AI-generated Voice Recognition

Project Charter Document

Ynon Friedman- 207498437
Guy Ben Ari - 209490473

Afeka Engineering College

אפקה
המכללה האקדמית
להנדסה בתל אביב

Instructor - Revital Marom Elgrabli

Table of Contents

1. Executive summary

Summary of the Problem:

The prevalence of fake voice audio in today's digital landscape poses a significant challenge. With the advancements in artificial intelligence and voice synthesis technologies, it has become increasingly difficult to distinguish between genuine and fabricated voice recordings. This creates a pressing need for a robust solution that can effectively recognize and authenticate voice content to ensure trust and accuracy in voice-based systems and applications.

Project Essence:

The essence of the project is to develop an advanced AI system capable of recognizing fake voice audio with a high degree of accuracy. Leveraging cutting-edge machine learning techniques and sophisticated algorithms, the project aims to analyze voice recordings and detect signs of manipulation or synthesis. By doing so, the project seeks to provide a comprehensive solution to mitigate the risks associated with fake voice audio, enabling organizations and individuals to authenticate and verify the authenticity of voice-based content effectively.

Planned System:

The project, titled "AI Voice Authenticator," will involve the creation of a powerful and user-friendly system that utilizes AI algorithms, deep learning models, and extensive training datasets. The system will analyze various characteristics and patterns within voice recordings, including speech patterns, pitch modulation, and spectral analysis, among others. By comparing these features against a database of genuine voice samples, the system will identify anomalies or inconsistencies indicative of fake voice audio.

The AI Voice Authenticator will be designed as a scalable and accessible solution, allowing users to easily upload voice recordings for analysis. Upon submission, the system will provide real-time feedback on the authenticity of the audio, offering a confidence score or probability of the recording being genuine or manipulated. Furthermore, the system will incorporate continuous learning capabilities to adapt and stay up to date with emerging techniques used in generating fake voice audio.

In conclusion, the AI Voice Authenticator project addresses the pressing need for an advanced AI system to combat the increasing prevalence of fake voice audio. By accurately detecting and authenticating voice recordings, the project aims to restore trust and reliability in voice-based communication and applications. With the integration of cutting-edge technologies and rigorous research, the AI Voice Authenticator project strives to make a significant impact in safeguarding the integrity of voice-based content.

2. Background, the problem and challenges, the existing state that requires a systematic solution

In today’s age we can see more and more examples of usage of AI in order to fake the voices of people or celebrities. This phenomena is common in the internet culture where use of free tools (such as fakeyou.com) in order to apply the voice of famous celebs or fictional characters from movies and tv to create sketches, jokes and memes in the best cases. And in the worst cases creating hate speech and offensive content. This advancement of technology also places the security system of many organizations at risk, as people with dark intentions could use this technology to fake the voice of someone from upper management or another important person and break into the system with it.

There are a few existing technologies of detecting fake speech, and their origin are in competitions whose goal is to solve the defensive deficit (for example, ASVspoof), these technologies are based on a wide range of ideas and models and reach different levels of accuracy for different faking tools.

To recognise voice that was created by a machine/AI, via any tool as a ‘catch all’, in our solution we will attempt to find and create a system that will function as a generalist solution. We will further tweak the system we create so it will be able to function as a security layer that can be inserted into existing large systems to defend them from faked-voice based attacks or scams. We’ll additionally demand that our system could correctly flag faked speech even if the recording sample has background noise or other factors that could inhibit the process, such as a very thick accent.

3. Objectives, Metrics, and Goals (Project's objectives)

As a direct derivative of defining the gaps, problems, and requirements for a solution, this section defines the objectives of the system that will characterize and be established within the framework of the final project.
The objectives will be detailed with an emphasis on three components:

3.1.1 - Objective

Preventing "spoof" attacks, which can pose significant harm to both the general population, including adults and children, as well as companies and organizations.

Main Goal: Creating an artificial intelligence-based system designed to identify whether speech is generated by a machine or by a human.

3.1.2 - Goals

A. Establishing a reliable model.

B. Ensuring adaptability to system improvements according to technological advancements in the industry.

C. Enhancing the model for the existing world.

D. Conducting extensive testing on the model using diverse datasets whenever possible.

E. Safeguarding users against "spoof" attacks.

F. Training the model on the existing dataset.

3.1.3 - Metrics

A. Accuracy of predictions or model's overall performance should be above 85% (precision, recall, F1 score).

B. Model's compatibility with new technologies or version updates, measured by successful integration and minimal disruptions during system upgrades.

C. Model's performance on real-world data, such as accuracy and robustness in diverse scenarios.

D. Model's performance on various datasets, measured by accuracy, generalization capability, and performance consistency.

E. Robustness of the model against spoofing attempts, measured by its ability to detect and prevent fraudulent or malicious inputs.

F. Training performance and convergence, measured by metrics like loss reduction, convergence speed, and the model's ability to learn from the given dataset.

4. Methodology:

This chapter outlines the engineering methods and tools required for the design and implementation of the planned system. It includes:

4.1. Functional Requirements

A. Voice Classification: The system needs to accurately classify sounds as either human speech or spoofed audio.

B. Real-time Processing: The system should be capable of processing voice inputs in real-time, providing prompt results for timely decision-making.

C. Voice Screening: The system should be able to filter out background noises present in audio files.

D. One System fits all: The system should be adaptable to various voice inputs, including different languages, accents, speech styles, and different audio file formats (mp3, mp4, wav, etc.).

E. Edge-case Handling: The system needs to handle errors, such as low-quality audio or distorted speech, to ensure reliable and accurate voice classification.

F. Multilingual Support: The system should be able to handle voice inputs in multiple languages, allowing for accurate classification and processing.

G. Speaker Identification: The system should have the capability to identify and distinguish between different speakers, enabling speaker-specific analysis and recognition.

H. Integration with External Systems: The system should be designed to integrate with other external systems or applications, allowing for seamless data exchange and interoperability.

4.2. Non-Functional Requirements

A. Accuracy: The system should achieve a high level of accuracy in voice classification, minimizing false positives (FP) and false negatives (FN) results.

B. Reliability: The system should be sufficiently reliable to provide services to important entities, assisting in user identification and verification.

C. Availability: The system should be reliable, with high availability and minimal downtime, ensuring continuous voice recognition services without disruptions.

D. Performance: The system should have optimal performance, with low retrieval and response times.

E. Usability: The system should be user-friendly, with intuitive interfaces and easy integration into existing workflows or systems, making it accessible and convenient for users to interact with.

F. Scalability: The system should be scalable to accommodate increasing volumes of voice data and growing user demands without compromising performance or reliability.

G. Security: The system should implement robust security measures to protect voice data, ensure confidentiality, prevent unauthorized access, and mitigate potential vulnerabilities or attacks.

H. Ethical Considerations: The system should adhere to ethical principles and guidelines, respecting privacy rights, avoiding biases or discrimination, and promoting fairness and transparency in its operations.

I. Maintainability: The system should be designed with ease of maintenance in mind, allowing for efficient updates, bug fixes, and enhancements to ensure long-term sustainability and support.

J. Interoperability: The system should be compatible with different hardware and software environments, enabling seamless integration with various platforms or devices.

K. Compliance: The system should comply with relevant regulations, industry standards, and data protection laws to ensure legal and ethical use of voice data.

L. Performance Monitoring and Analytics: The system should incorporate monitoring mechanisms and analytics capabilities to track performance, identify bottlenecks, and optimize system efficiency and effectiveness.

M. Documentation: The system should be accompanied by comprehensive documentation, including user manuals, technical specifications, and system documentation, to aid users and administrators in understanding and operating the system effectively.

4.2. Implementation of Profiling and Planning Methods

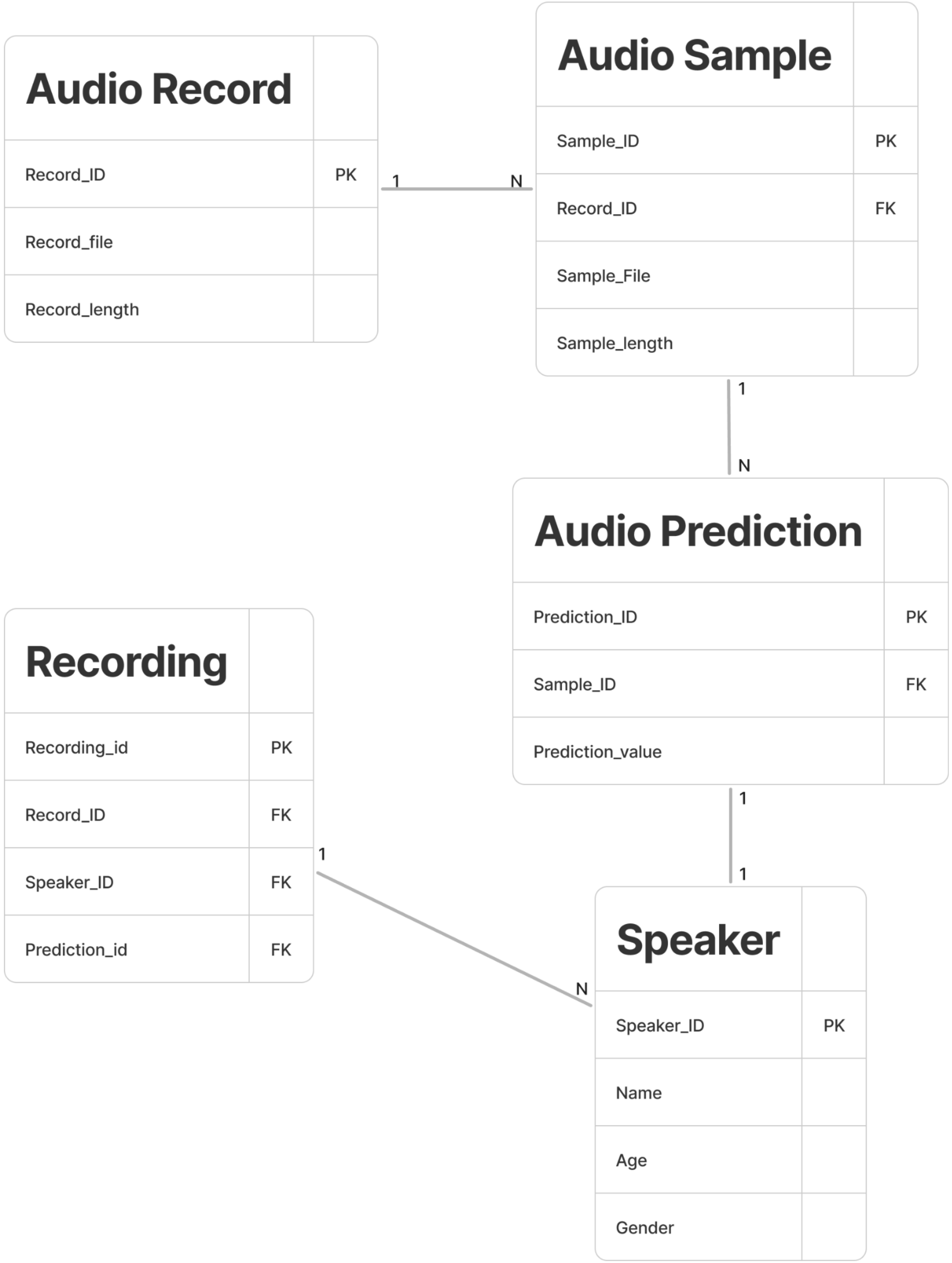
Requirements Analysis:Conduct stakeholder interviews and analyze existing voice authentication systems to gather functional and non-functional requirements.

