24

5.1 Proposed Framework Architecture For Deepfake Audio Detection. . . . . . 55

# Liste des tableaux

4.1 Comparison of State-of-the-Art Methods for Detecting Synthetic Speech . . 49

5.1 List of speakers, their professions, and nationalities. . . . . . . . . . . . . . 52

5.2 Dataset Distribution . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

5.3 Overall Performance of Different Models (5-fold Cross-Validation) . . . . . 61

# Chapitre 1

# General Introduction to Deepfake Audio Detection

## Introduction

Artificial Intelligence (AI) has revolutionized the world. However, this revolution has also opened doors for manipulation. Deepfakes, a type of synthetic media, have emerged as a growing concern. These AI-powered technologies can create highly realistic audio or video recordings that manipulate the appearance or voice of a person. This chapter explores the many forms of deepfakes and the difficulties they provide with some examples and history and addresses the threat they pose. Next, we'll focus on deepfake audio recognition, identify the issue of distinguishing natural speech from artificial voice, and finally present the thesis statement.

## 1.1 Context description

This section introduces deepfakes, exploring their definition, types, and the development in their creation. We will examine the cybersecurity threats posed by deepfakes and highlight real-world fraud examples. This context sets the stage for our exploration of deepfake audio detection techniques.

### 1.1.1 DeepFake Definition

A new technology that makes an unsettling impression has emerged in the rapidly changing digital landscape : deepfake technology. The term "deepfake" – a combination of "deep learning" and "fake" – refers to synthetic media where images, videos, or audio clips are manipulated using advanced AI-based tools. These alterations create highly realistic yet entirely generated representations that challenge our perception of reality. From their inception to their current state, deepfakes have rapidly evolved, becoming increasingly sophisticated and easily accessible to the general public [1].

### 1.1.2 DeepFake Types

Many varieties exist in DeepFake. The following are a few of the well-known ones :

- **Textual DeepFake :** In the early days of machine learning and Natural Language Processing (NLP), the idea of a machine tackling creative efforts like writing seemed like science fiction. Fast forward to 2017, and the landscape has dramatically shifted by the introduce of Transformers like (Generative Pre-trained Transformer) GPT [2]. Decades of tireless work by researchers, data scientists, and countless contributors have resulted in powerful language models and libraries. These advancements have paved the way for AI-generated writing that rivals human quality in terms of conciseness and clarity.

  This evolution highlights the remarkable capabilities of AI in domains once considered exclusively human. However, the development of AI writing tools also raises important questions. We need to consider the potential impact of these tools on authorship, plagiarism, and the very nature of creative expression in the digital age [1].

- **Video and Image Deepfake :** For deepfake creators with malicious intent, the ability to fabricate realistic videos and photographs is their primary weapon. In our current social media-driven world, where visuals reign supreme, deepfake videos are particularly dangerous. Unlike text, videos and photos have the power to capture attention, illustrate stories, and shape narratives in a way that text simply cannot.

  The threat posed by deepfake videos may even surpass that of manipulated text, given the current capabilities of AI in video generation. Advancements in AI have made video manipulation more sophisticated and potentially more harmful than natural language manipulation. One example is MarioNETte [3], a program developed by the Seoul-based software company Hyperconnect in 2021. This program exemplifies the power and potential misuse of deepfake technology. MarioNETte allows users to create deepfake videos of historical figures, celebrities, and political leaders. The program works by having another person mimic the facial expressions of the target individual, which are then seamlessly integrated into a deepfake body. As for image deepfake we have the DALL-E [4] an AI system that can produce realistic artwork and images from a natural language description.

This ability to create highly realistic deepfake videos or images of well-known personalities underscores the significant risks associated with this technology. Deepfake can be used to spread misinformation, damage reputations, and sow discord in society [1].

- **Audio Deepfake :** The world of deepfakes extends far beyond manipulated videos and photos. AI have also unlocked the ability to clone a human voice with surprising accuracy. To achieve this, deepfakes leverage a data repository containing audio recordings of the target individual. These algorithms can then analyze and learn from this data, meticulously replicating the person's unique vocal characteristics, including cadence, intonation, and even accent.

  The creation of deepfake audio has become even more accessible with the release of commercial programs like Lyrebird, Deep Voice, and Elevenlabs. These programs demonstrate the growing accessibility of this technology. While initial recordings are required to train the AI, subsequent interactions become surprisingly efficient. With just a few additional phrases, the AI can become adept at mimicking your voice and accent with impressive fidelity. As you provide more recordings, the deepfake program strengthens its ability to convincingly reproduce your voice, allowing it to narrate text in your tone and style.

  This ease of use and impressive level of realism highlight the growing concerns surrounding deepfake audio. The potential for misuse in areas like identity theft, impersonation scams, and the spread of misinformation is significant, making the development of robust detection methods a pressing priority [1].

### 1.1.3 DeepFake Audio Types

While deepfakes involving videos frequently make headlines, audio manipulation poses a serious and distinct risk. Deepfake audio, sometimes referred to as synthetic audio or voice cloning, uses machine learning to produce lifelike audio forgeries. Since audio deepfakes target the human auditory system, they may be more difficult to detect than video deepfakes, which rely on visual manipulation. We'll examine the various varieties of deepfake audio in this section :

- **Text-to-Speech (TTS) :** This technique involves generating entirely new speech from scratch based on a text script. Advanced AI models can mimic the voice characteristics (pitch, timbre, accent) of a target person to create realistic synthetic speech.

- **Voice Conversion :** This method focuses on altering a person's already-existing audio recordings. Deep learning algorithms have the ability to alter voice content while maintaining the identity of the original speaker. This makes it possible for someone to pretend to say things they never said in a forgery.
- **Emotion Fake :** This method seeks to change a speaker's emotional intonation. In spite of the fact that the original recording may have shown a different emotion, AI models are capable of altering audio to make someone sound happy, sad, or furious. This can drastically alter a message's impact and meaning.
- **Scene Fake :** The term describes the process of adjusting the speech's surrounding sound. Deepfake, for instance, could be used to mask edits or splicing by adding background noise or making it appear as though someone is giving a speech in a different setting (such as a political rally rather than a calm room).

### 1.1.4 The Rise of DeepFake

Deepfake first gained significant attention in the late 2010s, capturing the public's imagination and concern. Initially, they were mostly limited to entertainment and friendly jokes. However, it wasn't long before their potential for harm became evident. Today, technology underscores a critical direction in the digital era : the thinning line between truth and fiction.

The creation of deepfake is rooted in deep learning, a subset of AI that mimics the neural networks of the human brain. By ingesting vast amounts of data – images, video clips, or voice recordings – these algorithms learn to recreate and alter human likenesses with startling accuracy. This technology has progressed rapidly, thanks to advancements in AI and the increasing availability of data and computational power [5].

- **The Evolution of Deepfakes :** While the concept of AI-powered image manipulation existed as early as the 1990s within academic circles, deepfakes entered the public eye in the mid-2010s. The arrival of powerful neural networks and Generative Adversarial Networks (GANs) in 2014 marked a turning point. These GANs, pioneered by Ian Goodfellow, laid the groundwork for deepfake by enabling sophisticated and realistic manipulations.

  Initially, creating deepfakes was a technical difficulty, requiring significant processing power and expertise. This limited their use to researchers and tech-savvy individuals.

Early deepfakes, often fell short of the convincing realism seen today. However, technological advancements have democratized deepfake creation. Open-source projects and user-friendly applications have emerged, making it easier for anyone to generate believable deepfakes. This accessibility has fueled a significant rise in the production and online distribution of deepfake content.

- **Crossing the Line from Believable to Indistinguishable :** Deepfake is no longer just convincing, reaching a level where they are often indistinguishable from reality. This dramatic jump in realism stems from advancements in AI algorithms, the rise in computational power, and the plenty of data for training these models. Systems using Deep Learning (DL) to generate deepfakes have gotten incredibly good at recognizing and reproducing human facial expressions, facial features, and even speech patterns.

  One particularly concerning development is the ability to clone someone's audio. This opens the door to live deepfake, where someone can appear as another person during phone calls or streams. The potential for misuse in such a scenario extends far beyond entertainment, raising significant concerns in areas like politics or security.

  The social media landscape further fuels the rise of deepfakes. These platforms, where audio can be easily shared and viewed by millions, provide the perfect breeding ground for deepfakes to spread. This widespread dissemination, combined with our natural tendency to trust what we hear, makes deepfakes a powerful tool for malicious actors to spread misinformation.

### 1.1.5 Deepfakes as a Cybersecurity Threat

Deepfakes have opened a new, unsettling chapter in cybercrime. The very technology designed to entertain can now trick and manipulate with malicious intent. Deepfakes have become a powerful tool for identity theft and social engineering scams, exploiting our trust in familiar faces and voices. Imagine a shockingly realistic phone call from a loved one requesting urgent financial aid, only to discover it's a deepfake used by scammers. Or consider the potential for deepfakes to impersonate CEOs issuing false directives in videos, leading to massive financial fraud. These scenarios are no longer science fiction ; they are real threats in our digital age. As deepfakes become more sophisticated, discerning genuine interactions from manipulative fabrications becomes increasingly difficult. This highlights the critical need to cultivate skepticism and implement verification methods in digital communication.

Deepfakes pose a unique threat to businesses, going beyond traditional cybersecurity concerns. Companies that are already battling to protect data and finances now face a new challenge : safeguarding their reputation and authenticity. Imagine the devastating impact of a deepfake video showcasing a company leader in a compromising situation. Even if exposed as false, such content can inflict lasting damage on brand image and stakeholder trust. Deepfakes can also be weaponized in elaborate phishing schemes, tricking employees into following seemingly legitimate orders from superiors, potentially leading to data breaches or financial losses. The threat extends beyond external actors. Businesses also need to aid internal security to prevent the creation and dissemination of deepfakes within the organization. In today's digital landscape, a comprehensive security strategy must encompass defenses against deepfakes. This includes a combination of technological safeguards, employee awareness campaigns, and careful verification protocols [6].

### 1.1.6 Deepfake Fraud Examples

While deepfakes hold promise for creative efforts and entertainment applications, their potential for misuse is a growing concern. Deepfake technology has become a weapon in the hands of malicious actors, enabling them to commit a variety of malicious activities. This section will deeper into some recent, real-world examples of how deepfakes have been used to commit fraud, highlighting the diverse ways this technology can be exploited and the significant financial and reputational damage it can cause. By examining these cases, we can gain a deeper understanding of the evolving threat landscape and the importance of developing robust defences against deepfake fraud.

- **CEO of world's biggest ad firm targeted by deepfake scam.** The head of the world's biggest advertising group was the target of an elaborate deepfake scam that involved an artificial intelligence voice clone. The CEO of WPP, Mark Read, detailed the attempted fraud in a recent email to leadership, warning others at the company to look out for calls claiming to be from top executives.

  Fraudsters created a WhatsApp account with a publicly available image of Read and used it to set up a Microsoft Teams meeting that appeared to be with him and another senior WPP executive, according to the email obtained by the Guardian. During the meeting, the impostors deployed a voice clone of the executive as well as YouTube footage of them. The scammers impersonated Read off-camera using the meeting's chat window. The scam, which was unsuccessful, targeted an "agency leader", asking them to set up a new business in an attempt to solicit money and personal details. "Fortunately the attackers were not successful," Read wrote in the email. "We all need to be vigilant to the techniques that go beyond emails to take advantage of virtual meetings, AI and deepfakes."

A WPP spokesperson confirmed the phishing attempt bore no fruit in a statement : "Thanks to the vigilance of our people, including the executive concerned, the incident was prevented." WPP did not respond to questions on when the attack took place or which executives besides Read were involved [1].

- **Fraudsters Cloned Company Director's Voice In 35 Million Heist, Police Find.** In early 2020, a branch manager of a Japanese company in Hong Kong received a call from a man whose voice he recognized—the director of his parent business. The director had good news : the company was about to make an acquisition, so he needed to authorize some transfers to the tune of 35 million. A lawyer named Martin Zelner had been hired to coordinate the procedures and the branch manager could see in his inbox emails from the director and Zelner, confirming what money needed to move where. The manager, believing everything appeared legitimate, began making the transfers.

  What he didn't know was that he'd been duped as part of an elaborate swindle, one in which fraudsters had used "deep voice" technology to clone the director's speech, according to a court document unearthed by Forbes in which the U.A.E. has sought American investigators' help in tracing 400,000 of stolen funds that went into U.S.-based accounts held by Centennial Bank. The U.A.E., which is investigating the heist as it affected entities within the country, believes it was an elaborate scheme, involving at least 17 individuals, which sent the pilfered money to bank accounts across the globe.

  Little more detail was given in the document, with none of the victims' names provided. The Dubai Public Prosecution Office, which is leading the investigation, hadn't responded to requests for comment at the time of publication. Martin Zelner, a U.S.-based lawyer, had also been contacted for comment, but had not responded at the time of publication [2].

- **Obama's message to the public.** Most of the more convincing deepfakes have used imposters to impersonate the source's speech and mannerisms, such as this video developed by BuzzFeed and actor Jordan Peele combining After Effects CC and FakeApp. Peele's jaw was superimposed over Obama's, with a jawline that

[1] CEO of world's biggest ad firm targeted by deepfake scam https://www.theguardian.com/technology/article/2024/may/10/ceo-wpp-deepfake-scam

[2] Fraudsters Cloned Company Directors Voice In 35 Million Heist, Police Findhttps://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

matched Peele's mouth motions replacing Obama's. After that, FakeApp was used to improve the footage with almost 50 hours of automated processing [3].

- **Zuckerberg deepfake where he speaks frankly.** Artist Bill Posters uploaded this on Facebook-owned Instagram in June in reaction to Facebook's failure to remove the clip of Nancy Pelosi, displaying Mark Zuckerberg bragging about how the site "owns" its followers. The video was created as part of Posters and Daniel Howe's Spectre project, which was produced for Sheffield Doc Fest to highlight how one may use social media to deceive people. It was created using the VDR (video conversation substitution) software from Israeli 'Firm Canny AI, which is being pushed with a deepfake singalong with several international leaders.

  The posters used the hashtag deepfake to call attention to it. While the video seems convincing in silent mode, the voice gives it away, demonstrating that a competent actor is still required to create realistic deepfake instances. However, with Lyrebird and Adobe VoCo proposing AI voice generation, it may not be much until one can simply add passable sounds to deepfakes [4].

- **Yang Mi travels in time.** A video featuring Yang Mi, one of China's renowned current performers, pasted into the 1983 Hong Kong tv series The Legend Of The Condor Heroes went viral a few years ago, clocking up 240 million views before being taken off by Chinese authorities. Its maker, a Yang Mi admirer, apologized on Weibo, saying he made the film as a caution to promote awareness of the innovation.

  While the film and television industries are likely to react negatively to deepfakes at first, it is also feasible to see how the sector could ultimately accept the innovation and profit from it by enabling viewers to perform director on home updates by tricking dialogue, inserting alternate scenes, or even playing characters themselves. There will also be a slew of celebrity cameos in video games [5].

- **The Nancy Pelosi slowed-down video.** It was not a deepfake in the traditional sense ; instead, it illustrated why its possible misuse has become so dreaded in geopolitics. The 2019 clip was slowed down by 25 percent and video changed the pitch to make it appear as though Nancy Pelosi, the United States House of Representatives speaker, was gurning her words.