System Design:Develop a comprehensive system design, including component identification, data flow, and interfaces, utilizing industry-standard design techniques.

Development and Coding:Implement the voice authentication system using selected programming languages and frameworks, following an agile development approach.

Testing and Quality Assurance:Employ a comprehensive testing strategy, including, integration, system, and acceptance testing, to ensure accuracy and reliability

4.2.1. ERD diagram



Audio Record	
Record_ID	PK
Record_file	
Record_length	

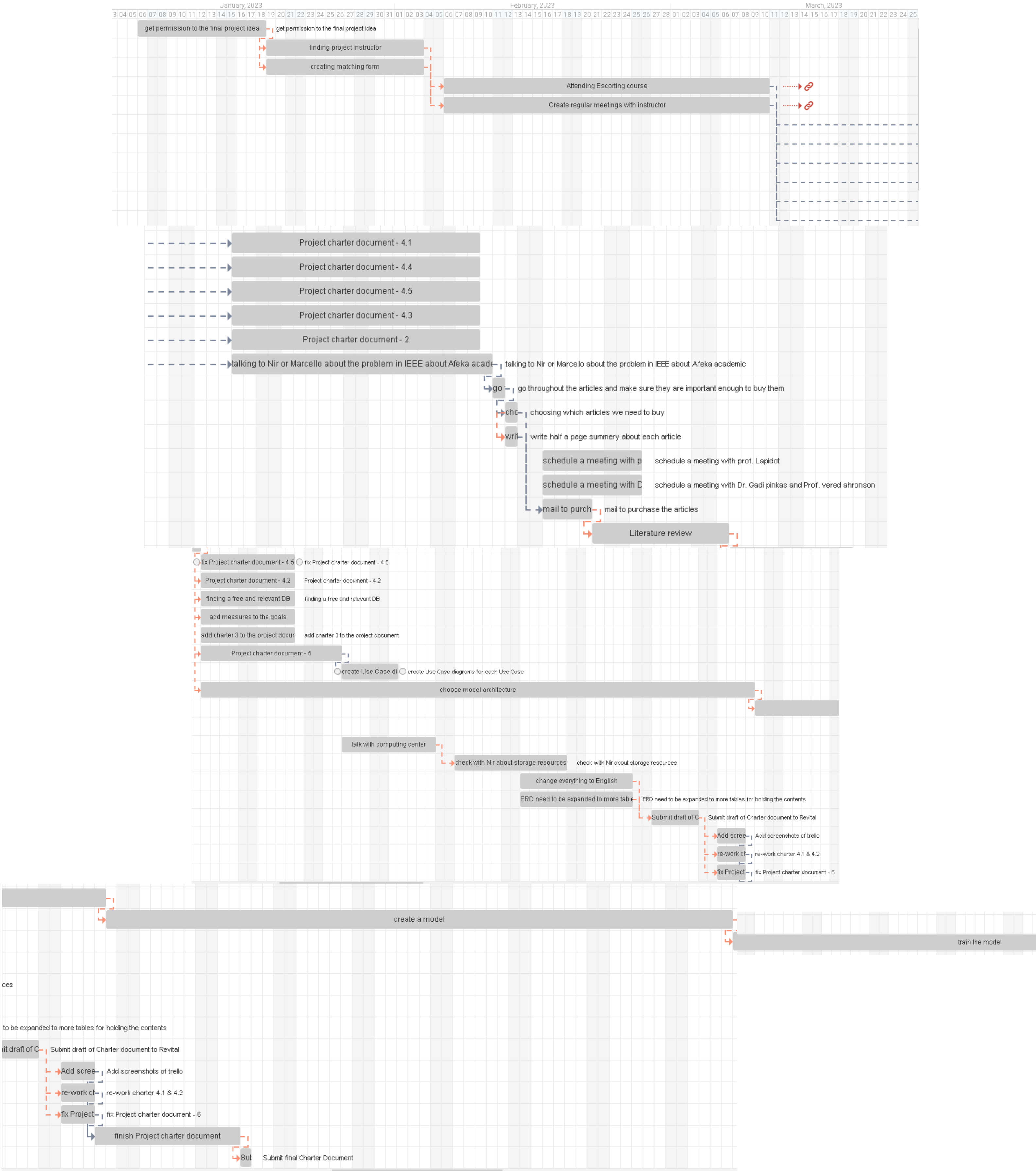
Audio Sample	
Sample_ID	PK
Record_ID	FK
Sample_File	
Sample_length	

Audio Prediction	
Prediction_ID	PK
Sample_ID	FK
Prediction_value	

Recording	
Recording_id	PK
Record_ID	FK
Speaker_ID	FK
Prediction_id	FK

Speaker	
Speaker_ID	PK
Name	
Age	
Gender	

4.3. GANTT chart



January, 2023

February, 2023

March, 2023

3 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

get permission to the final project idea — get permission to the final project idea

finding project instructor

creating matching form

Attending Escorting course

Create regular meetings with instructor

Project charter document - 4.1

Project charter document - 4.4

Project charter document - 4.5

Project charter document - 4.3

Project charter document - 2

talking to Nir or Marcello about the problem in IEEE about Afeka acade — talking to Nir or Marcello about the problem in IEEE about Afeka academic

go — go throughout the articles and make sure they are important enough to buy them

cho — choosing which articles we need to buy

wri — write half a page summery about each article

schedule a meeting with p — schedule a meeting with prof. Lapidot

schedule a meeting with D — schedule a meeting with Dr. Gadi pinkas and Prof. vered ahronson

mail to purch — mail to purchase the articles

Literature review

fix Project charter document - 4.5 — fix Project charter document - 4.5

Project charter document - 4.2 — Project charter document - 4.2

finding a free and relevant DB — finding a free and relevant DB

add measures to the goals

add charter 3 to the project docur — add charter 3 to the project document

Project charter document - 5

create Use Case di — create Use Case diagrams for each Use Case

choose model architecture

talk with computing center

check with Nir about storage resources — check with Nir about storage resources

change everything to English

ERD need to be expanded to more table — ERD need to be expanded to more tables for holding the contents

Submit draft of C — Submit draft of Charter document to Revital

Add scree — Add screenshots of trello

re-work ch — re-work charter 4.1 & 4.2

fix Project — fix Project charter document - 6

create a model

train the model

ces

to be expanded to more tables for holding the contents

it draft of C — Submit draft of Charter document to Revital

Add scree — Add screenshots of trello

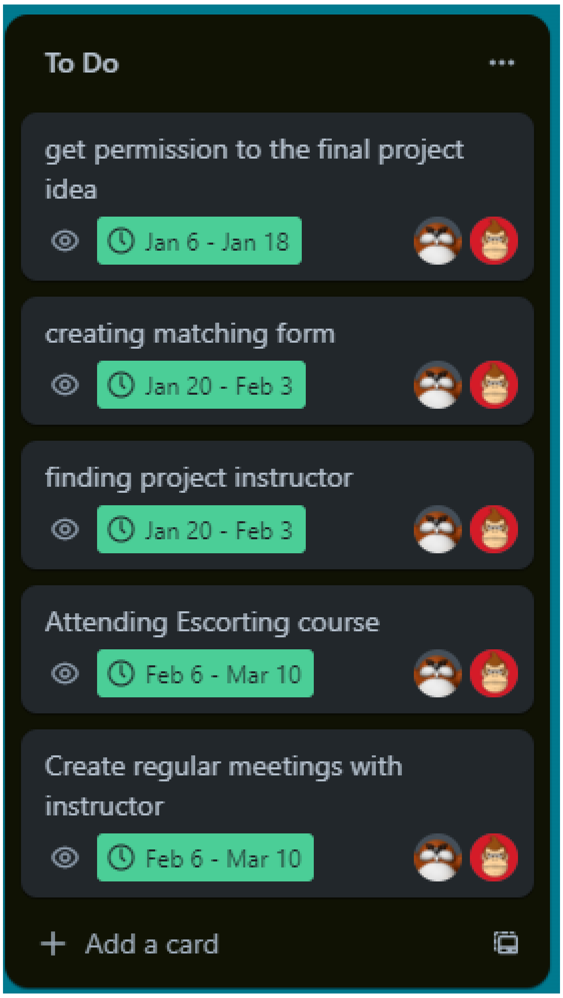
re-work ch — re-work charter 4.1 & 4.2

fix Project — fix Project charter document - 6

finish Project charter document

Sul — Submit final Charter Document

4.4. Trello list



To Do

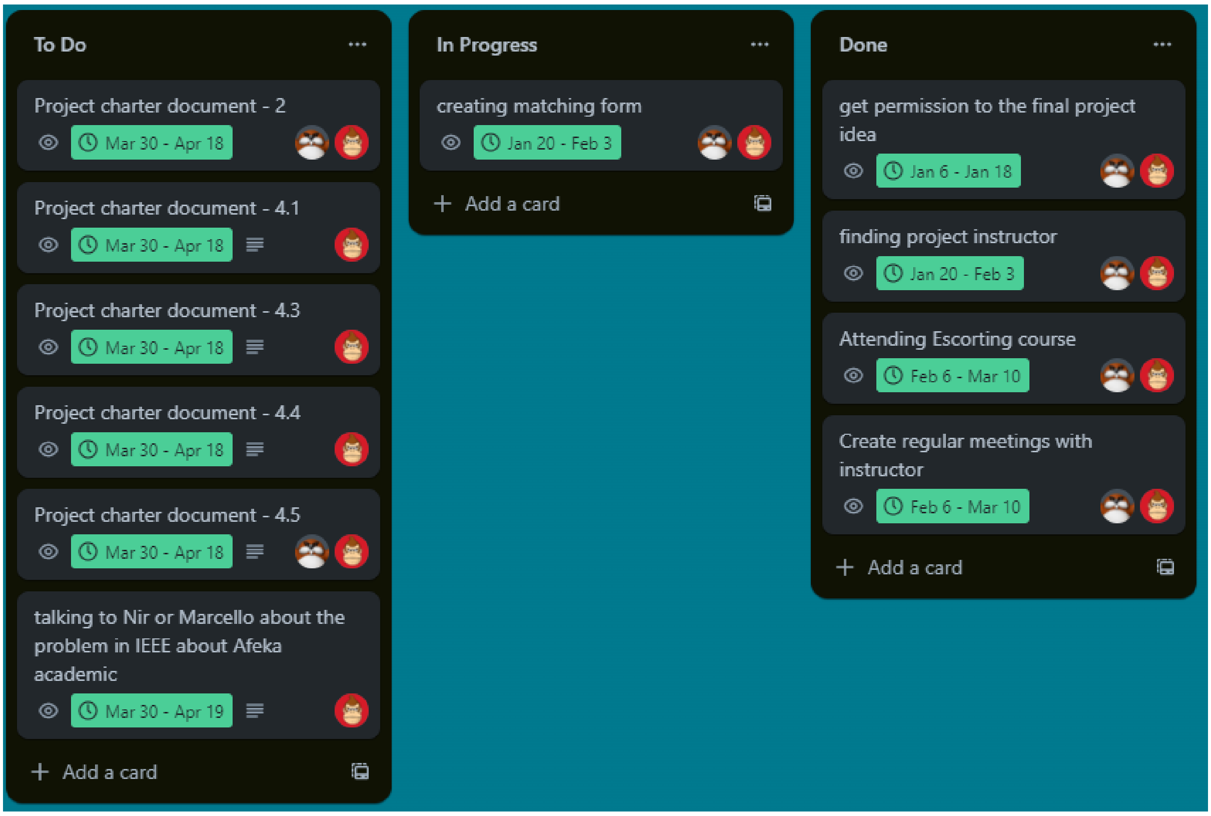
get permission to the final project idea
Jan 6 - Jan 18

creating matching form
Jan 20 - Feb 3

finding project instructor
Jan 20 - Feb 3

Attending Escorting course
Feb 6 - Mar 10

Create regular meetings with instructor
Feb 6 - Mar 10



To Do

Project charter document - 2
Mar 30 - Apr 18

Project charter document - 4.1
Mar 30 - Apr 18

Project charter document - 4.3
Mar 30 - Apr 18

Project charter document - 4.4
Mar 30 - Apr 18

Project charter document - 4.5
Mar 30 - Apr 18

talking to Nir or Marcello about the problem in IEEE about Afeka academic
Mar 30 - Apr 19

In Progress

creating matching form
Jan 20 - Feb 3

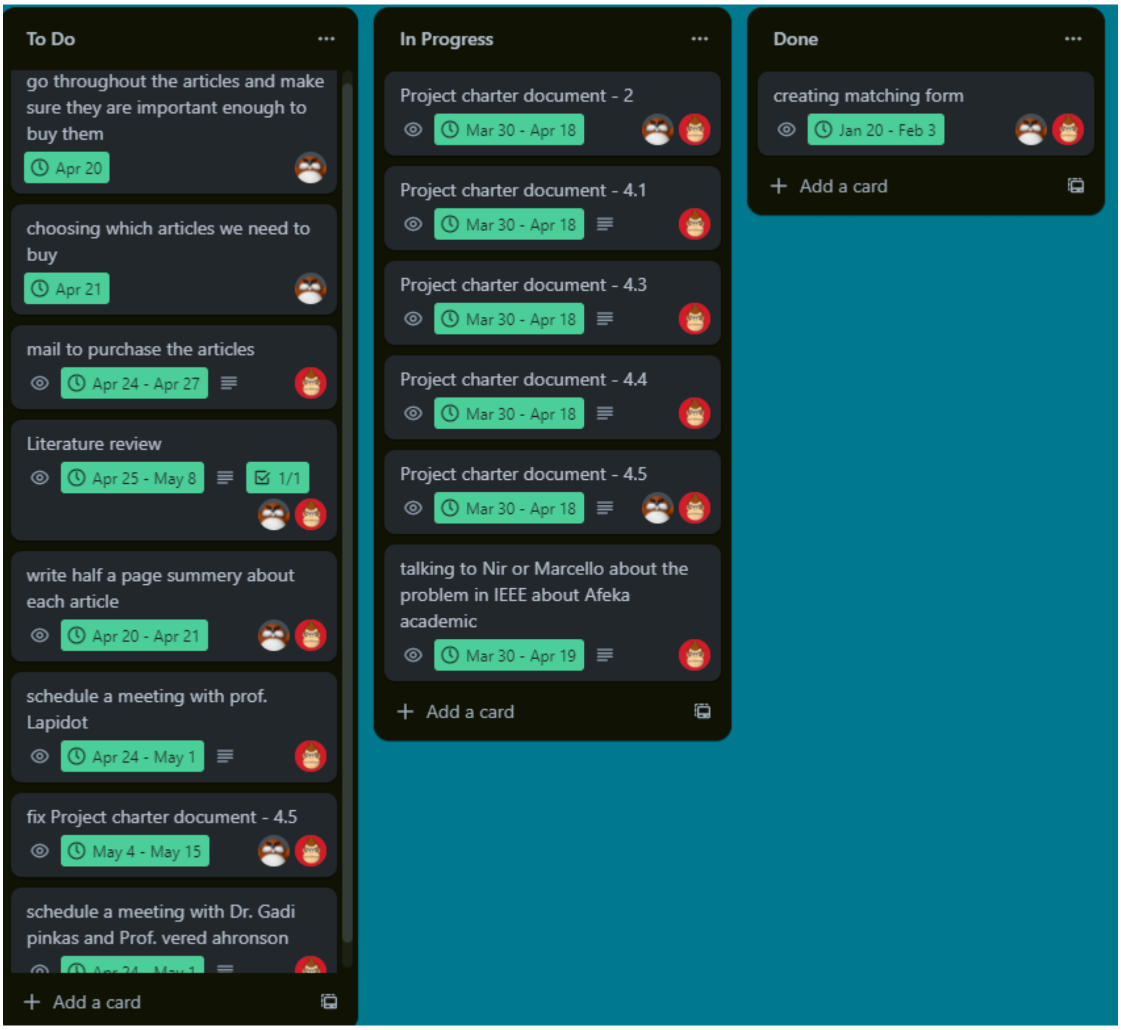
Done

get permission to the final project idea
Jan 6 - Jan 18

finding project instructor
Jan 20 - Feb 3

Attending Escorting course
Feb 6 - Mar 10

Create regular meetings with instructor
Feb 6 - Mar 10



To Do

go throughout the articles and make sure they are important enough to buy them
Apr 20

choosing which articles we need to buy
Apr 21

mail to purchase the articles
Apr 24 - Apr 27

Literature review
Apr 25 - May 8 ☑ 1/1

write half a page summery about each article
Apr 20 - Apr 21

schedule a meeting with prof. Lapidot
Apr 24 - May 1

fix Project charter document - 4.5
May 4 - May 15

schedule a meeting with Dr. Gadi pinkas and Prof. vered ahronson
Apr 24 - May 1

+ Add a card

In Progress

Project charter document - 2
Mar 30 - Apr 18

Project charter document - 4.1
Mar 30 - Apr 18

Project charter document - 4.3
Mar 30 - Apr 18

Project charter document - 4.4
Mar 30 - Apr 18

Project charter document - 4.5
Mar 30 - Apr 18

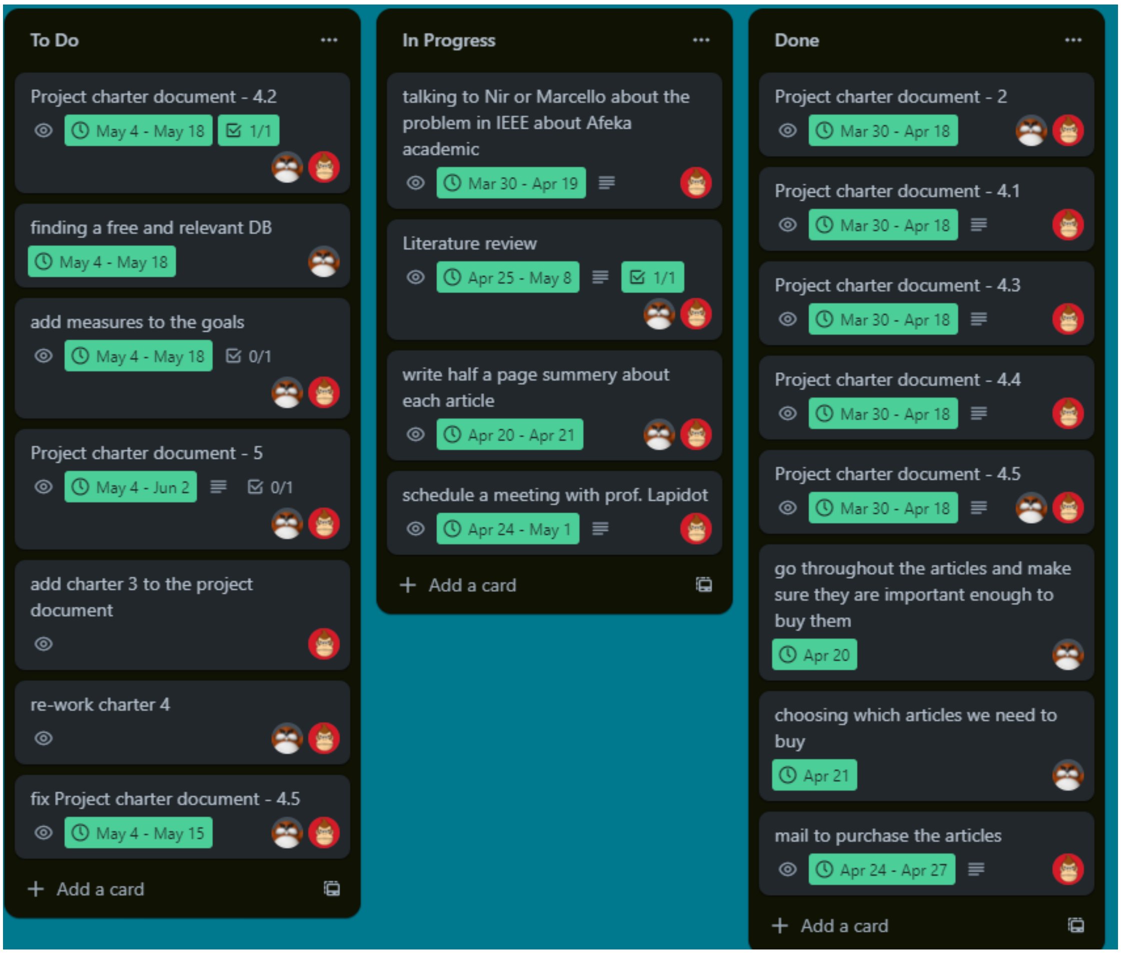
talking to Nir or Marcello about the problem in IEEE about Afeka academic
Mar 30 - Apr 19

+ Add a card

Done

creating matching form
Jan 20 - Feb 3

+ Add a card



To Do

Project charter document - 4.2
May 4 - May 18 ☑ 1/1

finding a free and relevant DB
May 4 - May 18

add measures to the goals
May 4 - May 18 ☑ 0/1

Project charter document - 5
May 4 - Jun 2 ☑ 0/1

add charter 3 to the project document

re-work charter 4

fix Project charter document - 4.5
May 4 - May 15

+ Add a card

In Progress

talking to Nir or Marcello about the problem in IEEE about Afeka academic
Mar 30 - Apr 19