[3] This PSA About Fake News From Barack Obama Is Not What It Appearshttps://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed#.gcxNolpGL

[4] A deepfake video of Mark Zuckerberg presents a new challenge for Facebook https://edition.cnn.com/2019/06/11/tech/zuckerberg-deepfake/index.html

[5] Chinese A-lister falls victim to 'deepfake' video stunt https://www.techinasia.com/chinese-alister-falls-victim-deepfake-video-stunt

The tape was shared worldwide, and after the video was fact-checked and found to be fraudulent, Facebook declined to remove it, saying it had decreased its circulation. The post was later taken down, although it's unclear who was responsible and who took it down [6].

- **The Mandalorian Luke Skywalker deepfake.** Star Wars fans were ecstatic when Luke Skywalker appeared in The Mandalorian's season 2 finale. However, once the space dust had cleared, viewers were keen to point out problems in the digital reconstruction of a youthful Mark Hamill. YouTuber Shamook tried his hand at deepfaking a Luke Skywalker from the Return of the Jedi age, with stunning results.

  Shamook had been recruited by no one other than Industrial Light and Magic, the renowned visual effects company responsible for bringing the Star Wars universe to life. We're interested in watching how deepfake technology shapes the universe far, far away [7].

## 1.2 Problem Statement

The rapid advancement of deepfake audio technology is swiftly diminishing trust in the authenticity of the human voice, posing a significant threat to security and communication across multiple domains. Sophisticated Machine Learning (ML) techniques are being used by malicious actors more often to create extremely convincing synthetic audio. This allows them to spread false information, conduct fraud, and impersonate people with concerning ease.

This manipulation of audio recordings using Voice Conversion (VC) has deep implications for Islamic scholars, where it could cause misinformation, misguidance, reputation damage, and impact on religious Fatwas and decisions. For journalism, the credibility of audio evidence is crucial, as well as for finance, where voice authentication is used to secure transactions. In the legal sphere, deepfake audio can undermine the integrity of court proceedings, while in the world of national security, it can be used to spread propaganda. The potential for harm is immense, as deepfakes can be weaponized to damage reputations, manipulate public opinion, and even encourage violence.

[6] Fact check : "Drunk" Nancy Pelosi video is manipulated https://www.reuters.com/article/world/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2B1/

[7] Mandalorian's Luke Skywalker Without CGI : Mark Hamill, Deep Fake & Deaging https://screenrant.com/mandalorian-luke-skywalker-mark-hamill-no-cgi-deepfake-look/

## 1.3 Motivation

The motivation of our research thesis can be summarized in the following points.

- **Mitigate The Harm That Could Affect Islamic Scholars :** Islamic scholars are particularly vulnerable to deepfake. Misinformation could spread under the covers of religious statements if audio recordings are maliciously altered. The loss of trust in academic authority can have severe consequences for Muslim communities, as it may have an impact on religious rulings known as fatwas and decisions. Deepfake may also be utilized to harm a scholar's reputation, which could lead to conflict in society and religion.
- **Help Protect Financial Transactions That Relays On Voice Authentication :** Voice authentication is a crucial security measure in financial transactions, and deepfake can be used to compromise this security.
- **Protect Individuals and Organizations :** Deepfake audio can be used for identity theft, financial fraud, and reputation damage.
- **Safeguard Democracy and the Information Ecosystem :** Deepfake audio can be used to spread misinformation and propaganda and manipulate public opinion. Detection is crucial for maintaining a healthy democracy and combating fake news.
- **Restore Trust in Audio Evidence :** Deepfake audio can undermine the reliability and integrity of audio evidence in legal proceedings and investigations. Reliable detection can help restore trust in this critical form of evidence.
- **Advance the Field of Audio Forensics :** Developing new techniques for detecting fake speech can push the boundaries of audio forensics and contribute to a safer digital environment.

## 1.4 Objectives

This research aims to address the growing threat of deepfake audio by achieving the following objectives.

- **Develop a Benchmark for Deepfake Audio Detection in Arabic :** Current standards for deepfake audio detection are frequently trained on English-language datasets. Due to the unique characteristics of the Arabic language, certain benchmarks may not translate properly to Arabic audio. Thus, our goal is to create a thorough benchmark made especially for evaluating Arabic deepfake audio detection.

- **Design a Framework for Deepfake Audio Detection :** The accuracy and durability of the deepfake detection techniques used today frequently have limits and do not apply to the Arabic language. Our goal is to provide a novel architecture for deepfake audio recognition for the Arabic language to overcome these constraints.
- **Create a High-Quality Arabic Deepfake Audio Dataset :** The lack of large-scale, high-quality Arabic deepfake audio datasets delays the development and evaluation of robust deepfake detection models. Therefore, we aim to create a comprehensive Arabic deepfake audio dataset.

## 1.5 Outline

The remainder of this thesis is organised as follows :

**Chapter 2 : "Artificial Intelligence Background"** : This chapter provides the technical foundation for understanding deepfake audio detection. It covers essential concepts in artificial intelligence, machine learning, and deep learning.

**Chapter 3 : "Audio Fundamentals"** : This chapter provides the fundamentals of audio processing. Additionally, it deepens into the specific audio features and manipulation techniques relevant to deepfake detection.

**Chapter 4 : "State of the art in deepfake audio detection"** : This chapter comprehensively reviews existing research on deepfake audio detection. It examines various approaches, methodologies, and algorithms that have been proposed to address this challenge. It also identifies gaps and limitations in current research, highlighting areas where further investigation is needed.

**Chapter 5 : "Results and Contribution"** : This chapter presents the main contributions of our research alongside the results of our deepfake audio detection framework, highlighting the effectiveness of our framework.

## Conclusion

This chapter has introduced deepfake, its types, and its growing presence. We've examined the cybersecurity threats posed by deepfake audio, highlighting real-world examples that emphasize the need for robust detection methods. Chapter 2 delves into the essential concepts of AI, Machine Learning (ML), and Deep Learning (DL).

# Chapitre 2

# Artificial Intelligence Background

## Introduction

The fundamental concepts and technologies that underlie the development of deepfake audio detection systems will be presented in this chapter. We begin by exploring the concept of Artificial Intelligence (AI) and its various subfields, including Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP).

## 2.1 Definition of Artificial Intelligence

Artificial intelligence (AI) empowers computers and machines to mimic human intelligence and problem-solving abilities. This technology can be used independently or integrated with other technologies to automate tasks that traditionally require human input. Examples of AI in our everyday lives include digital assistants, GPS navigation, self-driving cars, and AI-powered content creation tools like ChatGPT.

Within the field of computer science, AI encompasses machine learning and deep learning, which involve developing algorithms that learn and adapt from data to make increasingly precise predictions or categorizations.

AI has experienced periods of heightened interest in the past, but the release of ChatGPT signifies a major milestone. While previous advancements in generative AI focused on computer vision, the latest breakthroughs have revolutionized natural language processing (NLP). Today, generative AI can learn and replicate various data types, including human language, images, video, code, and even molecular structures.

The applications of AI are constantly expanding, but with the growing enthusiasm surrounding AI tools in business, ethical considerations and responsible AI practices have become paramount [7].

### 2.1.1 Types of Artificial Intelligence

This subsection is dedicated to discussing the types of AI, described as follows.

- **Weak AI :** Also known as narrow AI or Artificial Narrow Intelligence (ANI) Weak AI, also called narrow AI or ANI, is a type of artificial intelligence designed and trained to excel at specific tasks. Although it's called "narrow," it's far from weak and powers many of the AI applications we use today. Some notable examples of weak AI include Apple's Siri, Amazon's Alexa, IBM Watsonx, and the technology behind self-driving cars [7].

- **Strong AI :** Strong AI encompasses two theoretical forms : artificial general intelligence (AGI) and artificial super intelligence (ASI). AGI refers to a hypothetical machine with human-level intelligence, possessing self-awareness, consciousness, and the capacity for problem-solving, learning, and future planning. ASI, or superintelligence, would surpass human capabilities in intelligence and ability [7]. Although strong AI remains purely theoretical without any existing real-world applications, AI researchers are actively investigating its potential development. For now, the closest representations of ASI are found in science fiction, such as the character HAL from the film "2001 : A Space Odyssey."

### 2.1.2 Subfields of Artificial Intelligence

- **Machine learning (ML) :** Is subset of artificial intelligence (AI) and computer science, leverages data and algorithms to enable AI systems to mimic human learning processes, gradually enhancing their accuracy over time [8].

- **Deep learning (DL) :** Is subset of artificial intelligence (AI) and computer science, leverages data and algorithms to enable AI systems to mimic human learning processes, gradually enhancing their accuracy over time [9].

## 2.2 Machine Learning

The basic concept of machine learning in data science involves using statistical learning and optimization methods that let computers analyse datasets and identify patterns. Machine learning techniques leverage data mining to identify historic trends and inform future models [10]. Machine Learning types can be found in Figure 2.1.

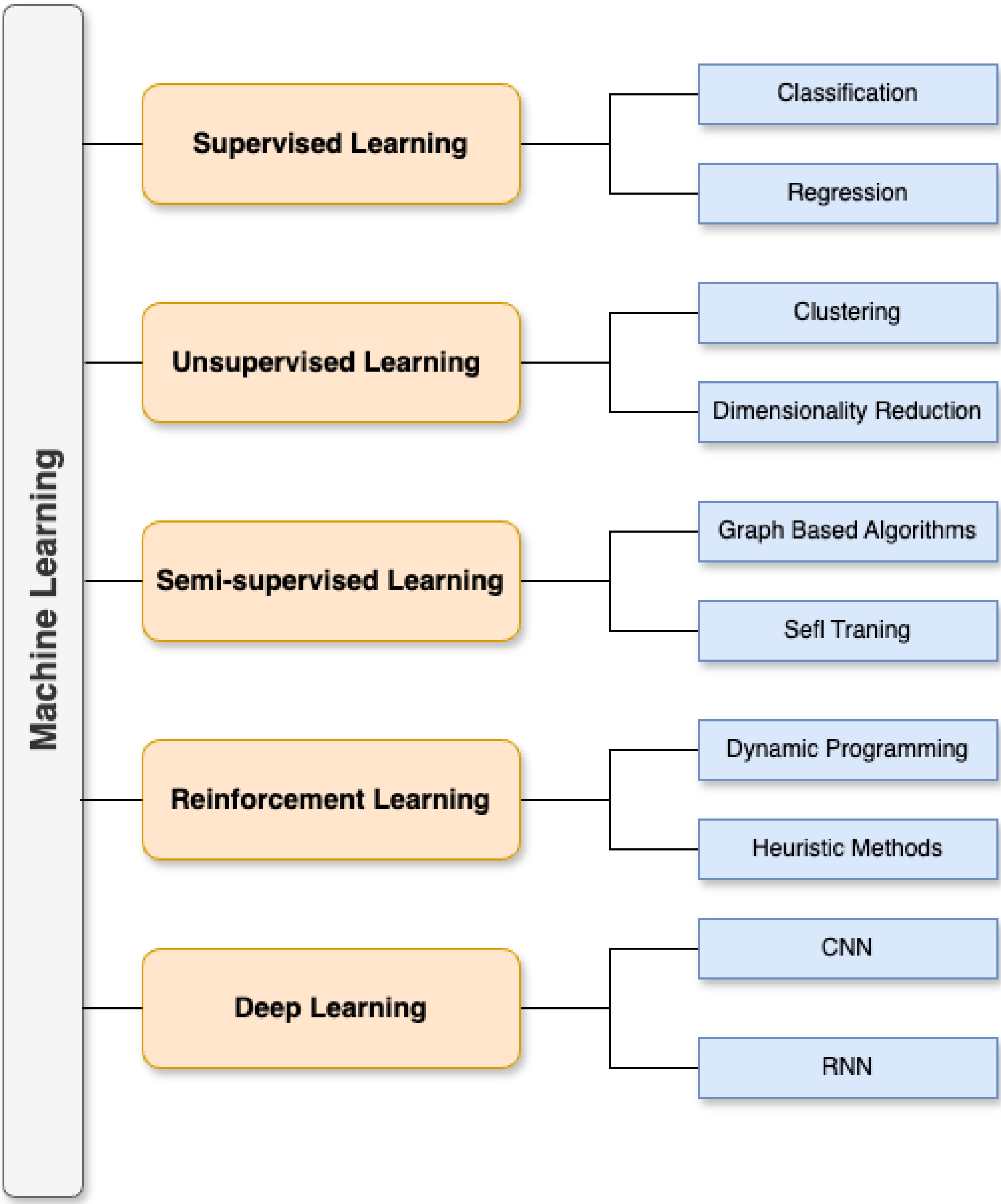

**Figure 2.1:** Machine learning Types

### 2.2.1 Supervised learning

Supervised machine learning utilizes labeled datasets to train algorithms for accurate data classification or outcome prediction. As input data is introduced, the model refines its internal parameters to optimize its fit. This process, part of cross-validation, safeguards against overfitting (excessive complexity) or underfitting (insufficient complexity). Supervised learning enables organizations to address diverse real-world challenges at scale, exemplified by filtering spam emails. Common techniques employed in supervised learning encompass neural networks, naïve Bayes, linear regression, logistic regression, random forests, and support vector machines (SVM). These are some popular algorithms of Supervised learning [8] :

- **Linear Regression :** Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting straight line that minimizes the sum of squared errors between the observed data and the predicted values.

- **Logistic Regression :** Logistic regression is a statistical model that analyzes the relationship between one or more independent variables and a binary dependent variable. It is used for classification problems, where the output is either 0 or 1, indicating the presence or absence of a particular characteristic.

- **Decision Trees :** A decision tree is a tree-like model that represents a series of decisions and their possible consequences. It is used for both classification and regression tasks, where the internal nodes represent feature tests, and the leaf nodes represent the final predicted classes or values.

- **Random Forests :** Random Forests are an ensemble learning method that combines multiple decision trees to improve prediction accuracy and control overfitting. Each tree is constructed using a random subset of features, and the final prediction is made by aggregating the predictions of all the individual trees.

- **XGBoost :** XGBoost is a powerful and efficient implementation of the gradient boosting algorithm. It is a tree-based ensemble machine learning algorithm that iteratively builds a series of weak decision trees and combines them to create a strong predictive model. XGBoost is known for its scalability, high performance, and ability to handle a wide range of data types and distributions.

- **Support Vector Machines :** Support Vector Machines are a powerful class of algorithms for classification and regression tasks. SVMs find the optimal hyperplane that maximizes the margin between the classes in the feature space. The data points closest to the hyperplane are called support vectors, and the goal is to maximize the distance between them and the hyperplane. SVMs can handle high-dimensional data and are effective for non-linear problems by using kernel tricks.

- **Naive Bayes :** Naive Bayes is a collection of algorithms based on Bayes' theorem and the assumption of independence between features. Despite its simplicity, it performs well on many classification tasks. It calculates the probability of each class given the feature values, and assigns the class with the highest probability. Naive Bayes is widely used in text classification, spam filtering, and sentiment analysis.