Literature review
Apr 25 - May 8 ☑ 1/1

write half a page summery about each article
Apr 20 - Apr 21

schedule a meeting with prof. Lapidot
Apr 24 - May 1

+ Add a card

Done

Project charter document - 2
Mar 30 - Apr 18

Project charter document - 4.1
Mar 30 - Apr 18

Project charter document - 4.3
Mar 30 - Apr 18

Project charter document - 4.4
Mar 30 - Apr 18

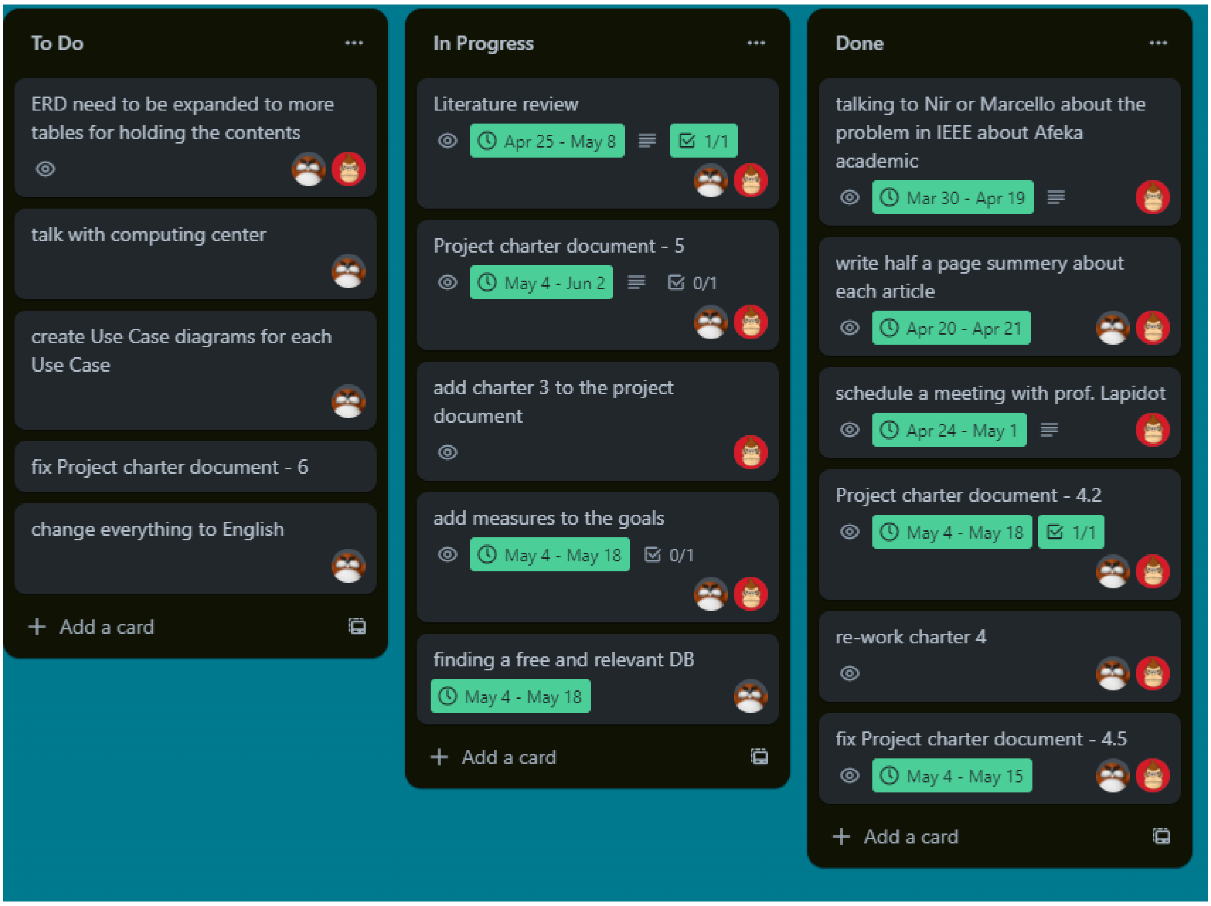
Project charter document - 4.5
Mar 30 - Apr 18

go throughout the articles and make sure they are important enough to buy them
Apr 20

choosing which articles we need to buy
Apr 21

mail to purchase the articles
Apr 24 - Apr 27

+ Add a card



To Do

- ERD need to be expanded to more tables for holding the contents
- talk with computing center
- create Use Case diagrams for each Use Case
- fix Project charter document - 6
- change everything to English

+ Add a card

In Progress

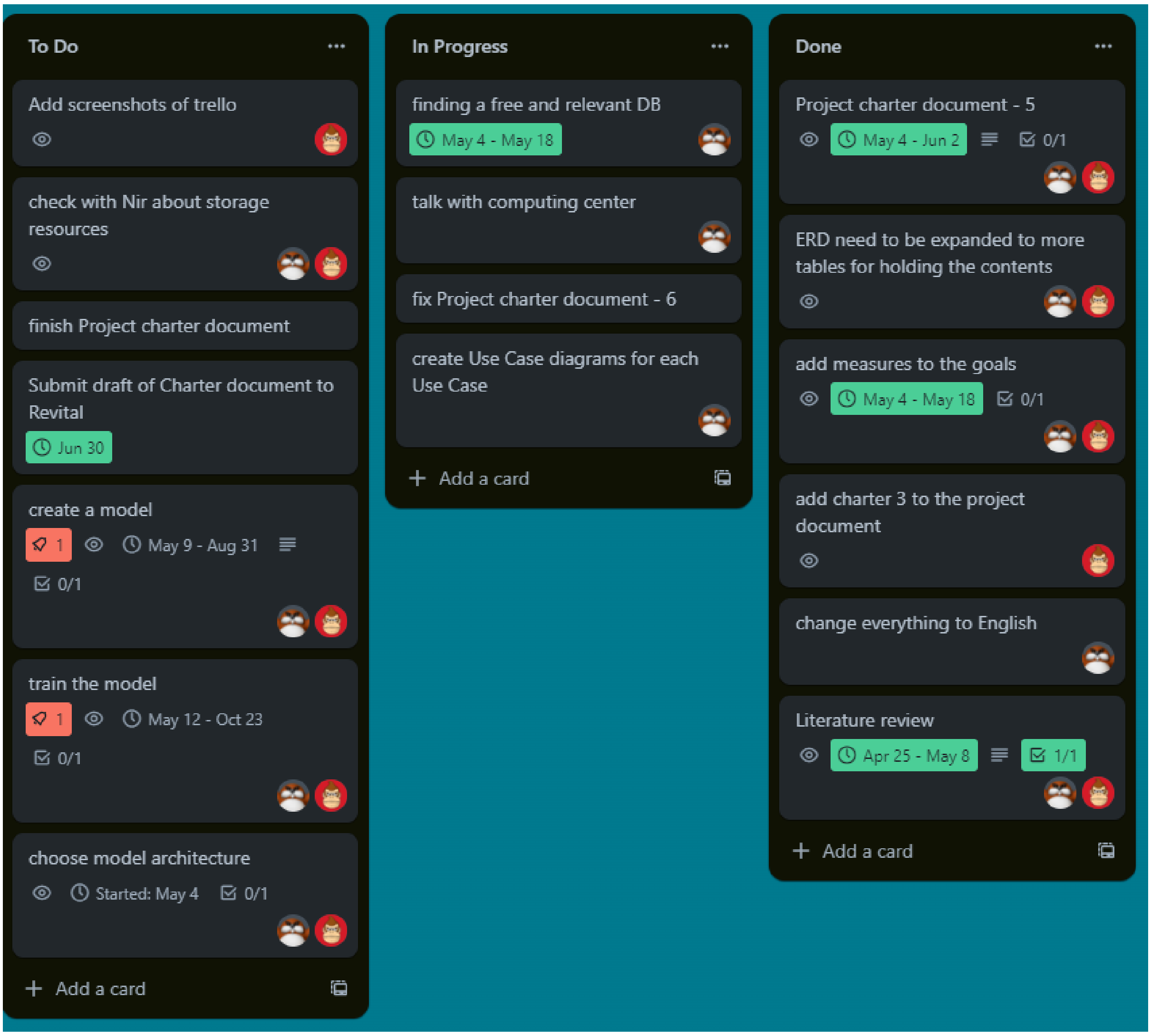
- Literature review — Apr 25 - May 8 — ☑ 1/1
- Project charter document - 5 — May 4 - Jun 2 — ☑ 0/1
- add charter 3 to the project document
- add measures to the goals — May 4 - May 18 — ☑ 0/1
- finding a free and relevant DB — May 4 - May 18

+ Add a card

Done

- talking to Nir or Marcello about the problem in IEEE about Afeka academic — Mar 30 - Apr 19
- write half a page summery about each article — Apr 20 - Apr 21
- schedule a meeting with prof. Lapidot — Apr 24 - May 1
- Project charter document - 4.2 — May 4 - May 18 — ☑ 1/1
- re-work charter 4
- fix Project charter document - 4.5 — May 4 - May 15

+ Add a card



To Do

- Add screenshots of trello
- check with Nir about storage resources
- finish Project charter document
- Submit draft of Charter document to Revital — Jun 30
- create a model — 1 — May 9 - Aug 31 — ☑ 0/1
- train the model — 1 — May 12 - Oct 23 — ☑ 0/1
- choose model architecture — Started: May 4 — ☑ 0/1

+ Add a card

In Progress

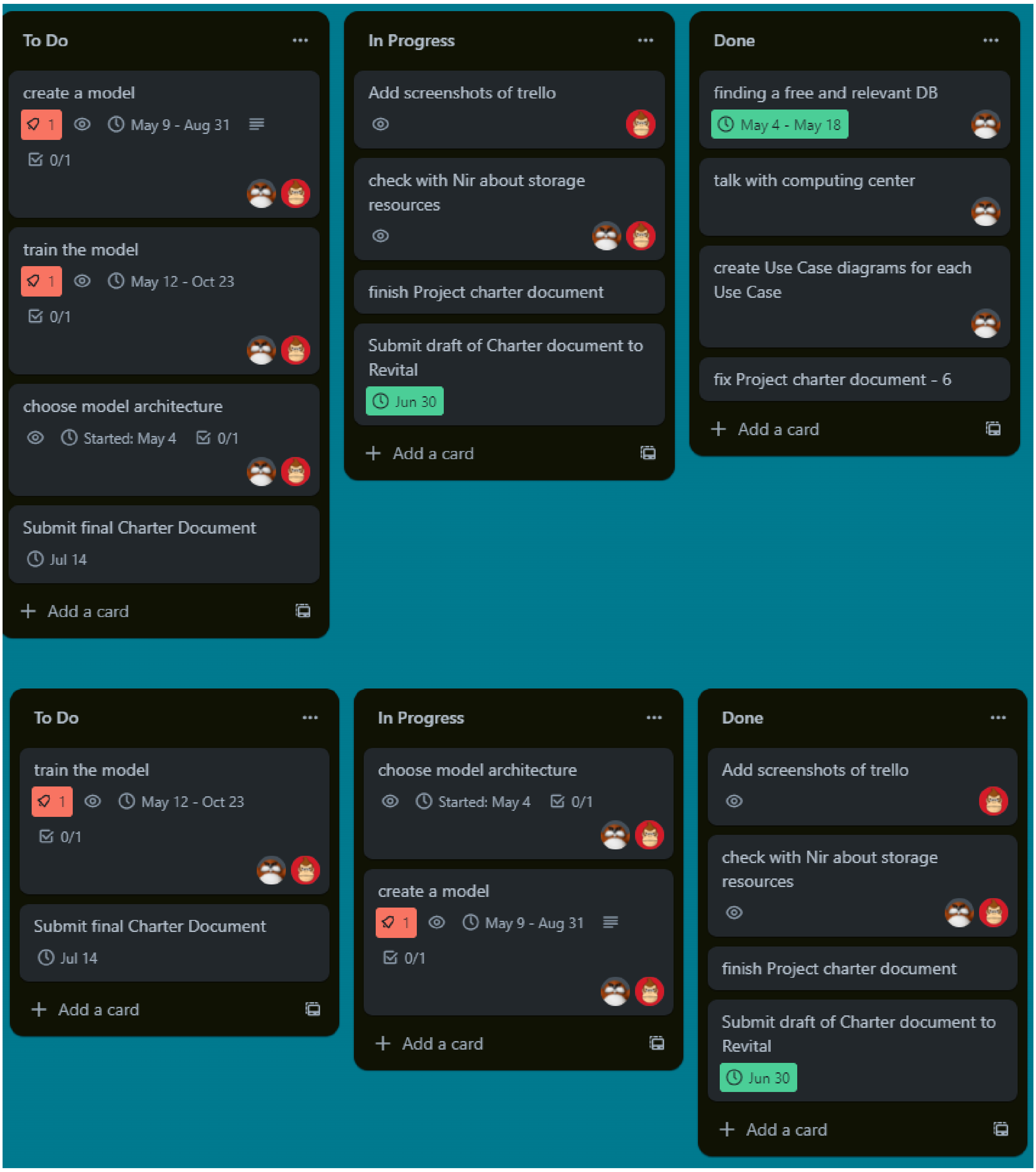
- finding a free and relevant DB — May 4 - May 18
- talk with computing center
- fix Project charter document - 6
- create Use Case diagrams for each Use Case