- **K-Nearest Neighbors (KNN) :** KNN is a non-parametric algorithm that classifies new instances based on their similarity to the k nearest neighbors in the training data. The algorithm computes the distances between the new instance and all the training instances, selects the k closest neighbors, and assigns the class label based on a majority vote of these neighbors. KNN is simple and effective, but can be computationally expensive for large datasets.

- **Gradient Boosting Machines (GBM) :** Gradient Boosting Machines are an ensemble learning technique that combines multiple weak decision tree models in an iterative, additive manner. Each new tree is trained to predict the residuals of the previous ensemble, gradually improving the overall model. GBMs are powerful and can handle complex, non-linear relationships, but can be prone to overfitting if not properly regularised.

### 2.2.2 Unsupervised learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets (subsets called clusters). These algorithms discover hidden patterns or data groupings without the need for human intervention. This method's ability to discover similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction. Principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, and probabilistic clustering methods.

These are some popular algorithms of Unsupervised learning [8] :

- **K-Means Clustering :** K-Means is one of the most widely used clustering algorithms. It partitions the data into K clusters by iteratively assigning data points to the closest cluster centroid and updating the centroids based on the assigned points. The goal is to minimize the sum of squared distances between data points and their assigned cluster centroids.

- **Hierarchical Clustering :** Hierarchical clustering algorithms build a hierarchy of clusters, either by merging smaller clusters into larger ones (agglomerative) or by dividing larger clusters into smaller ones (divisive). The result is typically visualized as a dendrogram, which represents the nested grouping of data points based on their similarity or distance.

- **Principal Component Analysis (PCA) :** PCA is a dimensionality reduction technique that transforms the original high-dimensional data into a lower-dimensional representation by finding the directions (principal components) that maximize the variance in the data. It can be used for data visualization, noise filtering, and feature extraction.

- **t-SNE (t-Distributed Stochastic Neighbor Embedding) :** t-SNE is a non-linear dimensionality reduction algorithm that is particularly well-suited for visualizing high-dimensional data in a low-dimensional space (typically 2D or 3D). It aims to preserve the local and global structure of the data, making it useful for identifying clusters and patterns.

### 2.2.3 Semi-supervised learning

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it's too costly to label enough data [10].

- **Self-Training (Self-Learning) :** Self-training is a wrapper algorithm that uses a small amount of labeled data to train a model initially. This model is then used to make predictions on the unlabeled data. The most confidently predicted unlabeled instances are added to the training set, and the process is repeated iteratively.

- **Graph-Based Methods :** Graph-based methods represent the data as nodes in a graph, with edges reflecting the similarity between instances. Labels are propagated from the labeled nodes to the unlabeled nodes based on the graph structure.

### 2.2.4 Reinforcement machine learning

Reinforcement machine learning is a machine learning model that is similar to supervised learning, but the algorithm isn't trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

The IBM Watson® system that won the Jeopardy ! challenge in 2011 is a good example. The system used reinforcement learning to learn when to attempt an answer (or question, as it were), which square to select on the board, and how much to wager—especially on daily doubles [8].

- **Q-Learning :** Q-Learning is a model-free reinforcement learning algorithm that learns an optimal action-selection policy for an agent interacting with its environment. It uses a Q-function that estimates the expected future reward for taking a particular action in a given state. The Q-values are iteratively updated based on the agent's experiences and rewards received.

- **SARSA (State-Action-Reward-State-Action) :** SARSA is another model-free reinforcement learning algorithm, similar to Q-Learning. However, instead of using the maximum Q-value for the next state, it updates the Q-value based on the actual action taken in the next state. SARSA is an on-policy algorithm, meaning it evaluates the same policy it is learning.

## 2.3 Deep Learning

A deep neural network (DNN) is technically defined as a neural network with three or more layers, but in reality, most DNNs have far more. Trained on extensive datasets, DNNs excel at identifying and classifying phenomena, discerning patterns and relationships, evaluating possibilities, and making predictions and decisions. While a single-layer neural network can offer basic predictive capabilities, the additional layers in a DNN enhance and refine these outcomes, yielding greater accuracy.

Deep learning technology empowers a wide range of applications and services that advance automation, enabling analytical and physical tasks to be performed without human involvement. It is the driving force behind everyday conveniences such as digital assistants, voice-activated TV remotes, and fraud detection systems, as well as cutting-edge technologies like self-driving cars and generative AI [11].

#### 2.3.0.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are composed of artificial neurons, known as units, which are organized into a series of layers. The number of units within a layer can vary significantly, ranging from just a few to millions, depending on the complexity of the patterns the network needs to learn from the dataset. Typically, an ANN consists of an input layer, one or more hidden layers, and an output layer. The input layer receives external data for analysis, which then flows through the hidden layers, undergoing transformations that make the information meaningful for the output layer. The output layer ultimately delivers the network's response to the input data [12]. ANN architecture can be found in Figure 2.2.

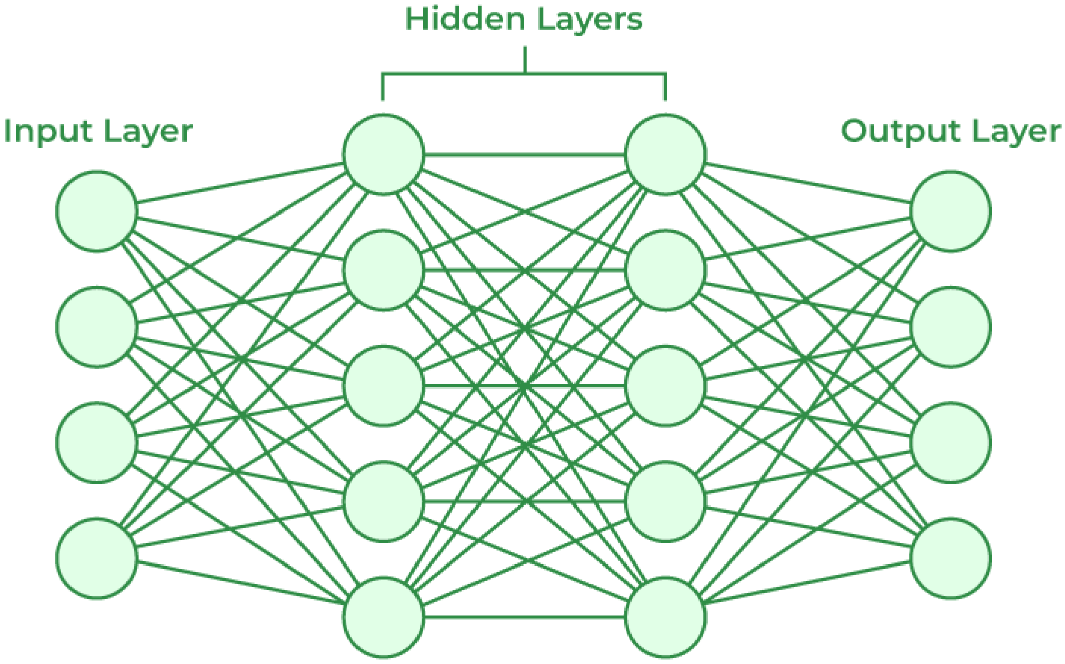

**Figure 2.2:** ANN architecture

#### 2.3.0.2 Convolution Neural Network

Convolutional Neural Networks (CNNs) are a specialized type of artificial neural network (ANN) designed to efficiently extract features from grid-like datasets, such as images or videos, where spatial patterns are crucial [12].

The architecture of a CNN comprises multiple layers :

- Input Layer : Receives the raw input data, typically an image or video. Convolutional Layer : Applies filters to the input to detect specific features like edges or textures.
- Pooling Layer : Downsamples the output of the convolutional layer, reducing the data's dimensionality and computational load.

- Fully Connected Layer : Processes the extracted features and makes a final prediction or classification.

The CNN learns to optimize its filters through a process of backpropagation and gradient descent, adjusting its parameters to improve its accuracy in recognizing patterns and making predictions. CNN architecture can be found in Figure 2.3.

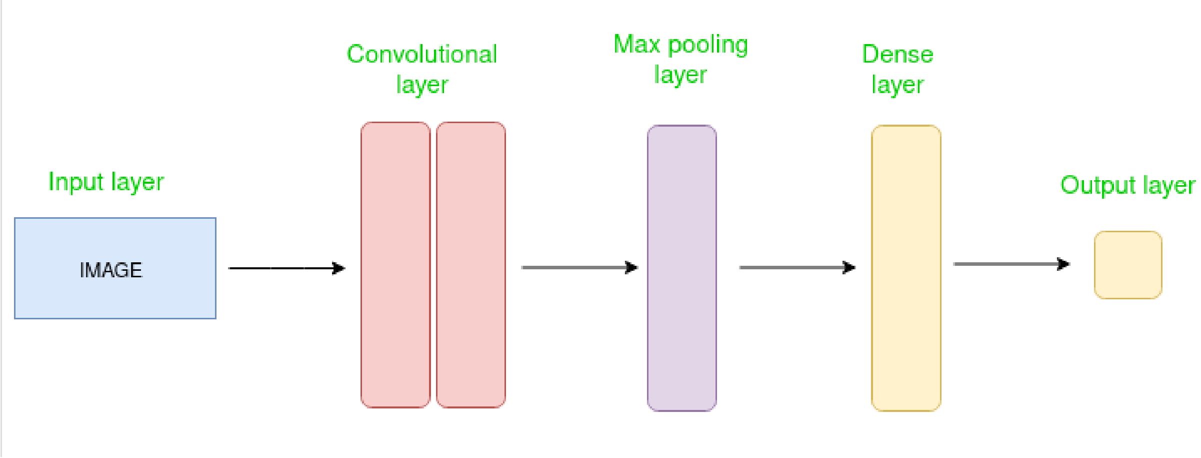

**Figure 2.3:** CNN architecture

#### 2.3.0.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a distinct type of neural network designed to process sequential data, where the output from one step becomes the input for the next. Unlike traditional neural networks, where inputs and outputs are independent, RNNs maintain a "hidden state" or "memory state" that allows them to retain information from previous inputs. This memory is crucial for tasks like language modeling, where predicting the next word in a sentence relies on understanding the context of the preceding words.

RNNs achieve this memory through the use of a hidden layer that carries information forward through the sequence. A key advantage of RNNs is their ability to reuse the same parameters for each input, simplifying the model's complexity compared to other neural networks. This parameter sharing allows the RNN to learn patterns and dependencies across sequential data effectively. RNN architecture can be found in Figure 2.4 [9].

#### 2.3.0.4 Long short-term memory (LSTM)

Long Short-Term Memory (LSTM) networks are a popular type of Recurrent Neural Network (RNN) designed to address the vanishing gradient problem, a challenge where information from distant past inputs can be lost or diluted. LSTMs introduce a mechanism called "cells" within their hidden layers, equipped with three gates (input, output,

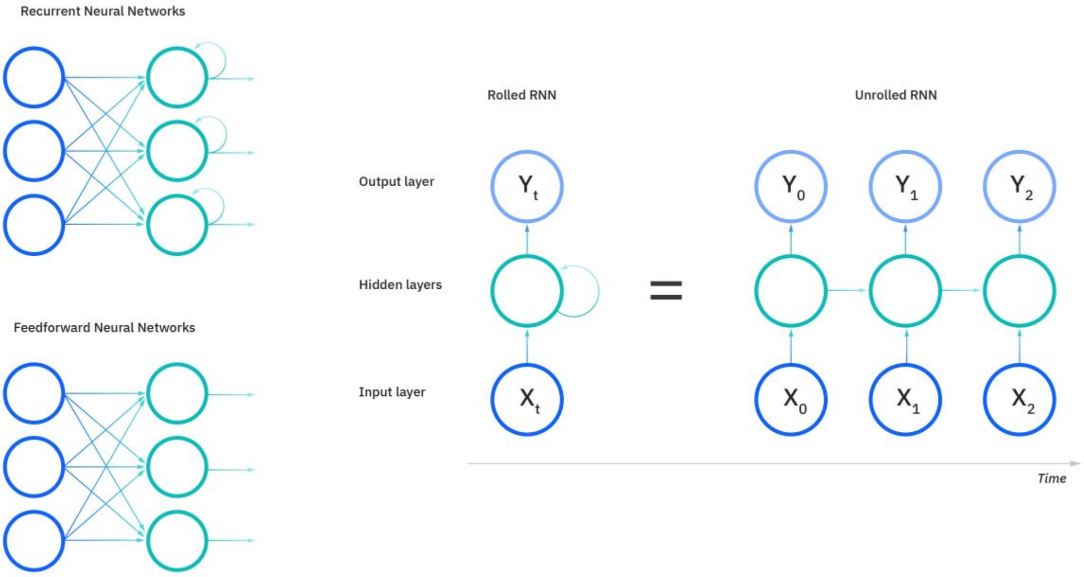

**Figure 2.4:** RNN architecture

and forget gates). These gates regulate the flow of information, enabling the network to selectively retain or discard data as it processes sequential input.

For instance, when predicting the word "peanut butter" in the sentence "Alice is allergic to nuts. She can't eat peanut butter," the LSTM's memory of the earlier mention of "nuts" is crucial. However, in standard RNNs, this connection might weaken if the relevant information appeared several sentences earlier. LSTMs overcome this limitation by actively controlling the flow of information through the gates, allowing them to maintain long-term dependencies and make accurate predictions even when the relevant context is distant.

The forget gate determines which information to discard from the cell state, the input gate decides what new information to store, and the output gate controls what information is passed on to the next time step. This dynamic control allows LSTMs to effectively utilize context from both recent and distant past inputs, making them particularly well-suited for tasks like natural language processing and time series analysis [9]. LSTM architecture can be found in Figure 2.5

#### 2.3.0.5 Transformers

The transformer model, introduced in the 2017 paper "Attention is All You Need," marked a significant turning point in AI, as it is now widely used in various applications, including Large Language Model (LLM) training.

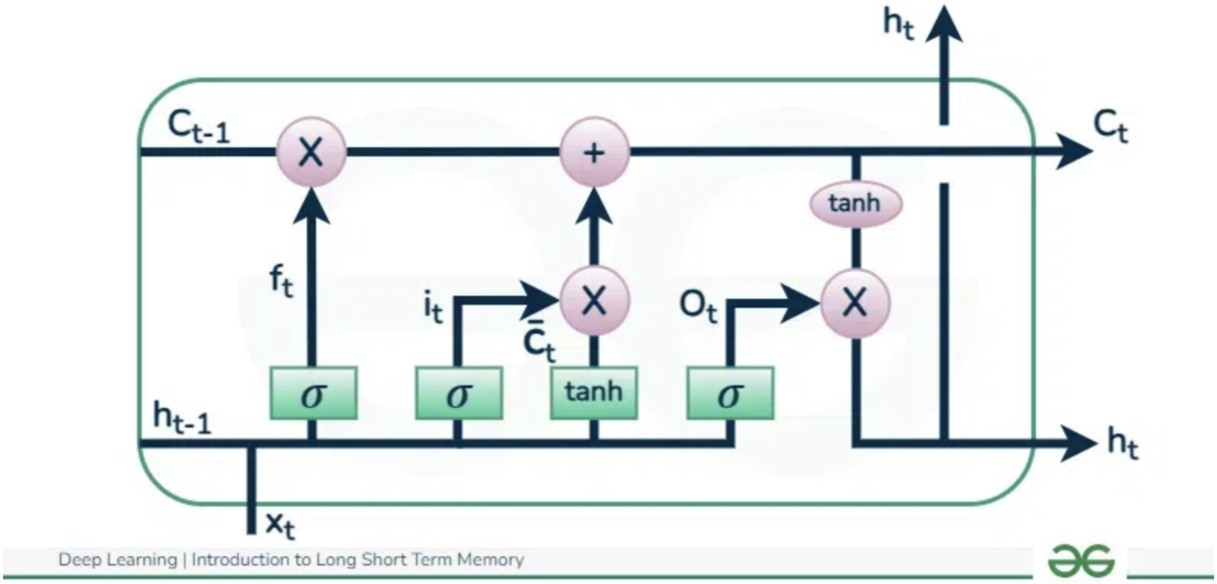

**Figure 2.5:** LSTM architecture

These models excel in real-time text and speech translation, facilitating communication for travelers and aiding researchers in areas like drug design and DNA analysis. They are also valuable in finance and security for anomaly detection and fraud prevention. Additionally, vision transformers have proven effective in computer vision tasks.

OpenAI's ChatGPT, a popular text generation tool, utilizes transformer architectures for prediction, summarization, question answering, and more. This is due to transformers' ability to focus on relevant input text segments. The "GPT" in various versions of the tool stands for "generative pre-trained transformer." These text-based generative AI tools benefit from transformer models as they efficiently predict the next word in a sequence, drawing from extensive and complex datasets.

Another notable model, BERT (Bidirectional Encoder Representations from Transformers), also leverages the transformer architecture. Since 2019, BERT has been employed for the majority of English-language Google searches and has expanded to over 70 other languages [2] Transformers architecture can be found in Figure 2.6.

## 2.4 Generative AI : The Engine Behind Deepfakes

Generative AI, a branch of artificial intelligence, specializes in developing models that can produce innovative and unique content. This encompasses a wide range of media, including images, music, text, and, importantly, audio. These models acquire knowledge by learning the patterns, styles, and structures present in existing data, and then leverage

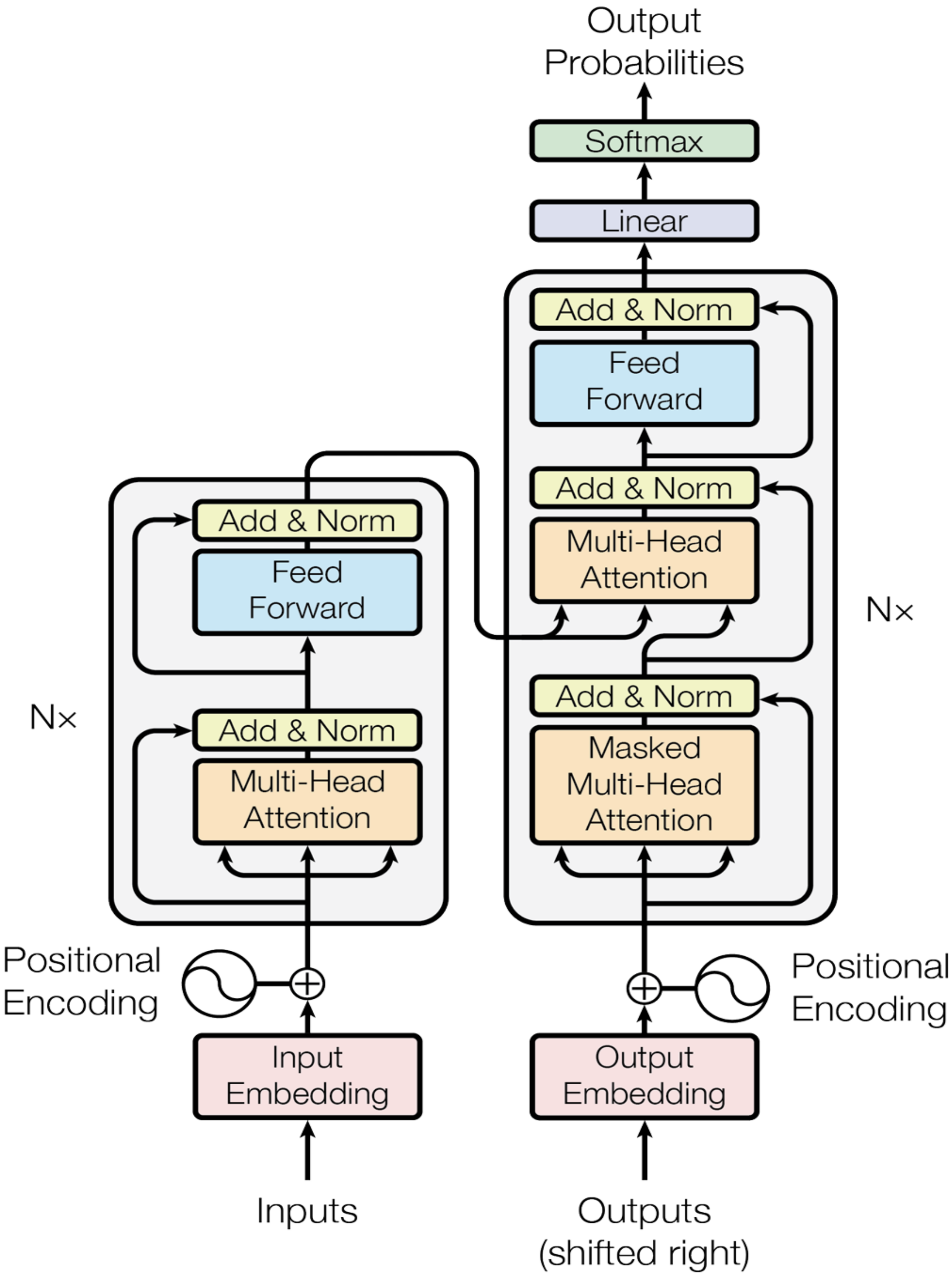

**Figure 2.6:** Transformers architecture [2].

this understanding to generate new content that closely resembles the original training data.

Comprehending the inner workings of generative AI is essential for deepfake audio detection. It uncovers the fundamental processes employed in creating these highly realistic fakes. By meticulously examining the traits of deepfakes produced by various models, researchers can devise more sophisticated detection methods capable of pinpointing subtle imperfections and inconsistencies that expose their artificial origin [13].

#### 2.4.0.1 Generative AI Models

- **Diffusion models :** Also referred to as denoising diffusion probabilistic models (DDPMs), diffusion models are generative models that identify vectors in latent space using a two-stage training process : forward diffusion and reverse diffusion. Forward diffusion gradually introduces random noise into training data, while reverse diffusion removes the noise to reconstruct the original data samples. The generation of new data involves running the reverse denoising process from entirely random noise. Training a diffusion model typically takes longer than training a variational autoencoder (VAE) model. However, due to the two-step process, diffusion models can train hundreds or even countless layers, leading to superior output quality in generative AI models. Furthermore, diffusion models are classified as foundation models due to their large scale, high-quality outputs, flexibility, and suitability for generalized use cases. Nonetheless, the reverse sampling process makes running foundation models a time-consuming endeavor [14].

- **Variational autoencoders (VAEs) :** Variational Autoencoders (VAEs) comprise two neural networks : an encoder and a decoder. The encoder receives an input and transforms it into a compact, concentrated representation of the data. This condensed version retains the essential information needed for the decoder to accurately reconstruct the original input while discarding any irrelevant details. The encoder and decoder collaborate to learn an efficient and simplified representation of the latent data. This enables users to easily sample new latent representations, which can then be processed by the decoder to generate novel data. Although VAEs can produce outputs like images more rapidly, the level of detail in their generated images is typically lower compared to the outputs of diffusion models [15].

- **Generative adversarial networks (GANs) :** Introduced in 2014, Generative Adversarial Networks (GANs) were widely recognized as the predominant approach among the three methods before diffusion models gained recent prominence. GANs operate through a competitive dynamic between two neural networks : a generator that produces new samples and a discriminator that learns to differentiate the generated content as either authentic (originating from the domain) or fabricated (generated). [16]

- **VITS (Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech) :** VITS is a deep learning model specifically designed for text-to-speech (TTS) synthesis. This versatile model stands out for its ability to generate high-fidelity speech waveforms directly from textual input. VITS's proficiency in producing natural-sounding speech makes it a popular choice for a variety of applications, including voice cloning, speech synthesis, and the creation of audiobooks and other narrated content. Additionally, VITS offers advantages over conventional TTS systems by incorporating elements from both variational autoencoders (VAEs) and generative adversarial networks (GANs) within its architecture. This combination allows VITS to capture the intricacies and nuances of human speech patterns, resulting in highly realistic and intelligible synthetic speech [17].

- **CREPE : A CONVOLUTIONAL REPRESENTATION FOR PITCH ESTIMATION :** CREPE is a deep learning model designed with a specific purpose : estimating the pitch of an audio signal. Unlike traditional methods that rely on handcrafted features or intricate signal processing pipelines, CREPE leverages a convolutional neural network (CNN) architecture. This CNN operates directly on the raw audio waveform, eliminating the need for manual feature engineering or complex preprocessing steps. This streamlined approach makes CREPE a more efficient and effective solution for pitch estimation tasks [18].

- **Retrieval-based Voice Conversion** : RVC is a deep learning model that excels in voice cloning and conversion. It builds upon the VITS architecture, originally designed for text-to-speech synthesis. The distinguishing feature of RVC is its integration of VITS with retrieval-based methods, enabling it to achieve high-quality voice conversions with minimal training data. In essence, RVC leverages the strengths of both VITS (for generating natural-sounding speech) and retrieval-based techniques (for efficient voice conversion with limited data) to create a powerful and flexible tool for voice modification tasks.

### 2.4.1 Applications of Generative Artificial Intelligence

Generative AI is a versatile technology that enhances the workflow of professionals across various fields, including creatives, engineers, researchers, and scientists. Its applications are virtually limitless, impacting all industries and individuals.

These models have the remarkable ability to take inputs in various forms, such as text, images, audio, video, or code, and transform them into new content in any of these modalities. For instance, they can generate images from text descriptions, compose music from visual cues, or even transcribe video content into text.

Here are the most popular generative AI applications :

- **Language :** Text forms the foundation for numerous generative AI models and is regarded as the most mature domain in this field. Large language models (LLMs) are a prime example of language-based generative AI, widely utilized for diverse tasks such as essay writing, code generation, translation, and even deciphering genetic sequences.
- **Audio :** Music, audio, and speech represent rapidly growing domains within generative AI. Some examples of its capabilities include models that can compose songs and audio clips from text inputs, identify objects within videos and generate corresponding sound effects, and even craft personalized music pieces.
- **Visual :** A prominent application of generative AI lies in the visual domain, encompassing the creation of 3D images, avatars, videos, graphs, and other illustrations. This technology offers flexibility in generating images with diverse artistic styles and provides techniques for editing and refining produced visuals. Generative AI models are utilized to construct graphs of novel chemical compounds and molecules, aiding in drug discovery, produce lifelike images for virtual and augmented reality experiences, craft 3D models for video games, design logos, enhance existing images, and much more.
- **Automotive industry :** In the automotive sector, generative AI is poised to revolutionize car development and simulation by creating realistic 3D environments and models. Additionally, synthetic data generated by AI is being used to train autonomous vehicles. This approach of virtually road testing self-driving capabilities in a simulated 3D world enhances safety, efficiency, and adaptability while significantly reducing risk and associated costs.
- **Field of natural sciences :** Generative AI is significantly impacting the field of natural sciences. Within healthcare, generative models are facilitating medical research by designing novel protein sequences that contribute to drug discovery. Additionally, medical professionals benefit from the automation of tasks like medical scribing, coding, imaging, and genomic analysis. In the field of meteorology, generative models enable the creation of detailed simulations of the planet, leading to more accurate

weather forecasts and predictions of natural disasters. These applications collectively enhance public safety and empower scientists to anticipate and prepare for environmental events more effectively.

- **Entertainment industry :** Across the entertainment industry, encompassing video games, film, animation, world building, and virtual reality, generative AI models are proving invaluable in streamlining the content creation process. Creators are increasingly embracing these models as tools to augment their creativity and enhance their workflows.

### 2.4.2 Benefits of generative AI

The most evident and significant advantage of generative AI lies in its ability to boost efficiency. By producing content and answers on demand, it has the potential to streamline or automate time-consuming tasks, reduce expenses, and liberate employees to focus on more valuable work.

However, generative AI offers a plethora of additional benefits for both individuals and organizations, some of them are :

- **Enhanced creativity :** Generative AI tools can stimulate creativity by automating brainstorming sessions, generating numerous original variations of content. These variations can serve as initial drafts or references, aiding writers, artists, designers, and other creative professionals in overcoming creative blocks.

- **Improved (and faster) decision-making :** Generative AI excels at analyzing vast datasets, uncovering patterns, and distilling valuable insights. It can then generate hypotheses and recommendations based on these insights, empowering executives, analysts, researchers, and other professionals to make more informed, data-driven decisions with greater speed and efficiency.

- **Dynamic personalization :** Generative AI enables dynamic personalization in applications such as recommendation systems and content creation. By analyzing user preferences and history, it can generate customized content in real time, resulting in a more tailored and engaging experience for each user.

- **Constant availability :** Generative AI operates tirelessly, offering uninterrupted availability 24/7 for tasks such as customer support chatbots and automated responses, ensuring consistent support and engagement.

### 2.4.3 Challenges, limitations and risks

Generative AI, despite its rapid advancement, comes with considerable challenges and risks for developers, users, and the public. A major concern is the phenomenon of "hallucinations," where AI generates outputs that appear factual but are nonsensical or entirely inaccurate. A well-known instance involved a lawyer who utilized a generative AI tool for research and received fictional case examples with fabricated quotes and attributions.

Some experts consider hallucinations an unavoidable trade-off for achieving creativity in AI models. However, developers can employ preventative measures, known as guardrails, to limit the model's reliance on unverified or untrusted data sources. Continuous evaluation and fine-tuning can also contribute to reducing hallucinations and improving the overall accuracy of AI-generated content.

- **Inconsistent outputs :** Generative AI models can sometimes produce inconsistent outputs even when given the same inputs due to their inherent variability. This inconsistency can be problematic for certain applications, like customer service chatbots, where consistent responses are essential. However, users can address this issue through prompt engineering, a process of iteratively refining and combining prompts to achieve the desired outcomes consistently in their generative AI applications.

- **Bias :** Generative models can inadvertently learn societal biases present in their training data, labeled data, external sources, or even from human evaluators involved in model tuning. This can result in the generation of biased, discriminatory, or offensive content. To mitigate this risk, developers must prioritize diverse training data, implement guidelines that proactively prevent bias during both training and fine-tuning stages, and continuously assess model outputs for bias alongside accuracy.

- **Lack of explainability and metrics :** A significant challenge with many generative AI models is their "black box" nature, making it difficult or impossible to comprehend their decision-making processes. Even the engineers and data scientists responsible for creating the underlying algorithms may struggle to explain the inner workings and how specific results are reached. Explainable AI practices and techniques can aid practitioners and users in understanding and trusting the processes and outputs of these models.