+ Add a card

Done

- Project charter document - 5 — May 4 - Jun 2 — ☑ 0/1
- ERD need to be expanded to more tables for holding the contents
- add measures to the goals — May 4 - May 18 — ☑ 0/1
- add charter 3 to the project document
- change everything to English
- Literature review — Apr 25 - May 8 — ☑ 1/1

+ Add a card



To Do

- create a model
1 · May 9 - Aug 31 · 0/1
- train the model
1 · May 12 - Oct 23 · 0/1
- choose model architecture
Started: May 4 · 0/1
- Submit final Charter Document
Jul 14
- + Add a card

In Progress

- Add screenshots of trello
- check with Nir about storage resources
- finish Project charter document
- Submit draft of Charter document to Revital
Jun 30
- + Add a card

Done

- finding a free and relevant DB
May 4 - May 18
- talk with computing center
- create Use Case diagrams for each Use Case
- fix Project charter document - 6
- + Add a card

To Do

- train the model
1 · May 12 - Oct 23 · 0/1
- Submit final Charter Document
Jul 14
- + Add a card

In Progress

- choose model architecture
Started: May 4 · 0/1
- create a model
1 · May 9 - Aug 31 · 0/1
- + Add a card

Done

- Add screenshots of trello
- check with Nir about storage resources
- finish Project charter document
- Submit draft of Charter document to Revital
Jun 30
- + Add a card

4.5. Severity scale 1-low 5-high, Probability scale 1-low 5-high

Index	Major risks	Probability scale	Severity scale	Ways of mitigation	Damage intensity
1	Use of an unreliable dataset or a dataset of a smaller then needed size	2	5	I. Finding a large and reliable dataset II. Finding datasets from multiple origins III. Altering and expanding the dataset IV. Combining datasets from multiple origins to create a diverse and wide singular dataset	10
2	Training errors like Overfitting & Underfitting	3	2	I. Altering the model by adding 'noise' II. Changing the parameters used III. Alter the chosen features in the model IV. Choosing a different model architecture V. Use of feature-select algorithms	6
3	This field is cutting edge and new and as such there will be a lack of papers & information sources about it	2	3	I. Conduction of a thorough literature review II. Researching faking tools and investigating them III. Having 'a finger on the pulse' as the development process continues in order to be up to date with the tech throughout the project	6
4	A lack of professional and industry standard in writing code of this field	3	2	I. Adherence to regular machine learning writing conventions with respect to needed changes II. Analysis of similar algorithms and extraction of pattern III. Setting a personal standard and keeping to it throughout development	6
5	Difficulties in validating 'live' information from the web and updating new findings in real time	4	3	I. Usage of proper datasets. II. Creating an independent database III. Creating a model that doesn't rely on catching specific faking	12

				tools but rather can work for any	
6	Permission to use recorded voice	1	5	I. Requesting appropriate permissions II. Finding other and accessible sources III. Use of existing datasets which have existing permissions	5
7	High prediction errors like FP or low accuracy scores like F1	4	4	I. Improving the model and polishing it over an extended period of time II. Use a wide range of faked voices as well as real voice samples III. Alter the parameters, features, architecture and ect. In order to improve the needed scores	16

4.6. Defining possible limitations

4.6.1 As voice spoofing technology advances, our detection methods become obsolete.

4.6.2 The voice file format (.mp4, .mp3 ect) convention changing in the future to one our system cannot extract details from

4.6.3 The system cannot detect the origin of the spoof, only find if it is spoofed or not.

5. Results & High Level Design

5.1.1 System contents: block chart that will be a derivative of the requirements chapter, gap filling, fulfilling requirements of system implementation

Gap filling- To fill the gaps that exist on the academic level and as the field is cutting edge, we performed a deep literature review and have read many papers on the field. Additionally we examined multiple sites of datasets to accumulate a large and reliable dataset which we could use during the model creation and training stages.

Fulfilling requirements of system implementation- to implement the system we require a deep understanding of machine learning, a field we have been educated in as part of our college degree. In addition we’re required to present a level of organized working which we fulfill, and access to datasets which we have received.

5.1.1 The main product of the system is an AI that can recognize if a voice sample is a real human voice or a machine generated one. It can be used as an additional security measure in banks or phones or other systems, and prevent future attacks fueled by faked voice.

5.1.2. The system will give security service to institutes or systems that recognize customers or secure themselves through voice recognition. As wella s provide a defense against phone call scams.

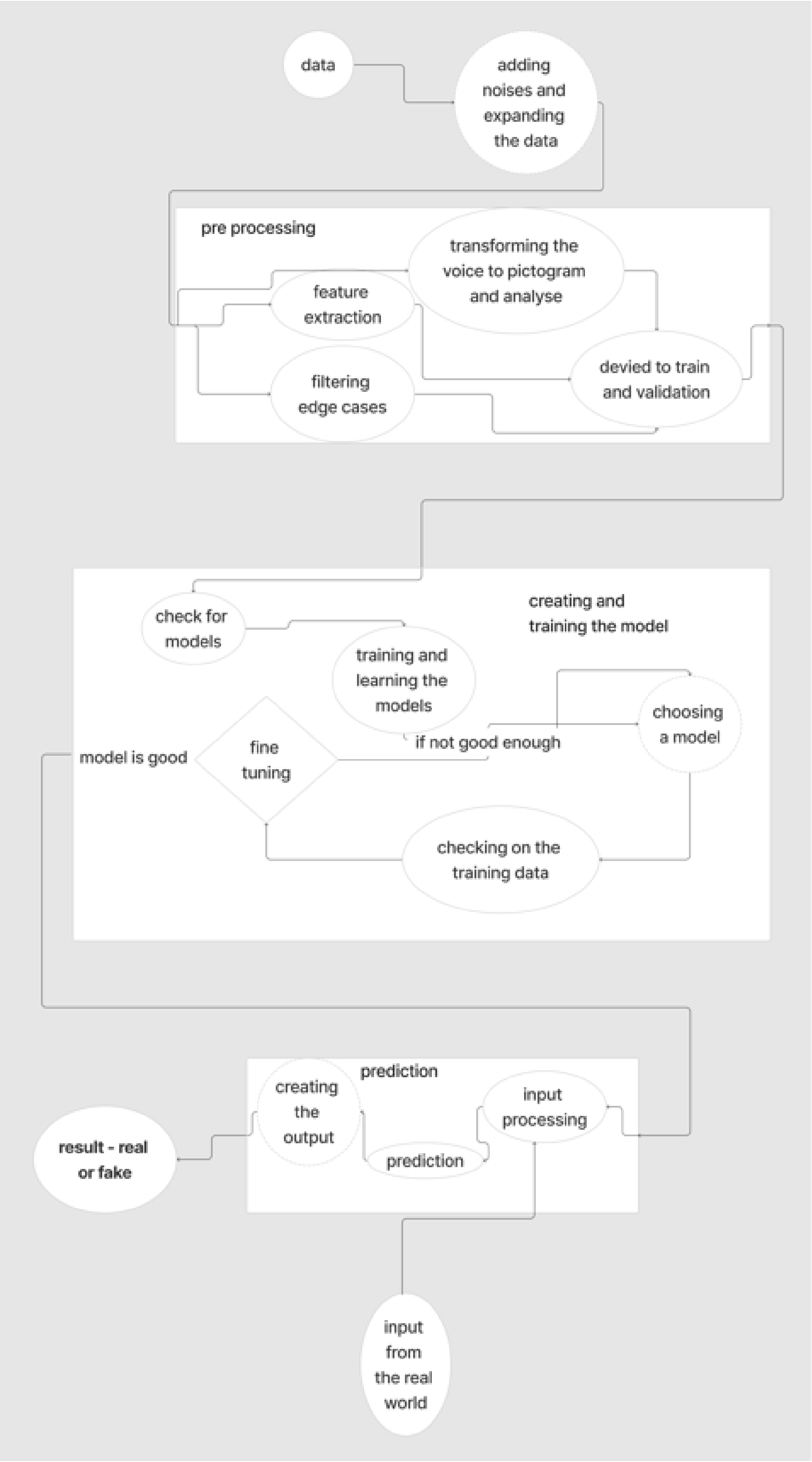
5.2 Product and Module Specifications:

5.2.1 General description of the expected products in the project, referring to the final system, its content, components, and operation. The description will be presented and detailed in a table, according to the required topics.

General description	Contents	Components	Ways it’s used
data	A database and augmentations that will provide a wide base to train the model	A number of databases taken from other academic resources in the field, as well as augmentations written by us	The information is duplicated and modified by the augmentations in order to create "noise" and improve learning, and was then transferred to the pre-processing stage
Pre processing	Fitting feature extraction, filtering edge cases, turning voice to image, image analysis, turning image to voice, dividing data to training and validation	Filtering features by their influence, checking the voice samples with edge case values and removing them, converting voice to a pictogram to get more features, dividing the data at a 80/20 ratio to train and validation	The data gets initially filtered, and from each sample we extract the features we’ll use during the training phase, the data we keep and divide in a way that’ll allow use to train with lower risk to go into ‘over-fitting’
Training	Checking models for use, choosing ideal model, validating model, improvement of the model if benchmark isn’t met	Performing a wide array of checks on several models and equating them to one another, training the models and comparing their validation scores, choosing the best model, checking the chosen model on the validation data, improving the model if needed. Choosing the default if we must.	After receiving the prepped data from the preprocessing stage, the training module uses the data to train and validate a number of different ML models, and chooses the best one. The best one is determined by parameters such as accuracy, FP type errors, F1 score. All these scores are calculated together and the best model is chosen and passed to the prediction module.