  Additionally, evaluating and comparing the quality of generated content poses a challenge. Traditional metrics may fall short in capturing the subtleties of creativity, coherence, and relevance. Developing reliable and robust evaluation methods for generative AI remains an ongoing area of research.

- **Threats to security, privacy and intellectual property :** Generative AI models pose potential threats to security, privacy, and intellectual property. Malicious actors can exploit these models to craft deceptive phishing emails, fabricated identities, or other harmful content that can deceive users into compromising their security and privacy. Developers and users must exercise caution to ensure that the data fed into the model, whether during fine-tuning or through prompts, does not inadvertently reveal their own intellectual property (IP) or any information protected as IP by other organizations. Furthermore, vigilant monitoring of model outputs is crucial to identify any new content that might expose their own IP or infringe upon the IP rights of others.

- **Deepfakes :** Deepfakes are artificially generated or manipulated images, videos, or audio designed to deceive viewers and listeners into believing that someone said or did something they did not. These creations are a chilling example of how generative AI can be maliciously exploited.

  While many are aware of deepfakes used to tarnish reputations or spread disinformation, cybercriminals have also weaponized this technology for cyberattacks (such as voice phishing scams) and financial fraud.

  Researchers are actively developing AI models capable of more accurately detecting deepfakes. However, until these become widely available, educating users and promoting best practices like verifying content before sharing can help mitigate the harm caused by deepfakes.

# Conclusion

This chapter has equipped us with the foundational knowledge of AI, particularly machine learning and deep learning, all essential for understanding deepfakes and their detection, next chapter will explore the fundamentals of audio.

# Chapitre 3

# Audio Fundamentals

## Introduction

In this chapter, we are exploring the concept of the world of audio, examining the properties of sound, digital audio basics, and audio signal processing techniques.

## Overview

Sound, at its essence, is a physical phenomenon that manifests as vibrations traveling through a medium, such as air or water. These vibrations create pressure waves that propagate outward from the source, eventually reaching our ears. When these waves interact with the delicate structures within our ears, they are transformed into electrical signals that are interpreted by our brains as sound. The characteristics of sound are diverse and complex. The frequency of a sound wave, measured in Hertz (Hz), determines its pitch, with higher frequencies perceived as higher notes and lower frequencies as lower notes. The amplitude of a sound wave, often measured in decibels (dB), corresponds to its loudness or intensity. The distance between successive peaks of a sound wave is its wavelength, and the speed at which these waves travel through a medium is known as the velocity of sound. These fundamental properties of sound form the basis for understanding audio, which refers to the electronic representation of sound waves. This representation is typically achieved through a process called analog-to-digital conversion (ADC), where continuous analog sound waves are sampled at regular intervals and converted into discrete digital values. In the next sections, we will delve deeper into the intricacies of digital audio representation, the various techniques used to process audio signals, and the specific challenges associated with analyzing and manipulating audio data. This comprehensive understanding of audio fundamentals is essential for comprehending the complex landscape of deepfake audio detection.

## 3.1 Properties of Sound

Sound waves possess several key properties that shape our auditory experience and are crucial to understanding audio manipulation techniques used in deepfakes [19] :

- **Frequency :** Measured in Hertz (Hz), frequency is the number of cycles a sound wave completes in one second. It directly correlates with the perceived pitch of a sound. Higher frequencies produce higher-pitched sounds, while lower frequencies result in lower-pitched sounds. Changes in pitch are fundamental to human speech and can be manipulated in deepfakes to alter the perceived voice.
- **Amplitude :** This refers to the maximum displacement of particles in a medium as the sound wave passes through. Amplitude is directly related to the loudness or intensity of a sound, measured in decibels (dB). Variations in amplitude create dynamics in speech, and deepfakes may manipulate these to make synthetic speech sound more natural or to mask inconsistencies.
- **Wavelength :** This is the distance between two consecutive peaks or troughs of a sound wave. Wavelength and frequency are inversely proportional, meaning that higher-frequency sounds have shorter wavelengths, and vice versa. Manipulating wavelengths in deepfakes can subtly alter the timbre of a voice, affecting its perceived quality and character.
- **Velocity (Speed of Sound) :** The speed at which sound waves travel varies depending on the medium through which they propagate. In general, sound travels faster in solids than in liquids, and faster in liquids than in gases. This property is relevant to deepfake detection as it affects the timing and synchronization of audio components, which can be exploited to identify inconsistencies in manipulated recordings.

Understanding these properties of sound is essential for analyzing and manipulating audio signals effectively. In the context of deepfake audio detection, analyzing variations in frequency, amplitude, and other acoustic features can provide valuable clues for distinguishing between authentic and synthetic speech. In subsequent sections, we will explore how these properties are leveraged in various techniques for audio signal processing and deepfake detection.

## 3.2 Digital Audio Basics

Digital audio is the cornerstone of modern sound recording, processing, and reproduction. It offers numerous advantages over analog audio, including greater fidelity, resistance to noise and degradation, and ease of editing and manipulation. However, to fully grasp the implications of digital audio in the context of deepfake detection, it's crucial to understand its fundamental principles [19].

### 3.2.1 Analog vs. Digital Audio

Sound, in its natural form, is an analog signal, meaning it is continuous and can take on an infinite number of values within a given range. Analog audio signals are represented as continuous fluctuations in voltage or current, mirroring the continuous variations in air pressure that constitute sound waves. In contrast, digital audio is a discrete representation of sound, where the continuous waveform is divided into distinct samples at regular intervals. Each sample is assigned a numerical value representing its amplitude at that specific instant. The process of converting analog audio to digital involves two key steps :

- **Sampling :** This involves measuring the amplitude of the analog signal at fixed intervals, known as the sampling rate. The higher the sampling rate, the more accurately the digital signal represents the original analog waveform. Common sampling rates include 44.1 kHz (used for CDs) and 48 kHz (used for digital video).
- **Quantization :** This involves mapping the continuous amplitude values of each sample to a finite set of discrete levels. The number of levels is determined by the bit depth of the digital audio. A higher bit depth allows for a wider range of amplitude values to be represented, resulting in greater dynamic range and reduced quantization noise (the error introduced by rounding analog values to discrete levels).

### 3.2.2 Sampling Rate and Bit Depth

The choice of sampling rate and bit depth significantly impacts the quality of digital audio. A higher sampling rate captures more detail and nuances of the original sound, resulting in greater fidelity and wider frequency response. Conversely, a lower sampling rate can lead to aliasing, where high-frequency components of the sound are misinterpreted as lower frequencies. Bit depth determines the dynamic range (the difference between the loudest and quietest sounds) and the level of quantization noise. A higher bit depth allows for a wider dynamic range and reduces quantization noise, resulting in cleaner and more accurate audio reproduction.

### 3.2.3 Audio File Formats

Digital audio is typically stored in various file formats, each with its own characteristics and trade-offs :

- **Uncompressed Formats :**
  - **WAV (Waveform Audio File Format) :** A lossless format that retains the original audio quality but results in large file sizes.
  - **AIFF (Audio Interchange File Format) :** Similar to WAV, commonly used on Apple devices.
- **Lossy Compressed Formats :**
  - **MP3 (MPEG Audio Layer III) :** A popular lossy format that significantly reduces file size by discarding less perceptible audio information.
  - **AAC (Advanced Audio Coding) :** A more efficient lossy format than MP3, often used for streaming and mobile devices.
- **Lossless Compressed Formats :**
  - **FLAC (Free Lossless Audio Codec) :** Provides compression without losing audio quality, ideal for archiving and high-fidelity playback.
  - **ALAC (Apple Lossless Audio Codec) :** Apple's version of a lossless compressed format.

Understanding these digital audio basics will provide a solid foundation for comprehending how audio data is represented, processed, and manipulated, which is crucial for developing effective deepfake audio detection techniques.

## 3.3 Audio Signal Processing

### 3.3.1 Fundamental Concepts

- **Fourier Transform :** The Fourier Transform is a mathematical technique that decomposes a signal into its constituent frequency components. In audio signal analysis, it is used to convert the time-domain representation of an audio signal into the frequency domain, allowing for the analysis of the signal's frequency content. This transformation is crucial for understanding the spectral characteristics of audio signals and extracting relevant features. [20] There are two main types of Fourier Transforms :

  - **Continuous-Time Fourier Transform (CTFT) :** The CTFT is used for continuous-time signals that are defined for all values of time. It represents the signal as a continuous function of frequency, providing a frequency spectrum that is also continuous.
  - **Discrete Fourier Transform (DFT) :** The DFT is used for discrete-time signals, which are sampled at specific time intervals. It represents the signal as a sequence of complex numbers, each corresponding to a specific frequency component. The DFT is particularly important for digital signal processing, as it allows for the efficient computation of the frequency spectrum of a sampled signal.

- **Time Domain Representation :** The time domain representation of an audio signal shows the amplitude of the signal over time. It depicts how the signal's amplitude varies as a function of time. In this representation, the x-axis represents time, and the y-axis represents the amplitude or intensity of the signal at each time instant. Time domain representations are intuitive and directly reflect the waveform of the audio signal as it is perceived by the human ear. However, they do not provide direct information about the frequency content of the signal, which is crucial for many audio analysis tasks.
- **Frequency Domain Representation :** The frequency domain representation, obtained through the Fourier Transform, decomposes the audio signal into its constituent frequency components. It shows the distribution of the signal's energy across different frequencies. In the frequency domain representation, the x-axis represents frequency, and the y-axis represents the amplitude or strength of each frequency component present in the signal. This representation provides valuable information about the spectral characteristics of the audio signal, such as the dominant frequencies, harmonic structures, and the distribution of energy across different frequency bands.

The Fourier Transform and frequency domain representation has several important properties and applications in audio signal processing :

- **Frequency analysis :** By transforming an audio signal into the frequency domain, the Fourier Transform reveals the signal's frequency content, which is essential for tasks such as speech recognition, music analysis, and audio compression.
- **Filtering :** The frequency domain representation obtained through the Fourier Transform enables efficient filtering operations by allowing for the selective manipulation of specific frequency components.
- **Feature extraction :** Many audio features, such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectral features, are derived from the frequency domain representation obtained through the Fourier Transform.
- **Audio compression :** Techniques like MP3 encoding utilize the Fourier Transform to identify and remove or encode less audible frequency components, achieving efficient audio compression.
- **Convolution and deconvolution :** The Fourier Transform simplifies the computation of convolution and deconvolution operations, which are important for tasks such as filtering, system identification, and signal restoration.

### 3.3.2 Filtering

Types of Filters : Audio filters are used to selectively attenuate or amplify specific frequency components of an audio signal. The three main types of filters are : [21]

- **Low-pass filters :** These filters allow low-frequency components to pass through while attenuating high-frequency components above a specified cutoff frequency.
- **High-pass filters :** These filters allow high-frequency components to pass through while attenuating low-frequency components below a specified cutoff frequency.
- **Band-pass filters :** These filters allow a specific range of frequencies to pass through while attenuating frequencies outside that range.

**Applications of Filtering :** Filtering is used in various audio applications, such as noise reduction, equalization, and signal separation. For example, low-pass filters can be used to remove high-frequency noise, while high-pass filters can remove low-frequency rumble or hum from audio signals.

**Concept of Filtering in Audio Signal Processing :** In the context of audio signal processing, filtering is a fundamental operation that allows for the selective manipulation of specific frequency components. By applying appropriate filters, it is possible to isolate, enhance, or suppress certain aspects of an audio signal, enabling various signal processing tasks.

### 3.3.3 Feature Extraction

Extracting Important Features : Feature extraction is the process of deriving relevant and informative characteristics from audio signals. These features capture essential properties of the audio data and are used for tasks such as audio classification, speech recognition, and audio analysis. Common Features : [22]

- **Mel-Frequency Cepstral Coefficients (MFCCs) :** MFCCs are widely used features in audio and speech processing. They represent the short-term power spectrum of an audio signal, taking into account the non-linear perception of frequency by the human auditory system. MFCCs are effective for capturing the spectral envelope of audio signals and are commonly used in speech recognition, speaker identification, and audio classification tasks.

$$MFCC_k = \sum_{n=1}^{N} \log(S_n) \cos\left[\frac{\pi k}{N}(n-0.5)\right], \quad k = 1, 2, \ldots, K \tag{3.1}$$

- **Chroma Features :** Chroma features, also known as chromagrams, represent the distribution of energy across different pitch classes (notes) in an audio signal. They are particularly useful for analyzing harmonic content and are widely used in tasks such as chord recognition, key detection, and music information retrieval.

$$C_k = \sum_{n \in \mathcal{H}_k} S_n, \quad k = 1, 2, \ldots, 12 \tag{3.2}$$

- **Root Mean Square (RMS) :** The RMS value is a measure of the average power or energy of an audio signal. It is often used as a feature for audio analysis tasks, such as loudness estimation, dynamic range compression, and audio event detection.

$$RMS = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{3.3}$$

- **Spectral Centroid :** The spectral centroid represents the "center of mass" of the frequency spectrum of an audio signal. It provides information about the overall brightness or sharpness of the sound and is commonly used in timbre analysis and audio classification tasks.

$$C = \frac{\sum_{k=0}^{N-1} f_k S_k}{\sum_{k=0}^{N-1} S_k} \tag{3.4}$$

- **Spectral Bandwidth :** The spectral bandwidth is a measure of the frequency range occupied by the majority of the energy in an audio signal's spectrum. It can be used to characterize the spread or concentration of energy across different frequencies, which is useful for tasks such as instrument recognition and audio texture analysis.

$$BW = \sqrt{\frac{\sum_{k=0}^{N-1} (f_k - C)^2 S_k}{\sum_{k=0}^{N-1} S_k}} \tag{3.5}$$

- **Spectral Rolloff :** The spectral rolloff is the frequency below which a specified percentage (e.g., 85 present or 95 present) of the total spectral energy is contained. It can be used to estimate the perceived brightness or sharpness of an audio signal and is often employed in audio classification and music genre recognition tasks.

$$R_f = \min\left\{ f \mid \sum_{k=0}^{f} S_k \geq 0.85 \sum_{k=0}^{N-1} S_k \right\} \tag{3.6}$$

- **Zero-Crossing Rate :** The zero-crossing rate is the rate at which an audio signal crosses the zero amplitude level. It provides information about the noisiness or periodicity of the signal and can be useful for tasks such as speech/music discrimination, onset detection, and audio segmentation.

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} \mathbb{I}\{x_n x_{n+1} < 0\} \tag{3.7}$$

These features, along with others, are commonly extracted from audio signals and can be used individually or in combination to capture different aspects of the audio data, enabling various audio analysis and processing tasks.

## 3.4 Speech Processing

Speech processing is a multidisciplinary field that encompasses the analysis, synthesis, and modification of spoken language. It plays a crucial role in various applications, from voice assistants and automatic transcription to deepfake audio generation and detection. Understanding the core principles of speech processing is essential for comprehending how deepfakes manipulate and mimic human voices [23].

### 3.4.1 Phonetics and Phonology

- **Phonetics :** This branch of linguistics studies the physical properties of speech sounds, including their production, acoustic characteristics, and perception. It examines how individual sounds (phonemes) are articulated and how they combine to form words.
- **Phonology :** This branch of linguistics focuses on the systematic organization of sounds in a language. It explores the rules and patterns governing how phonemes are combined to form words and how these words are used in meaningful communication.