Prediction	The prediction module contains the input processing and the prediction model, the active prediction stage, and the output preparation stage.	The input processing component and the prediction model receive the prediction model from the training phase. Additionally, a voice sample provided to the system as input is processed based on the features utilized by the system. During the prediction phase, the system runs the model on the input and checks whether it is genuine or forged. In the output preparation phase, the system generates the prediction result as required.	Once the model is trained once, the prediction model is then used multiple times, and the main usage of the system is when a user inputs a voice sample in the appropriate format. The system, along with the prediction model, will return an output indicating whether the voice sample is forged or genuine.
Result	Output displays if input file is faked or not	A trained model that can provide a result	The model that was trained in the past examines the features of the file and the security level of the model to determine whether the file is genuine or not. Afterward, it updates the user with the result and the confidence level.

5.2.2 Block diagram


data
adding noises and expanding the data
pre processing
transforming the voice to pictogram and analyse
feature extraction
filtering edge cases
devied to train and validation
check for models
creating and training the model
training and learning the models
choosing a model
if not good enough
model is good
fine tuning
checking on the training data
prediction
creating the output
input processing
prediction
result - real or fake
input from the real world

5.2.3.1

Component	**Responsibility**	**Location**
Data	A dataset we can train the model on	Initial stage- preparation
Adding 'noise' and expanding the data	Expanding the data to add diversity	Initial stage- preparation
Feature extraction	The selection of features on which the model relies later on, to decide whether the file is genuine or not, is made at a later stage.	Second stage - pre processing
Filtering edge cases	The extraction of poor-quality audio files, such as unclear or heavily noisy background audio, may pose challenges in the prediction process.	Second stage - pre processing
Conversion to a pictogram and back	Expanding the feature selection and including additional tests using image-related features, rather than relying solely on audio, can enhance the system's capabilities. By incorporating image-related features, the system can perform additional checks and analyses that are not solely reliant on audio.	Second stage - pre processing
Division to train and valid	The remaining data is typically divided into a training set, which is used to train and learn from the model, and a validation set, on which the model provides expected predictions and compares them to the actual values to assess its	Second stage - pre processing

	accuracy.	
Model checking	Checking which model provides the best result	Third stage- training
training	Training the model	Third stage- training
Choose ideal model	Choosing the most accurate model	Third stage- training
Checking on validation data	Check the model on new data and check for overfitting	Third stage- training
Improvement of the model	Check model scores vs. benchmark scores and decide if it needs to be improved.	Third stage- training
Input	Receiving input from user	Fourth state- input
Input processing	Process the input to extract features	Fourth state- input
Predict	Use the prediction model to predict on the input and decide if it’s a faked voice or not	Fourth state- input
Output prep	Prepare the output of prediction	Fourth state- input
result	Return or print result	Fifth stage- output

5.2.3.2 The existing system does not have external integration but rather the embedding of the system within other systems that will provide it with input only for it to serve as a defense mechanism.

5.2.4 The engineering challenges include understanding large language model (LLM) models, selecting an appropriate architecture for the language model, creating a comprehensive yet not overly complex model that can be run on a regular computer, dealing with invalid voice files and correcting or filtering them.

5.2.5 Following the mapping and detailing of the content and block modules, initial functional analysis will be performed, and use case scenarios and flowcharts of the planned system will be presented.

5.2.5.1 use cases, Sequence and flow diagrams

Usage scenarios were used to illustrate and display the use cases

Use Case 1: Bank User Verification by Voice Recognition

Primary Flow:

1. The user initiates the bank account login process and chooses the voice recognition option.

2. The system prompts the user to say a specific passphrase or series of phrases.

3. The system records the user's voice and uses the voice recognition AI.

4. The voice recognition AI returns to the system if the voice is real or fake.

5. If the system confirms a match, the user is granted access to their account.

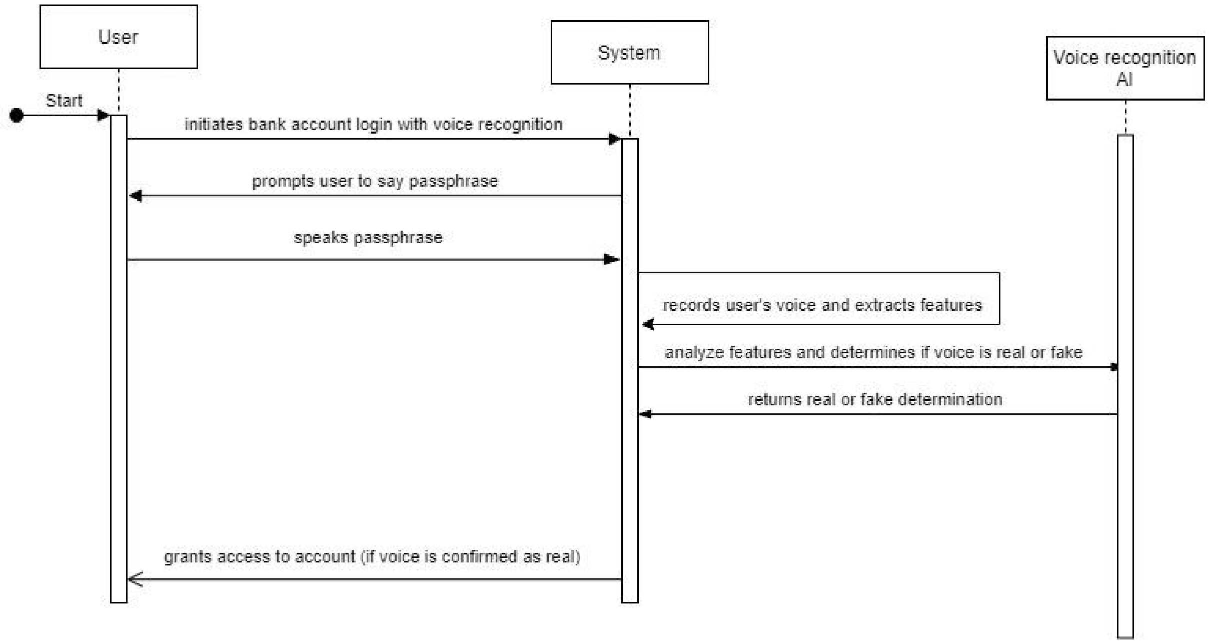
Alternative Flow:

1. If the system cannot recognize the voice due to background noise or other factors, the system prompts the user to repeat the passphrase or phrases.

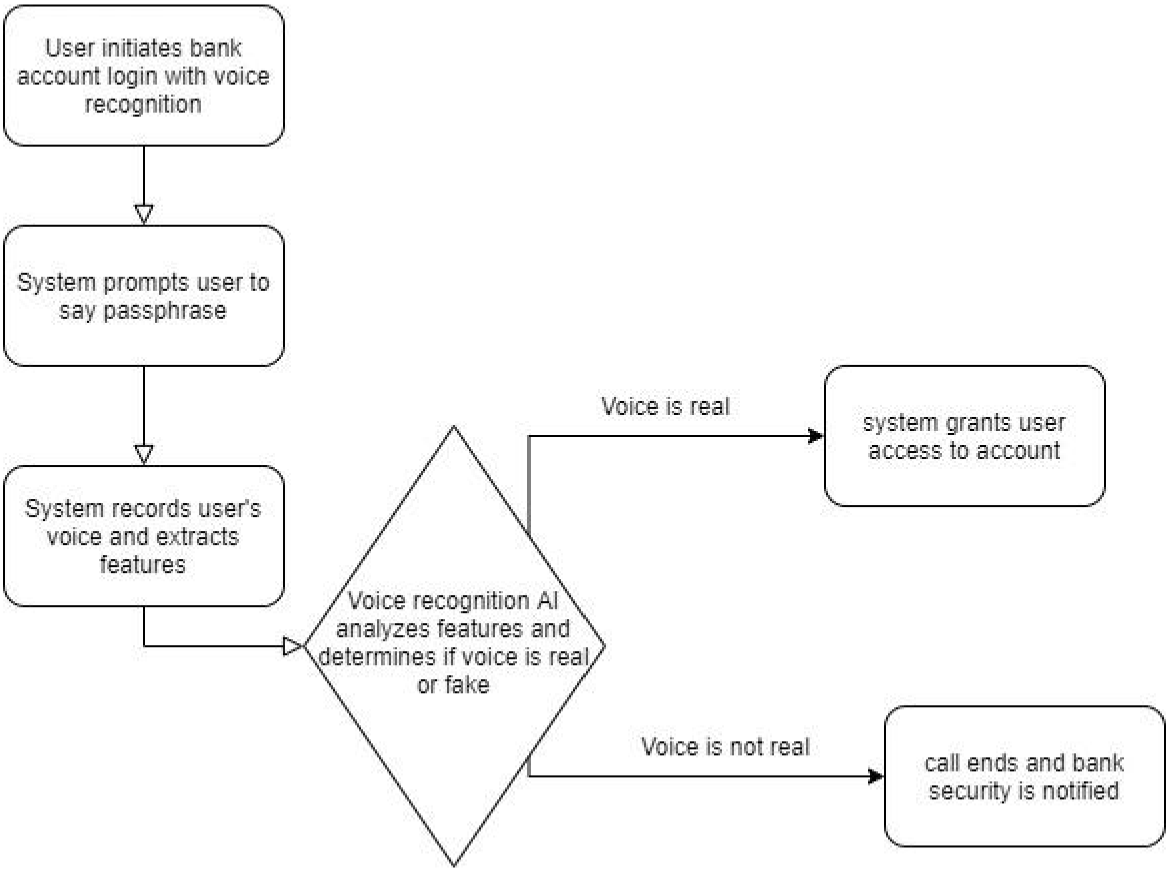
2. If the system cannot confirm a match after multiple attempts, the user is prompted to try again or use another form of authentication.

3. If the voice recognition AI returns to the system that the voice is fake, the call will end and the bank security will call to the phone number of the bank user.