Understanding phonetics and phonology is crucial for speech processing tasks, as it provides insights into the structure and organization of spoken language. This knowledge is leveraged in tasks like speech recognition, where the acoustic properties of speech sounds are used to identify words and phrases, and in text-to-speech synthesis, where the phonemic structure of a language guides the generation of natural-sounding speech.

### 3.4.2 Speech Recognition

Automatic speech recognition (ASR) is the process of converting spoken language into text. ASR systems typically consist of several components [23] :

- **Acoustic Modeling :** This component analyzes the acoustic features of speech signals to identify individual phonemes or sub-word units. It uses statistical models to predict the most likely sequence of phonemes given the acoustic input.
- **Language Modeling :** This component leverages linguistic knowledge to predict the most likely sequence of words based on the identified phonemes and the context of the utterance.
- **Decoding :** This component combines the outputs of the acoustic and language models to determine the most probable word sequence, effectively transcribing the spoken language into text.

Speech recognition technology is constantly evolving, with deep learning models like recurrent neural networks (RNNs) and Transformers playing an increasingly important role in achieving state-of-the-art performance.

### 3.4.3 Text-to-Speech (TTS) Synthesis

Text-to-speech (TTS) synthesis is the inverse of speech recognition, converting written text into spoken language. TTS systems analyze the text input, determine its linguistic structure, and generate corresponding acoustic signals that mimic human speech. TTS technology is used in various applications, including voice assistants, accessibility tools for the visually impaired, and, unfortunately, in the creation of deepfake audio. The ability to generate convincing synthetic speech using TTS models raises significant concerns about the authenticity and trustworthiness of audio recordings. In the context of deepfakes, TTS synthesis can be used to create fabricated speech that closely mimics the voice of a target individual, making it difficult to distinguish from genuine recordings. Understanding the intricacies of speech processing, including phonetics, phonology, speech recognition, and TTS synthesis, is essential for developing effective countermeasures against deepfake audio. By analyzing the unique characteristics of both human speech and synthetically generated audio, researchers can identify the subtle differences that reveal the artificial nature of deepfakes [23].

### 3.4.4 Voice Biometrics

Voice biometrics is the process of recognizing individuals based on their unique voice characteristics. It leverages the fact that each person's voice has distinct acoustic properties influenced by physiological and behavioral factors, such as vocal tract shape, articulation patterns, and speaking style. Voice biometrics has applications in various domains, including security, authentication, and forensics. Speaker Recognition : Speaker recognition is a fundamental task in voice biometrics, and it can be further categorized into two main types [24] :

- **Speaker Identification :** Speaker identification is the process of determining who is speaking from a set of known voice samples or speakers. It involves comparing an unknown voice sample against a database of enrolled voice models or templates. The system attempts to find the best match and identify the speaker from the available set of speakers.

- **Speaker Verification :** Speaker verification, also known as voice authentication or voiceprint recognition, is the process of confirming whether a speaker is who they claim to be. In this scenario, the speaker's identity is first claimed or provided, and the system compares the given voice sample against the pre-enrolled voice model associated with that claimed identity. The system then makes a binary decision, either accepting or rejecting the claim based on the similarity between the voice sample and the stored model.

Importance of Voice Biometrics in Security and Authentication : Voice biometrics has gained significant importance in security and authentication applications due to several key advantages :

- **Unique Identifiers :** Voice characteristics are unique to each individual, making them suitable for identification and verification purposes.
- **Non-Invasive :** Voice biometrics is a non-invasive and contactless authentication method, providing a convenient user experience.
- **Remote Authentication :** Voice-based authentication can be performed remotely, enabling secure access to systems and services without physical presence.
- **Multimodal Authentication :** Voice biometrics can be combined with other biometric modalities, such as face or fingerprint recognition, for enhanced security and accuracy.
- **Forensic Applications :** Voice recordings can be used in forensic investigations, law enforcement, and intelligence gathering for speaker identification and verification.

Features Used in Voice Biometrics : Effective voice biometric systems rely on extracting robust and discriminative features from voice samples. Some commonly used features in speaker recognition include :

- **Pitch :** The fundamental frequency of a person's voice, which is influenced by the length and tension of the vocal cords.
- **Tone :** Refers to the resonant frequencies or formants produced by the vocal tract, which contribute to the unique timbre or quality of a person's voice.
- **Cadence :** The rhythm, stress patterns, and intonation of speech, which can be influenced by factors like accent, emotion, and speaking style.
- **Spectral features :** Characteristics derived from the frequency spectrum of the voice signal, such as Mel-Frequency Cepstral Coefficients (MFCCs), which capture the spectral envelope and harmonic structure of the voice.

- **Temporal features :** Characteristics related to the time-varying nature of speech, such as energy contours, zero-crossing rates, and duration patterns.

Challenges in Extracting Robust Speaker-Specific Features : While voice biometrics offers several advantages, extracting robust and reliable speaker-specific features can be challenging due to various factors :

- **Intra-Speaker Variability :** A person's voice can vary due to factors like emotional state, health conditions, aging, and environmental conditions (e.g., background noise), making it difficult to capture consistent features.
- **Inter-Speaker Similarity :** Some speakers may have similar voice characteristics, especially if they are related or share similar physiological or linguistic backgrounds, complicating the discrimination process.
- **Channel and Noise Effects :** The recording conditions, transmission channels, and background noise can introduce distortions and artifacts that can degrade the quality of the extracted features.
- **Spoofing Attacks :** Voice biometric systems can be vulnerable to spoofing attacks, where an impostor attempts to mimic or synthetically generate a target speaker's voice, compromising the system's security.

To address these challenges, researchers and developers in the field of voice biometrics continuously work on improving feature extraction techniques, exploring robust and invariant features, and developing advanced machine learning algorithms and countermeasures against spoofing attacks. Additionally, multi-modal biometric systems that combine voice with other modalities, such as facial features or behavioral characteristics, can enhance the overall accuracy and reliability of biometric recognition systems.

## 3.5 Challenges in Audio Analysis

Despite the advancements in audio processing techniques, several challenges persist in analyzing audio signals, particularly in the context of deepfake detection. These challenges can significantly impact the accuracy and reliability of audio analysis tasks, requiring careful consideration and innovative solutions.

#### 3.5.0.1 Noise and Distortion

Background noise and signal distortion are ubiquitous in real-world audio recordings. Noise, stemming from various sources such as environmental sounds, microphone artifacts, or electronic interference, can obscure or mask the relevant acoustic information in the audio signal. Distortion, on the other hand, can arise from clipping, compression, or other non-linear effects that alter the original waveform, introducing unwanted artifacts and affecting the overall quality of the audio. Both noise and distortion can hinder the performance of audio analysis tasks, such as speech recognition, speaker identification, and deepfake detection. To mitigate these issues, various signal processing techniques are employed, including :

**Noise Reduction :** Methods like spectral subtraction, Wiener filtering, and deep learning-based denoising algorithms can be used to reduce background noise and enhance the desired signal.

**Signal Enhancement :** Techniques like equalization, dynamic range compression, and de-clipping can be used to improve the overall quality of the audio signal and compensate for distortions.

In the context of deepfake detection, effective noise reduction and signal enhancement are crucial for ensuring that the analysis focuses on the genuine characteristics of the audio, rather than artifacts introduced by noise or distortion.

#### 3.5.0.2 Variability in Speech

Human speech is incredibly diverse, with individuals exhibiting a wide range of accents, dialects, intonation patterns, and speaking styles. Factors like emotional state, age, gender, and health conditions can further influence the acoustic properties of speech. This inherent variability poses a significant challenge for audio analysis, as algorithms need to be robust enough to handle this diversity and accurately extract meaningful features from different voices and speaking conditions. To address this challenge, researchers employ various approaches, including :

**Speaker Adaptation :** This technique involves fine-tuning speech processing models to adapt to the specific characteristics of individual speakers or groups of speakers.

**Robust Feature Extraction :** Developing feature extraction techniques that are less sensitive to variations in speech can improve the performance of audio analysis tasks.

**Large and Diverse Training Data :** Training machine learning models on large and diverse datasets that encompass a wide range of speakers and speaking conditions can enhance their ability to generalize to unseen data.

#### 3.5.0.3 Security Concerns

The rise of deepfake audio technology raises significant security concerns, particularly in the context of spoofing attacks. Malicious actors can exploit deepfakes to impersonate individuals, bypass voice authentication systems, and spread misinformation. These attacks can have severe consequences, including financial fraud, identity theft, and damage to reputation. The development of reliable deepfake audio detection methods is crucial for mitigating these security risks. By analyzing the subtle differences between genuine and synthetic speech, detection algorithms can identify manipulated audio and prevent it from being used for malicious purposes. This ongoing research effort is essential for safeguarding the integrity of audio information and maintaining trust in digital communication. By addressing these challenges head-on, researchers are paving the way for more accurate, robust, and secure audio analysis systems that can effectively detect deepfakes and other forms of audio manipulation.

## Conclusion

This chapter has equipped us with the foundational knowledge of audio processing techniques, essential for understanding deepfakes and their detection. We are now prepared to explore the current state-of-the-art deepfake detection methods in the following chapter.

# Chapitre 4

# State of the Art in Deepfake Audio Detection

## Introduction

This chapter provides a comprehensive overview of the current state of the art in deepfake audio detection. It explores the diverse approaches being pursued by researchers, ranging from traditional audio forensics techniques to cutting-edge machine learning and deep learning models. We will examine the strengths and limitations of existing methods, identify key challenges and open questions in the field, and highlight emerging trends that show promise for the future of deepfake detection.

## 4.1 Traditional Audio Forensics Techniques

Before the advent of deep learning-based approaches, audio forensics relied on a variety of established techniques to analyze and authenticate audio recordings. These techniques primarily focused on identifying inconsistencies, artifacts, or patterns within the audio signal that could indicate tampering or manipulation.

Traditional audio forensics techniques can be broadly categorized into two main approaches [24] :

**Signal Analysis :** This involves examining the raw audio waveform and its spectral representation to identify anomalies or inconsistencies that might indicate tampering.

**Content Analysis :** This focuses on analyzing the content of the audio, such as the spoken words, acoustic environment, or background noise, to determine its authenticity.

#### 4.1.0.1 Techniques

Several traditional techniques have been employed in audio forensics [24] :

- **Audio Fingerprinting :** This involves creating a unique "fingerprint" of an audio recording based on its acoustic characteristics. By comparing fingerprints, forensic experts can identify identical or similar recordings and detect instances of re-use or manipulation.
- **Spectrographic Analysis :** This involves visually examining the spectrogram of an audio signal, which displays the distribution of frequencies over time. Experts can identify discontinuities, abrupt changes, or unnatural patterns in the spectrogram that might suggest editing or tampering.
- **Electrical Network Frequency (ENF) Analysis :** This technique analyzes the subtle fluctuations in electrical power frequency that are often embedded in audio recordings. By comparing the ENF patterns of a recording with known reference patterns, experts can verify the time and location of the recording and potentially detect tampering.
- **Auditory Analysis :** Trained forensic experts can listen critically to audio recordings, using their knowledge of speech patterns, acoustics, and production techniques to identify inconsistencies, artifacts, or anomalies that might indicate manipulation

#### 4.1.0.2 Limitations

While traditional audio forensics techniques have been valuable tools for authentication and tampering detection, they face limitations when confronted with sophisticated deepfakes :

- **Handcrafted Features :** Many traditional methods rely on handcrafted features that may not capture the subtle nuances and complex patterns introduced by deepfake algorithms.
- **Limited Generalization :** These techniques often require extensive knowledge and expertise in audio analysis and may not generalize well to novel manipulation techniques.
- **Time-Consuming :** Manual analysis of audio recordings can be time-consuming and labor-intensive, especially for large volumes of data.

As deepfake technology continues to advance, traditional methods are becoming increasingly inadequate for detecting sophisticated manipulations. This has led to a growing interest in exploring machine learning-based approaches, which offer the potential for more automated, adaptable, and accurate deepfake audio detection.

## 4.2 Current State-of-the-Art Detection Methods

In recent years, there has been a significant shift towards leveraging machine learning and deep learning techniques for deepfake audio detection, driven by their ability to learn complex patterns and adapt to evolving manipulation techniques. This section will review some of the prominent methods in this area, highlighting their key features, strengths, and limitations. A comparative analysis of the state-of-the-art methods for detecting synthetic speech can be found in Table 4.1

### 4.2.1 Machine Learning-Based Methods

Alegre el al. [25] presented a novel countermeasure to detect spoofing attacks in automatic speaker verification (ASV) systems. Spoofing attacks are when a bad actor tries to fool the ASV system by impersonating someone else's voice. The countermeasure uses local binary patterns (LBPs) to analyze speech signals and a one-class classification approach to distinguish between genuine and spoofed speech. The results showed that the countermeasure was effective in detecting all three types of spoofing attacks voice conversion attacks, speech synthesis and artificial signal attacks.

Jordan J. Bird and Ahmad Lotfi [26] addressed the growing implications of deepfake voice conversion. They proposed a robust machine learning approach to detect deepfake audio generated using Retrieval-based Voice Conversion (RVC). They extracted 26 audio features and then generated the DEEP-VOICE dataset, which consists of real human speech from eight well-known figures, they achieved effective performance using the XG-Boost model.

### 4.2.2 Deep Learning-Based Methods

Wu et al. [27] presented Quick-SpoofNet, an innovative approach leveraging one-shot learning and metric learning to detect audio deepfakes, including previously unseen attacks, in automatic speaker verification (ASV) systems. Addressing the challenge of generalizing to unseen spoofing attacks, which existing countermeasures struggle with due to assumptions of similar data distributions between training and test utterances. QuickSpoofNet was evaluated using the ASVspoof2019 dataset's logical access (LA) subpart for in-domain assessment with unseen attacks. For generalization, subsets from the ASVspoof2021 dataset were used, representing real-world scenarios with unseen attacks. The method achieved a remarkable overall Equal Error Rate (EER) and accuracy on the three ASVspoof2019 dataset parts (LA, PA, DF) and demonstrated robust performance against various voice cloning and conversion algorithms within this dataset.

Authors of [28] proposed a new method to detect synthetic speech, which is fake speech created by computers using text-to-speech (TTS) and voice conversion (VC) algorithms. The method is based on the idea that real speech has a more consistent pattern than synthetic speech. The method trains a model to recognize the patterns of real speech. The authors tested their method on the ASVspoof 2019, a dataset specifically designed for testing methods that detect fake speech. The results showed that their method was better at detecting synthetic speech than other existing methods. This is especially true for synthetic speech created by methods the model hadn't seen before, showing that the method is good at generalizing to new types of synthetic speech.

Zeinali1 et al. [29] proposed a new system where they extracted 3 features and built three different model variants based on ResNET to detect synthetic speech created by TTS and VC algorithms, the authors used the ASVspoof 2019 dataset to train their models, after the conducted experiments the fusion model showed promising results compared to other methods.