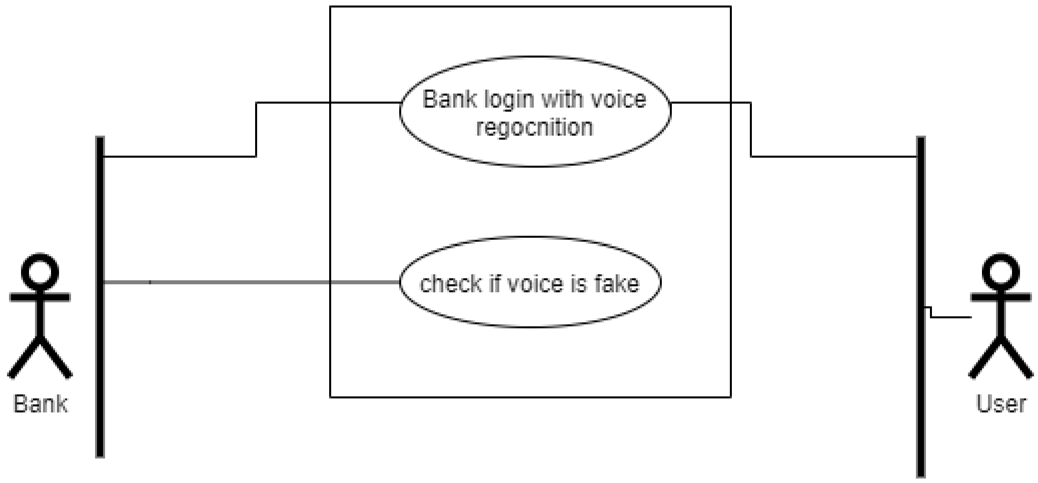
Sequence diagram


User
System
Voice recognition AI
Start
initiates bank account login with voice recognition
prompts user to say passphrase
speaks passphrase
records user's voice and extracts features
analyze features and determines if voice is real or fake
returns real or fake determination
grants access to account (if voice is confirmed as real)

Flow diagram


User initiates bank account login with voice recognition
System prompts user to say passphrase
System records user's voice and extracts features
Voice recognition AI analyzes features and determines if voice is real or fake
Voice is real
system grants user access to account
Voice is not real
call ends and bank security is notified

Use case diagram



Use Case 2: Elderly Spoofing Attack Prevention

Primary Flow:

1. An elderly person receives a phone call from someone claiming to be a family member or friend.

2. The person on the phone requests money or sensitive information, such as bank account details.

3. The elderly person becomes suspicious and activates the voice recognition AI system on their phone.

4. The system records the caller's voice and returns if the voice is real or fake.

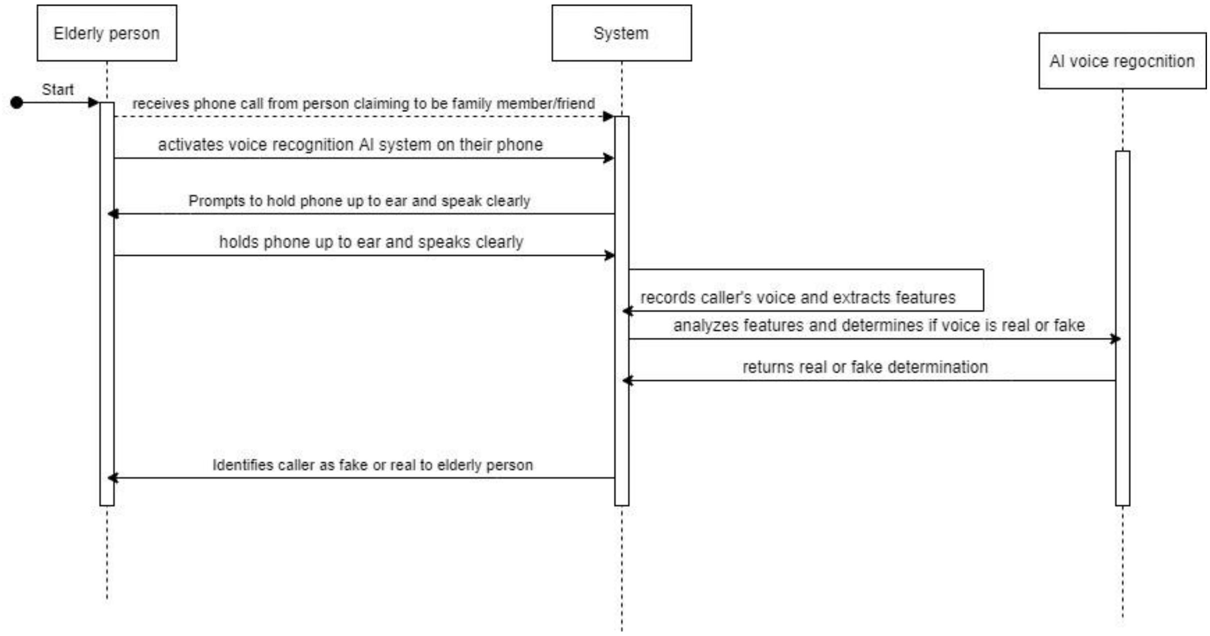
5. If the system confirms a match, the caller is identified as a genuine family member or friend and the elderly person can proceed with the conversation or end the call.

Alternative Flow:

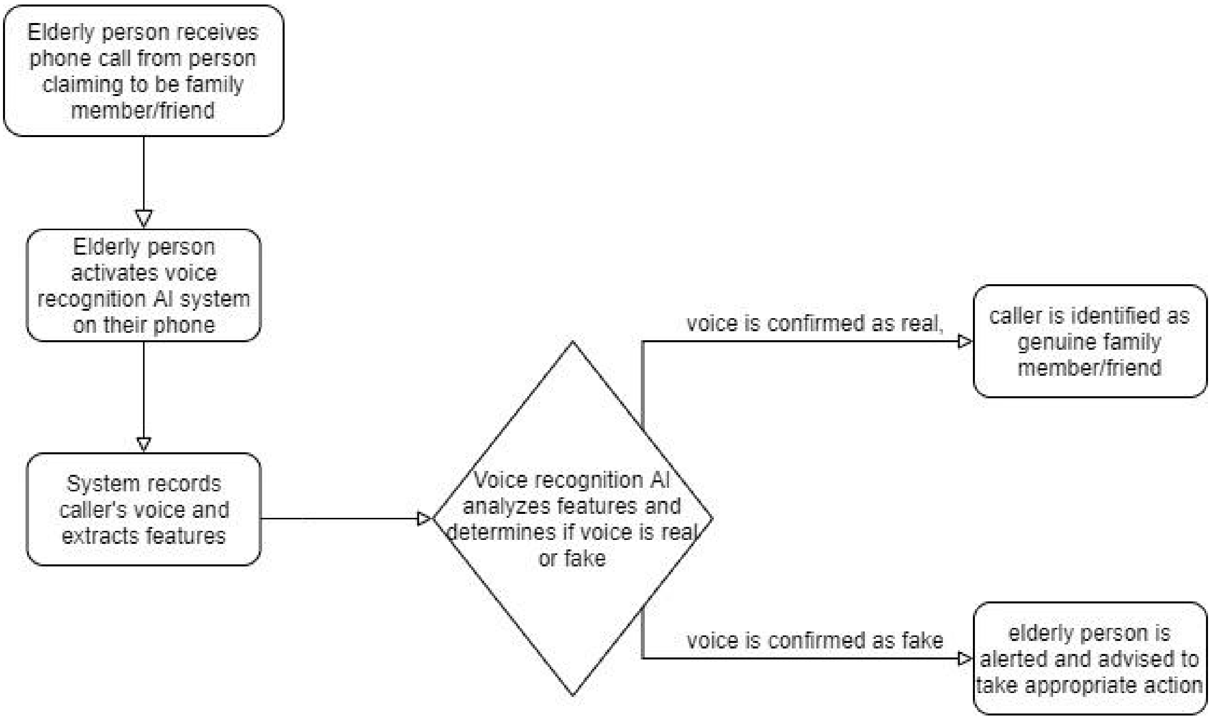
1. If the system cannot recognize the voice due to poor call quality or other factors, the system prompts the elderly person to ask the caller to repeat themselves or call back later.

2. If the system cannot confirm a match after multiple attempts, the elderly person is alerted that the caller may be a fraud and advised to take appropriate action, such as contacting the authorities.

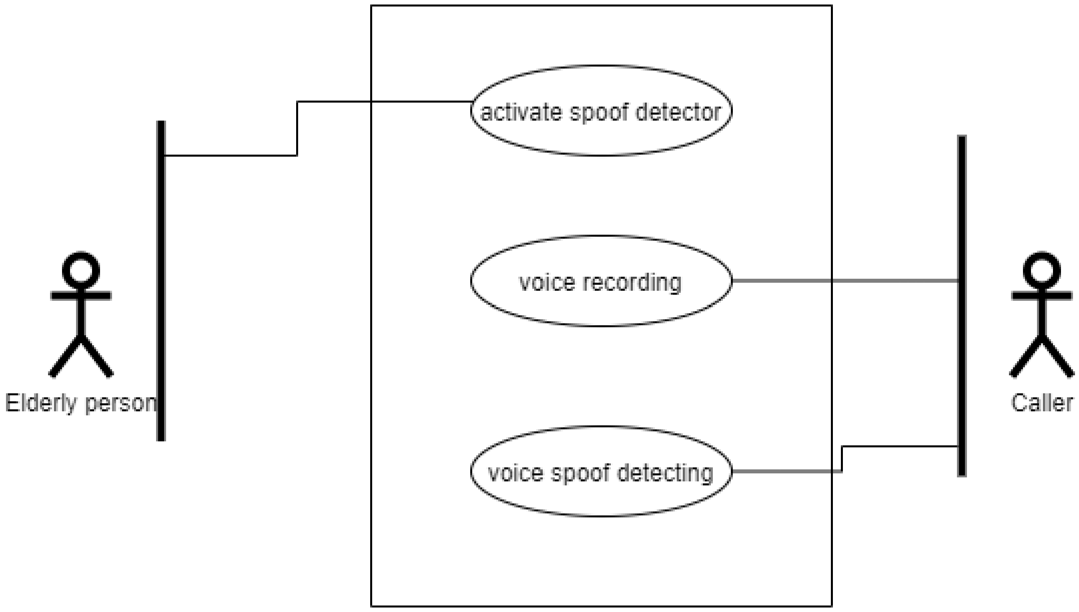
Sequence diagram


Elderly person
System
AI voice regocnition
Start
receives phone call from person claiming to be family member/friend
activates voice recognition AI system on their phone
Prompts to hold phone up to ear and speak clearly
holds phone up to ear and speaks clearly
records caller's voice and extracts features
analyzes features and determines if voice is real or fake
returns real or fake determination
Identifies caller as fake or real to elderly person

Flow diagram


Elderly person receives phone call from person claiming to be family member/friend
Elderly person activates voice recognition AI system on their phone
System records caller's voice and extracts features
Voice recognition AI analyzes features and determines if voice is real or fake
voice is confirmed as real,
caller is identified as genuine family member/friend
voice is confirmed as fake
elderly person is alerted and advised to take appropriate action

Use case diagram



Use Case 3: Prevention of False Accusations

Primary Flow:

1. An individual accuses someone of a crime based on their voice recording.

2. The accused individual's defense team requests access to the voice recording and submits it to the voice recognition AI system.

3. The system checks the voice sample and returns if the voice is real or fake.

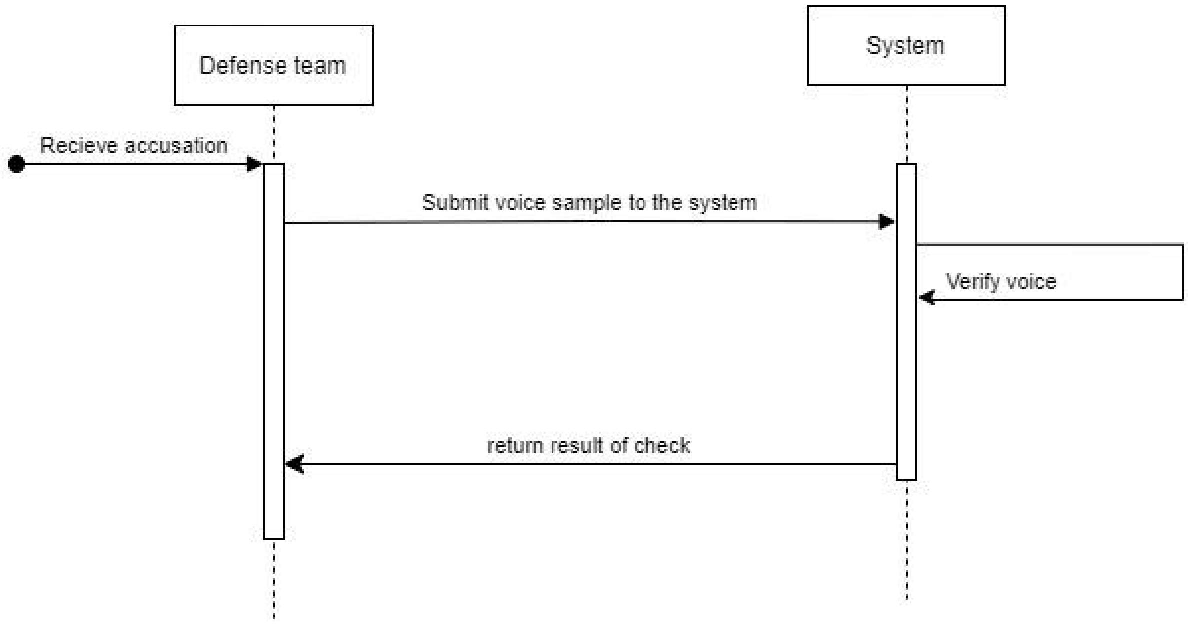
4. If the system identifies that the voice is fake, the accused individual may be exonerated and the actual suspect may be identified.