Spoofing attacks, which try to impersonate legitimate users using techniques like voice conversion, speech synthesis, replay, and impersonation, are a threat to biometric authentication, especially Automatic Speaker Verification (ASV) systems. To address this threat Parasu et al. [30] proposed the Light-ResNet architecture with spectrogram input features to improve generalization in spoofing detection. Evaluations conducted on various databases, including ASVspoof 2015, BTAS 2016 (replay), and ASVspoof 2017 V2.0, demonstrate that the Light-ResNet architecture consistently outperforms baseline systems in terms of generalization and spoofing detection performance

# Conclusion

We reviewed deepfake audio detection techniques, from traditional methods to cutting-edge machine learning and deep learning approaches. While traditional methods have

| Paper Title | Classifier | Key Features | Dataset(s) Used | Performance Metrics | Strengths | Limitations |
|---|---|---|---|---|---|---|
| Alegre et al. [25] | Machine Learning (One-class SVM) | - Local Binary Patterns (LBPs) - One-class classification | NIST'05, NIST'06 | - Reduced FAR from 55% to 4.1% for voice conversion - Reduced FAR to 0.2% for speech synthesis and artificial signals | - Generalization to unseen attack types | - May struggle with highly sophisticated attacks - Limited to spectro-temporal features |
| Jordan J. Bird and Ahmad Lotfi. [26] | Machine Learning (XGBoost) | - 26 audio features - Multiple ML models compared | DEEP-VOICE (custom dataset) | - 99.3% average accuracy - 0.995 precision, 0.991 recall | - Real-time detection - High performance across metrics | - Limited to RVC-based voice conversion - May not generalize to other types of voice synthesis |
| Khan and Malik. [27] | Deep Learning (Siamese LSTM) | - One-shot learning - Metric learning | ASVspoof2019-LA, ASVspoof2021-DF, VSDC-0PR | - 0.50% EER on ASVspoof2019-LA - 86.41% accuracy on ASVspoof2021-DF | - Generalizes to unseen attacks - High performance on in-domain data | - Computationally intensive - May require large support set for best performance |
| Wu et al. [28] | Deep Learning (Light CNN) | - Feature genuinization - Lightweight model | ASVspoof 2019 | EER - 0.25%, Precision - 97.29%, Accuracy - 98.50%, F1 - 95.50%, Recall - 93.20% | - Lightweight architecture | - Computationally intensive - struggle with other languages |
| Zeinali1 et al. [29] | Deep Learning (ResNET-based models) | - generalization against unknown attacks | ASVspoof 2019 | improved t-DCF and EER by 71% and 75% respectively | - Generalizes to unseen attack types | - Computationally intensive - struggle with other languages |
| Parasu et al. [30] | Deep Learning (Light-ResNET) | - generalization against unknown attacks | ASVspoof 2019 | improved t-DCF and EER by 71% and 75% respectively | - Generalizes to unseen attack types | - Computationally intensive - struggle with other languages |

**Tableau 4.1:** Comparison of State-of-the-Art Methods for Detecting Synthetic Speech

limitations, machine learning models offer new possibilities. Challenges remain, including generalization to unseen speakers and manipulation techniques. Emerging trends like multi modal approaches and self-supervised learning hold promise. Continued innovation is crucial to stay ahead of deepfake technology's evolving threat.

# Chapitre 5

# Contributions and Results

## Introduction

This chapter details the methodological framework employed to develop a robust and reliable deepfake audio detection system for the Arabic language. The primary research goal is to create a model capable of accurately distinguishing between genuine and synthetically Arabic-generated speech. To achieve this, we have taken a data-centric approach, emphasizing the importance of creating a high-quality, diverse, and representative Arabic dataset for training and evaluation.

## 5.1 Proposed Contributions

The primary contributions of our research, which are twofold : the creation of the Arabic Audio Deepfake dataset and the development of a machine learning-based framework for deepfake audio detection. These contributions aim to address the gap in resources and techniques for detecting audio deepfakes in the Arabic language, thereby advancing the field of audio forensics and security.

### 5.1.1 Contribution 1 : Dataset Creation - Arabic Audio Deepfake

One of the critical challenges in the field of deepfake detection, particularly in audio, is the lack of comprehensive and representative Arabic datasets. To address this, we have created the Islamic Scholars Arabic Audio Deepfake dataset. This dataset is specifically tailored to support the development and evaluation of deepfake detection algorithms for Arabic audio content. The creation of this dataset involved several key steps :

#### 5.1.1.1 Data Collection and Prepossessing

The dataset for this research was constructed by collecting publicly available audio recordings sourced from YouTube. These recordings include a variety of sources, including interviews, speeches, podcasts, and other forms of spoken content, ensuring a diverse representation of real-world speech scenarios. To capture the rich diversity within the Arabic language, we accurately selected recordings from five distinct speakers, each representing a unique region and dialect.

| Speaker | Profession | Nationality |
|---|---|---|
| Sheikh Muhammad bin Salih al-Uthaymeen[1] | Islamic Scholar | Saudi |
| Sheikh Ibn Baz[2] | Islamic Scholar | Saudi |
| Sheikh Salih bin Fawzan al-Fawzan[3] | Jurist and Professor | Saudi |
| Muhammad al-Ghazali[4] | Islamic Preacher and Thinker | Egyptian |
| Muhammad Ratib al-Nabulsi[5] | Islamic Preacher | Syrian |

[1] YouTube Video (Last accessed : 01-05-2024)
[2] YouTube Video (Last accessed : 01-05-2024)
[3] YouTube Video (Last accessed : 01-05-2024)
[4] YouTube Video (Last accessed : 01-05-2024)
[5] YouTube Video (Last accessed : 01-05-2024)

**Tableau 5.1:** List of speakers, their professions, and nationalities.

This selection ensures that the dataset encompasses the variations in pronunciation, intonation, and vocabulary that exist across different regions and dialects of the Arabic language, enhancing the model's ability to generalize to diverse real-world scenarios. For each speaker, we collected approximately 20 minutes of high-quality audio recordings. These recordings were carefully selected to minimize background noise and ensure optimal audio quality for subsequent analysis. The 20-minute recordings were then segmented into shorter, 1-minute samples, resulting in 20 individual audio files per speaker. This segmentation not only facilitates efficient data processing but also provides a standardized format for feature extraction and model training. Given the high quality of the source recordings, there was no need for additional noise reduction or audio enhancement during the preprocessing stage. The resulting dataset comprises a total of 100 real audio samples (5 speakers x 20 samples per speaker), providing a robust foundation for training and evaluating our deepfake audio detection system. In the following sections, we will explore the details of deepfake generation and data annotation, explaining how we created synthetic audio samples and labeled them accordingly to construct a comprehensive dataset for deepfake audio detection.

#### 5.1.1.2 Deepfake Generation

To augment our dataset with synthetic samples and create a robust training environment for our deepfake audio detection model, we employed the RVC (Retrieval-based Voice Conversion) model. RVC is a cutting-edge deep learning-based model designed for high-quality voice conversion tasks. By leveraging the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) architecture, RVC excels in producing natural and expressive voice conversions with relatively small amounts of training data. In our deepfake generation process, we utilized each of the five unique speakers as both a source and a target for voice conversion. The RVC model was employed to convert the voice of each speaker into the voices of the remaining four speakers, effectively creating a matrix of pairwise voice conversions. This approach resulted in a total of 400 distinct deepfake audio samples per speaker ( 20 minutes divided into 20 samples ), where each sample represents the source speaker's voice converted to one of the other four speakers. This approach covered a wide array of voice transformations across different regions and dialects of the Arabic language. The diversity of these synthetic samples is essential for training a robust deepfake detection model that can generalize effectively to real-world scenarios, where manipulated audio might originate from various sources and utilize different conversion techniques. The subsequent section will elaborate on the data annotation process, detailing how each audio sample (both real and synthetic) was meticulously labeled to create a balanced and informative dataset for training and evaluating our deepfake audio detection system.

#### 5.1.1.3 Data Annotation

To facilitate the training and evaluation of our deepfake audio detection model, a meticulous data annotation process was conducted. The goal of this process was to assign clear and consistent labels to each audio sample, indicating whether it was a genuine recording or a synthetically generated deepfake. Each of the five speakers in our dataset was assigned a unique abbreviation to facilitate labeling. For instance, Speaker 1 was labeled as "S1," Speaker 2 as "S2," and so on. To differentiate between real and deepfake samples, each audio file was labeled as either "real" or "fake." For example, a real audio sample from Speaker 1 would be labeled as "S1_real," while a deepfake sample of Speaker 1's voice converted to Speaker 2's voice would be labeled as "S1_fake_S2." This labeling scheme ensures a clear and unambiguous identification of each sample's origin and nature, enabling the machine learning model to learn the distinct characteristics of both genuine and deepfake audio during training. The annotation process was conducted automatically to ensure the creation of a high-quality dataset that can effectively train a reliable deepfake audio detection model. The resulting annotated dataset comprises a balanced distribution of real and deepfake audio samples, featuring a wide range of speakers, dialects. This diversity is essential for ensuring that the trained model is robust to variations in speech patterns and manipulation techniques, enabling it to generalize effectively to real-world scenarios.

#### 5.1.1.4 Dataset Statistics

The final dataset collected for this study comprises collection of real and deepfake audio samples, totaling 500 unique recordings. The distribution of the dataset is as follows :

**Real Audio Samples :** 100 samples (5 speakers x 20 samples per speaker)

**Deepfake Audio Samples :** 400 samples (20 speakers x 20 deepfakes per speaker)

Each audio sample has a duration of 1 minute, resulting in a total dataset length of 400 minutes (6 hours and 40 minutes) of audio data. The dataset is evenly distributed across the five selected speakers, ensuring that each speaker is represented equally in both the real and deepfake categories. This balanced and diverse dataset is a crucial asset for training and evaluating our deepfake audio detection model. It provides a comprehensive representation of real-world speech patterns across different dialects and a wide array of deepfake manipulations generated using the RVC model. By training on such a diverse dataset, we aim to develop a model that can generalize effectively to unseen audio samples and accurately distinguish between genuine and synthetic speech. In the following sections, we will delve into the feature extraction process, discussing the specific acoustic, prosodic, and linguistic features that we extract from these audio samples to characterize and differentiate between real and deepfake speech. These features will serve as the input to our machine learning models, enabling us to build a robust and reliable deepfake audio detection system.

| **Category** | **Number of Samples** | **Duration per Sample** |
|---|---|---|
| Real Audio | 100 (5 speakers × 20 samples) | 1 minute |
| Deepfake Audio | 400 (20 speakers × 20 deepfakes) | 1 minute |
| **Total** | **500** | |

**Tableau 5.2:** Dataset Distribution

### 5.1.2 Contribution 2 : Proposed Framework For Deepfake Audio Detection

Building on the Islamic Scholars Arabic Audio Deepfake (ISAAD) dataset, we propose a new framework for detecting deepfake audio. Our framework leverages advanced Machine Learning (ML) techniques to achieve high detection accuracy and robustness. The key components of the proposed framework are outlined in Figure 5.1.

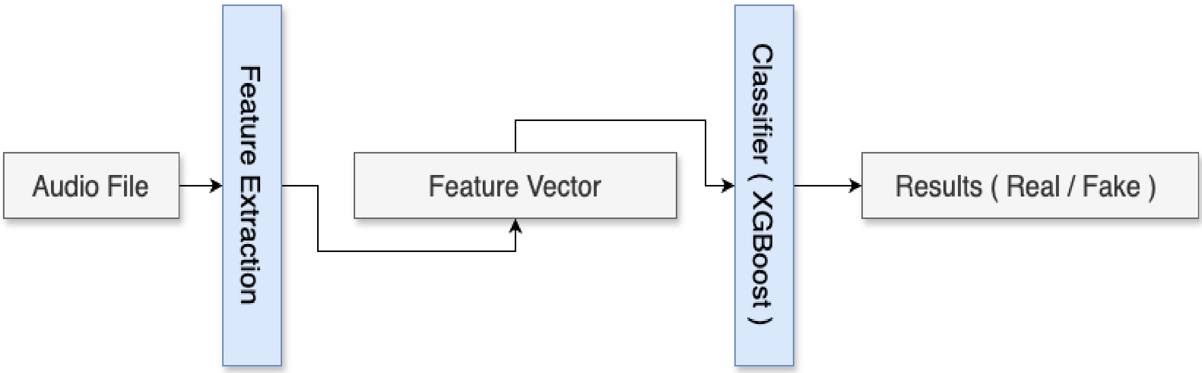

**Figure 5.1:** Proposed Framework Architecture For Deepfake Audio Detection.

#### 5.1.2.1 Feature Extraction

Feature extraction is a critical step in deepfake audio detection, as it involves transforming raw audio signals into meaningful numerical representations that can be used to train machine learning models. These features capture the essential characteristics of the audio, including spectral content, temporal patterns, and higher-level linguistic cues, making them invaluable for distinguishing between genuine and synthetic speech.

#### 5.1.2.2 Overview

The choice of features plays a pivotal role in the effectiveness of deepfake detection systems. Relevant features can capture the subtle artifacts and inconsistencies introduced by deepfake generation techniques, enabling the model to learn discriminative patterns between real and manipulated audio. In this study, we leverage a combination of low-level and high-level features, drawing inspiration from established practices in audio analysis and deepfake detection research.

#### 5.1.2.3 Feature Selection

The selection of features for this research was guided by a comprehensive review of the relevant literature. We focused on features that have been shown to be effective in previous studies on deepfake detection, as well as features that capture the specific characteristics of speech that are often manipulated in deepfakes. The chosen features can be broadly categorized into low-level feature and high-level features :

- **Low-Level Features :** These features capture the fundamental acoustic properties of the audio signal.
  - **Mel-Frequency Cepstral Coefficients (MFCCs) :** MFCCs are a widely used representation of the short-term power spectrum of a sound, based on a nonlinear Mel scale that approximates human auditory perception. They have been shown to be effective in capturing the spectral envelope of speech, which can reveal inconsistencies in synthetically generated audio.
  - **Spectral Features :** These features capture various aspects of the frequency distribution of the audio signal, such as spectral centroid, spectral bandwidth, spectral rolloff, and spectral flux. They provide insights into the tonal quality, energy distribution, and dynamics of the audio, which can be indicative of manipulation.
  - **Zero-Crossing Rate :** This feature measures the rate at which the audio signal crosses the zero-amplitude line. It is useful for distinguishing between voiced and unvoiced speech segments, which can be important for identifying unnatural transitions or discontinuities in deepfake audio.
- **High-Level Features :** These features capture higher-level linguistic and prosodic aspects of the audio signal.
  - **Prosodic Features :** These include pitch (fundamental frequency), intonation (variation in pitch), and energy (loudness). Variations in prosody can be indicative of emotional state or speaking style, and inconsistencies in these features can be a telltale sign of deepfake audio.
  - **Statistical Features :** These include measures like mean, variance, skewness, and kurtosis, which can capture the statistical distribution of various acoustic features. Deviations from expected statistical patterns can indicate anomalies in the audio signal.

These models were selected based on their proven effectiveness in classification tasks, their ability to handle tabular data, and their potential to capture complex patterns in the audio features.

#### 5.1.2.4 Feature Classification : XGBoost

After rigorous experimentation and evaluation, we found that XGBoost (Extreme Gradient Boosting) consistently outperformed other models on our dataset. XGBoost is an ensemble learning method that combines the predictions of multiple decision trees, leveraging the strengths of each individual tree to improve overall performance. This approach offers several advantages for deepfake audio detection :

- **High Accuracy** : XGBoost is known for its ability to achieve high accuracy on a wide range of classification tasks, including those with complex feature interactions.
- **Regularization** : It incorporates regularization techniques to prevent overfitting, ensuring that the model generalizes well to unseen data.
- **Handling Missing Values** : XGBoost can effectively handle missing values, which can be common in real-world audio datasets.
- **Feature Importance** : It provides insights into the relative importance of different features, which can aid in understanding the factors that contribute to deepfake detection.

Given XGBoost's strong performance and its suitability for tabular data, we selected it as the primary model for our deepfake audio detection system.

## 5.2 Experimental study

In this section, we present the experimental results obtained from the evaluation of our proposed framework for deepfake audio detection. We detail the experimental setup, dataset utilization, evaluation metrics, and comparative analysis with existing methods.

### 5.2.1 Training and Evaluation

To train and assess the performance of our proposed deepfake audio detection framework, we employed a rigorous methodology that involved model training, evaluation metrics, and cross-validation.

#### 5.2.1.1 Training Methodology

We utilized the XGBClassifier implementation from the XGBoost library in Python, initializing it with a fixed random_state for reproducibility. The model was trained on the extracted features (X_train) and corresponding labels (y_train), which indicate whether each sample is real or fake. For training, we opted for a straightforward approach, leveraging the default hyperparameters provided by the XGBoost library. These default settings often provide a good starting point and can yield competitive results in many cases.

#### 5.2.1.2 Evaluation Metrics

To comprehensively assess the performance of our framework, we employed a suite of evaluation metrics :

- **Accuracy :** The overall proportion of correctly classified samples (both real and fake).
$$Accuracy = \frac{Number\,of\,Correct\,Predictions}{Total\,Number\,of\,Predictions} \tag{5.1}$$
- **Mean Absolute Error (MAE) :** The average absolute difference between the predicted probabilities and the true labels.
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - y_i| \tag{5.2}$$
- **AUC (Area Under the ROC Curve) :** A measure of the model's ability to distinguish between real and fake samples across different classification thresholds.
$$AUC\,ROC = \int_0^1 ROC\,Curved\,(False\,Positive\,Rate) \tag{5.3}$$
- **MCC (Matthews Correlation Coefficient) :** A balanced measure of the quality of binary classifications, considering true and false positives and negatives.
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.4}$$
- **Precision :** The proportion of correctly predicted deepfakes out of all samples predicted as deepfakes.
$$Precision = \frac{True\,Positives\,(TP)}{True\,Positives\,(TP) + False\,Positives\,(FP)} \tag{5.5}$$
- **Recall :** The proportion of correctly predicted deepfakes out of all actual deepfakes.

$$Recall = \frac{True\,Positives\,(TP)}{True\,Positives\,(TP) + False\,Negatives\,(FN)} \tag{5.6}$$

- **F1-score :** The harmonic mean of precision and recall, providing a single metric that balances both.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.7}$$

These metrics provide a comprehensive view of the model's performance, considering both its overall accuracy and its ability to correctly identify deepfakes without misclassifying genuine audio.

#### 5.2.1.3 Evaluation protocol : Cross-Validation

To ensure the robustness and generalizability of our model, we employed 5-fold cross-validation. This technique involves dividing the dataset into 5 folds, training the model on 4 folds (80%), and evaluating it on the remaining fold (20%). This process is repeated 5 times, with each fold serving as the test set once. The final performance metrics are then averaged across all folds, providing a more reliable estimate of the model's performance on unseen data. By incorporating cross-validation, we can assess the model's ability to perform consistently across different data splits and reduce the risk of overfitting to the training set. This enhances the model's generalizability and ensures its effectiveness in real-world scenarios.

## 5.3 Experimental Results

This section presents our deepfake audio detection results for Arabic speech. We recap Chapter 1's objectives (robust detection system, Arabic dataset) and Chapter 4's methodology (feature extraction, XGBoost model). We showcase model performance, emphasizing XGBoost's effectiveness. We then discuss our contributions, including the novel Arabic deepfake dataset and the potential benchmark for Arabic speech deepfake detection. Finally, we address strengths, limitations, and future directions.

### 5.3.1 Experimental Setup

The experimental setup for this research involved a combination of hardware and software tools to facilitate both the generation of deepfake audio samples and the training and evaluation of the detection model.

#### 5.3.1.1 Hardware

- **Deepfake Generation** : We utilized the computational power of an NVIDIA Tesla T4 GPU hosted on Google Colab. This high-performance GPU accelerated the deepfake generation process, enabling us to efficiently create a large number of synthetic audio samples using the RVC model.
- **Model Training and Evaluation** : The training and evaluation of our XGBoost classifier were conducted on an Apple M1 PRO CPU. The M1 PRO chip's powerful machine learning capabilities allowed us to efficiently train and validate our model on the extracted audio features.

#### 5.3.1.2 Software and Libraries

- **Data Preprocessing, Feature Extraction, and Model Training** : For these tasks, we utilized the JetBrains DataSpell integrated development environment (IDE), which provides a convenient interface for data science workflows.
- **Machine Learning Framework** : We primarily relied on the TensorFlow framework for implementing our XGBoost classifier, leveraging its extensive machine learning capabilities and optimized performance.

#### 5.3.1.3 Implementation Details

For feature extraction, we utilized the librosa library, a popular Python package for audio and music analysis. Librosa provides a comprehensive set of functions for loading, visualizing, and analyzing audio signals. It also offers a wide range of built-in feature extraction functions, making it an ideal tool for our research. By leveraging the capabilities of librosa, we were able to efficiently extract a diverse set of low-level and high-level features from our dataset, capturing the nuances and complexities of both real and deepfake audio. These extracted features formed the foundation for training and evaluating our machine learning models for deepfake audio detection.

### 5.3.2 Performance of Different Models

To rigorously assess the effectiveness of various machine learning models in detecting deepfake Arabic audio, we conducted a comprehensive evaluation on our custom-created dataset. We trained and tested several models, including traditional algorithms, ensemble methods :

- Traditional Algorithms :
  - Logistic Regression
  - Decision Trees
  - Support Vector Machines (SVM)
  - Naive Bayes
  - K-Nearest Neighbors (KNN)
- Ensemble Methods :
  - Random Forest
  - Gradient Boosting Machines :
    - XGBoost
    - LightGBM

As evident from Table 5.3, the XGBoost model consistently outperformed all other models across all evaluation metrics. It achieved the highest accuracy, precision, recall, F1-score, AUC, and MCC, indicating its superior ability to discriminate between real and deepfake Arabic audio.

In particular, XGBoost's ensemble learning approach, which combines the predictions of multiple decision trees, seems to be particularly effective in capturing the complex patterns and nuances present in the audio features. Furthermore, its built-in regularization techniques help prevent overfitting, ensuring that the model generalizes well to unseen data.

**Tableau 5.3:** Overall Performance of Different Models (5-fold Cross-Validation)

| **Model** | **Accuracy** | **MAE** | **AUC** | **MCC** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|---|---|---|
| XGBoost | 0.9965 | 0.0035 | 0.9965 | 0.9930 | 0.9945 | 0.9985 | 0.9965 |
| LightGBM | 0.9956 | 0.0044 | 0.9956 | 0.9913 | 0.9940 | 0.9973 | 0.9956 |
| Random Forest | 0.9943 | 0.0057 | 0.9943 | 0.9886 | 0.9951 | 0.9935 | 0.9943 |
| Decision Tree | 0.9463 | 0.0537 | 0.9463 | 0.8926 | 0.9431 | 0.9500 | 0.9465 |
| k-Nearest Neighbors | 0.7965 | 0.2035 | 0.7965 | 0.5968 | 0.7668 | 0.8524 | 0.8073 |
| Naive Bayes | 0.7610 | 0.2390 | 0.7610 | 0.5380 | 0.7105 | 0.8814 | 0.7867 |
| Logistic Regression | 0.8817 | 0.1183 | 0.8817 | 0.7640 | 0.8688 | 0.8994 | 0.8838 |
| SVM | 0.6818 | 0.3182 | 0.6818 | 0.3737 | 0.6480 | 0.7966 | 0.7146 |

# Conclusion

This chapter has presented the framework for our Arabic deepfake audio detection system, emphasizing the importance of a custom dataset. We detailed the creation process of our new Arabic deepfake dataset, which serves as the foundation for training and evaluating our model.

# General Conclusion

In this thesis, we have explored the complex challenge of deepfake audio detection, focusing on the unique context of the Arabic language. Through the creation of a new Arabic deepfake dataset and the development of an effective framework deepfake audio detection model, we have made significant steps in advancing the field. Our research has not only demonstrated the feasibility of detecting deepfake Arabic audio with high accuracy but has also shed light on the specific features and techniques that are most effective in distinguishing real from manipulated speech.

The implications of our findings extend beyond the technical world. By contributing to the development of reliable deepfake audio detection tools, we empower individuals and organizations to better protect themselves from the potential harms of this technology. Our research also has broader societal implications, as it helps to safeguard trust in audio information and combat the spread of misinformation, which is crucial for maintaining a healthy democracy and informed public discourse.

## Strengths and Limitations

### Strengths

Our approach to deepfake audio detection in Arabic exhibits several notable strengths, stemming primarily from our contributions to the field :

- **Novel Arabic Dataset :** The creation of a custom Arabic deepfake audio dataset addresses a critical gap in existing resources, providing a valuable tool for training and evaluating detection models specifically tailored for the Arabic language. This dataset's diversity in speakers, dialects, and voice conversion techniques enhances the model's ability to generalize to real-world scenarios.

- **Effective XGBoost Model :** The XGBoost model, with its ensemble learning approach and regularization capabilities, has demonstrated exceptional performance in accurately classifying real and deepfake Arabic audio samples. This highlights the model's potential as a benchmark for Arabic speech deepfake detection.

- **Interpretability and Feature Importance :** The XGBoost model provides insights into the relative importance of different features, allowing us to identify the most discriminative acoustic, prosodic, and linguistic cues for deepfake detection. This information can be leveraged to refine feature engineering techniques and improve the explainability of the model's decision-making process.
- **Robustness and Generalization in Arabic speech :** Our model has shown robustness to variations in speaker characteristics and deepfake generation techniques in Arabic speech, demonstrating its potential for real-world deployment. The use of cross-validation further enhances the model's ability to generalize to unseen data.

### Limitations

While our approach offers promising results, there are some limitations that warrant consideration :

- **Dataset Bias :** The reliance on publicly available YouTube recordings for our dataset might introduce biases in terms of the types of speech, speaking styles, and recording quality represented. Expanding the dataset with more diverse and controlled samples could further improve the model's performance and generalization capabilities.
- **Hyperparameter Optimization :** While the XGBoost model performed well with default hyperparameters, there is potential for further improvement through systematic hyperparameter tuning. Exploring different combinations of learning rates, tree depths, and regularization parameters could lead to even higher accuracy and more robust detection.
- **Exploration of Alternative Architectures :** While XGBoost proved effective, other deep learning architectures, such as convolutional neural networks (CNNs) or transformers, have shown promise in audio analysis tasks. Exploring these alternative models could potentially reveal additional insights and improve detection performance.

By acknowledging these limitations, we can identify areas for future research and development, ultimately leading to more robust and generalizable deepfake audio detection methods for the Arabic language and beyond.

While our work represents a significant step forward, it is by no means the final word on deepfake audio detection. The ever-evolving nature of deepfake technology necessitates continuous research and development to stay ahead of the curve. Future research should focus on expanding and diversifying datasets, exploring new model architectures and feature extraction techniques, and addressing the challenges of generalization and adversarial attacks.

By fostering collaboration between researchers, policymakers, and industry stakeholders, we can collectively work towards a future where the authenticity and trustworthiness of audio information can be reliably verified, mitigating the harmful effects of deepfakes and ensuring the integrity of communication in the digital age.

# Bibliographie

[1] Jessica Toonkel. *Trust No One : Inside the World of Deepfakes.* Hodder Stoughton, 2022.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv :1706.03762*, 2017.

[3] Dmitriy Smirnov, Michael Gharbi, Matthew Fisher, Vitor Guizilini, Alexei A. Efros, and Justin Solomon. Marionette : Self-supervised sprite learning, 2021.

[4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[5] Nitesh Kumar Gaur, editor. *DeepFakes : Creation, Detection, and Impact.* Routledge, 2023.

[6] Graham Meikle. *Deepfakes.* Polity Press, 2020.

[7] Oliver Theobald. *Machine Learning for Absolute Beginners : A Plain English Introduction.* Independently published, first edition, 2016.

[8] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, 2nd edition, 2019.

[9] Charu C. Aggarwal. *Neural Networks and Deep Learning : A Textbook.* Springer, 2018.

[10] Andreas C. Müller and Sarah Guido. *Introduction to Machine Learning with Python : A Guide for Data Scientists.* O'Reilly Media, 2016.

[11] Aston Zhang, Zachary C. Lipton, Mu Li, and Alex J. Smola. *Dive into Deep Learning.* D2L.ai, 2021.

[12] Francois Chollet. *Deep Learning with Python (First Edition).* Manning Publications, 2017.

[13] David Foster. *Generative Deep Learning : Teaching Machines to Paint, Write, Compose, and Play.* O'Reilly Media, 2019.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[15] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4) :307–392, 2019.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[17] Jisu Kim, Laurent Elghaoui, Mehrdad Farajtabar, Yuandong Tian, Trevor Xiao, Lisha Chen, and Zaiwei Hong. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5561–5571. PMLR, 2021.

[18] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe : A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018.

[19] Panos Photinos. *The Physics of Sound Waves (Second Edition).* 2053-2563. IOP Publishing, 2021.

[20] Allen B. Downey. *Think DSP : Digital Signal Processing in Python.* Green Tea Press, Needham, Massachusetts, 2014. Copyright © 2014 Allen B. Downey.

[21] Jonathan M. Blackledge. *Digital Signal Processing.* Elsevier, second edition edition, 2006.

[22] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa : Audio and music signal analysis in python. In *SciPy*, 2015.

[23] Ken Pohlmann. *Principles of Digital Audio.* McGraw-Hill, 2000.

[24] Udo Zölzer. *Digital Audio Signal Processing.* Wiley, 2008.

[25] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *2013 IEEE Sixth International Conference on Biometrics : Theory, Applications and Systems (BTAS)*, pages 1–8, 2013.

[26] Jordan J. Bird and Ahmad Lotfi. Real-time detection of ai-generated speech for deepfake voice conversion, 2023.

[27] Awais Khan and Khalid Mahmood Malik. Securing voice biometrics : One-shot learning approach for audio deepfake detection, 2023.

[28] Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks, 2020.

[29] Moustafa Alzantot, Ziqi Wang, and Mani Srivastava. Deep residual neural networks for audio spoofing detection, 2019.

[30] Prasanth Parasu, Julien Epps, Kaavya Sriskandaraja, and Gajan Suthokumar. Investigating Light-ResNet Architecture for Spoofing Detection Under Mismatched Conditions. In *Proc. Interspeech 2020*, pages 1111–1115, 2020.