Alternative Flow:

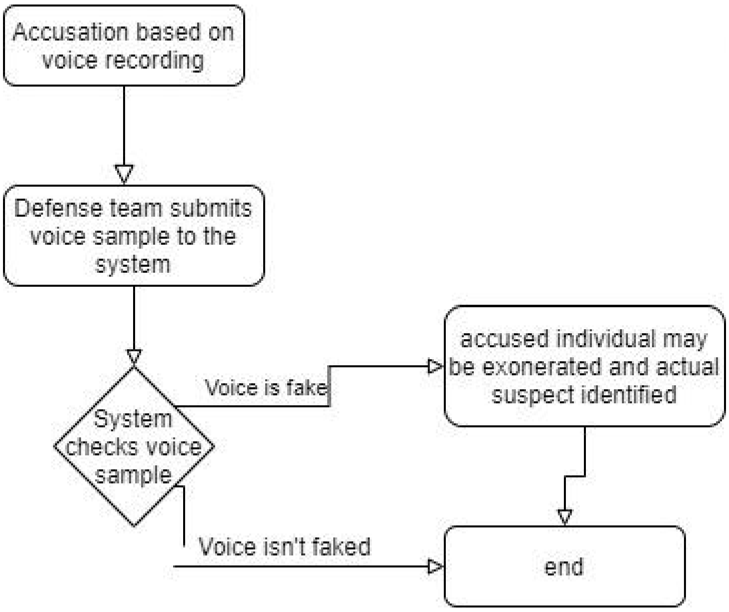
1. If the system cannot recognize the necessary voice due to poor recording quality or other factors, the defense team may need to submit additional voice samples or other evidence.

2. If the system does not identify a match with any known suspects or witnesses, further investigation may be required.

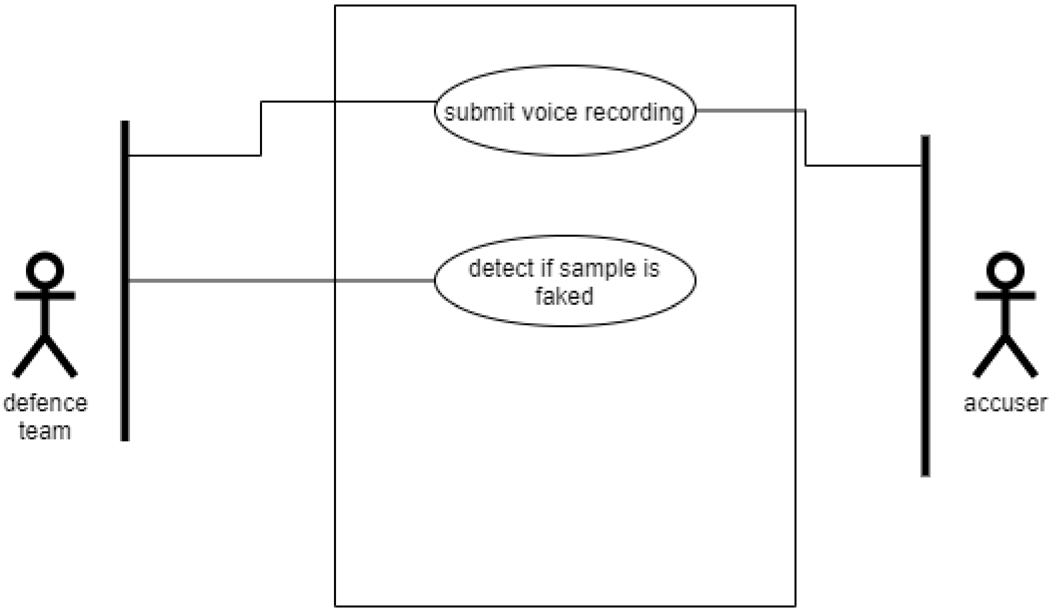
Sequence diagram



Flow diagram


Accusation based on voice recording
Defense team submits voice sample to the system
System checks voice sample
Voice is fake
accused individual may be exonerated and actual suspect identified
Voice isn't faked
end

Use case diagram



Use Case 4: Verification of Public Figures' Statements

Primary Flow:

1. A public figure makes a statement in a speech or interview.

2. The system checks if the voice recording is real or fake.

3. If the system confirms a match, the statement can be verified as coming from the public figure.

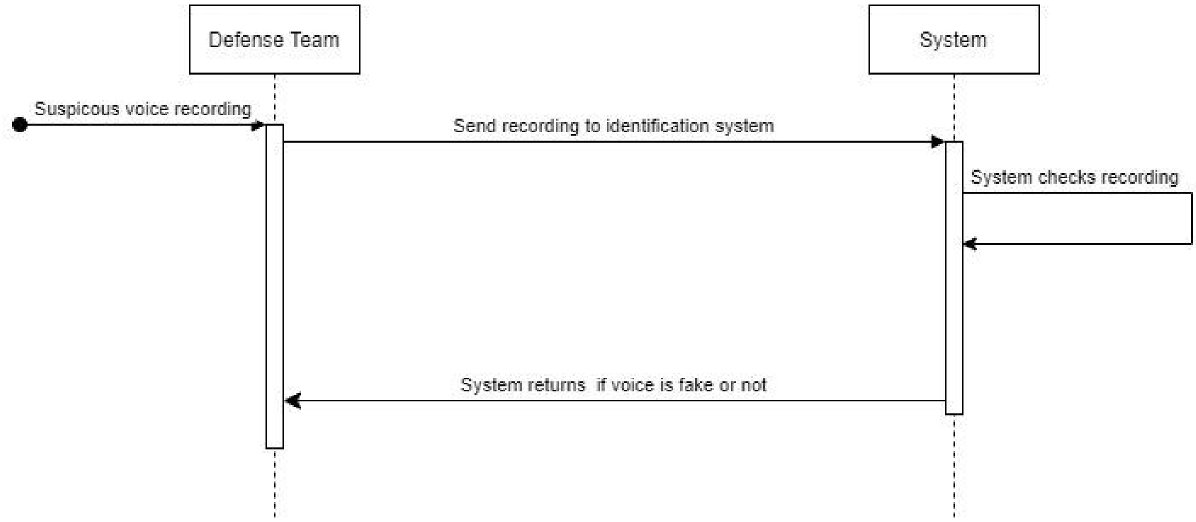
4. If the system does not confirm a match, the statement may be investigated for authenticity or accuracy.

Alternative Flow:

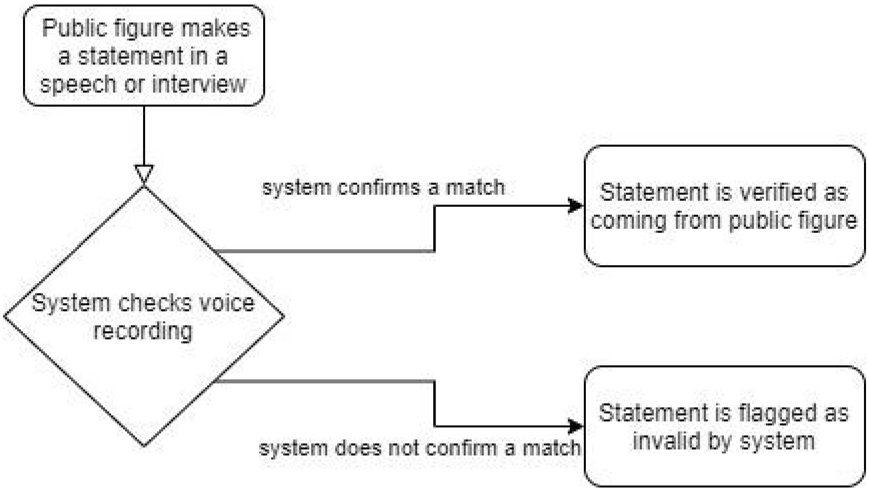
1. If the system cannot recognize the necessary voice due to poor recording quality or other factors, the statement may need to be verified through other means, such as through the public figure's representatives or through additional evidence.

2. If the system cannot confirm a match, the statement may need to be investigated further to determine its source and accuracy.

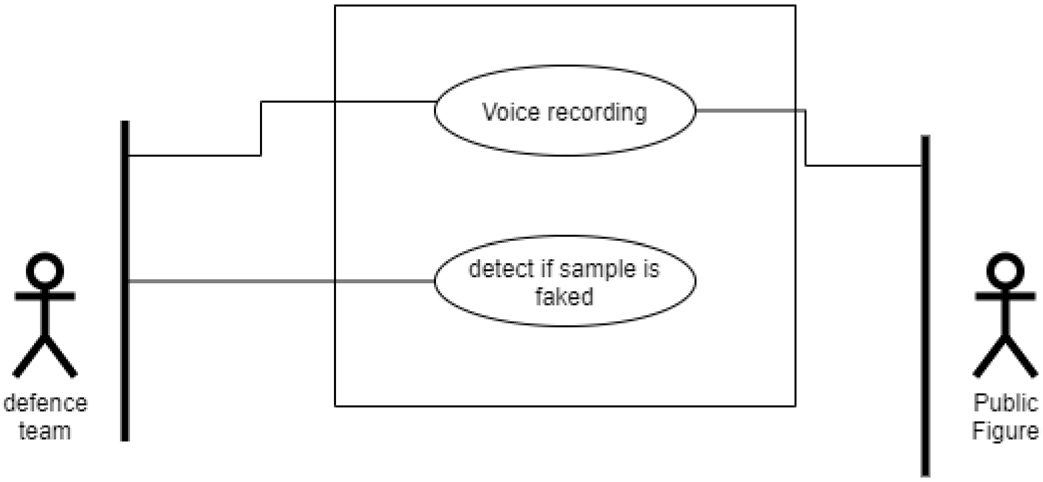
Sequence diagram


Defense Team
System
Suspicous voice recording
Send recording to identification system
System checks recording
System returns if voice is fake or not

Flow diagram


Public figure makes a statement in a speech or interview
system confirms a match
Statement is verified as coming from public figure
System checks voice recording
Statement is flagged as invalid by system
system does not confirm a match

Use case diagram



5.3 Presentation and Initial Analysis of Possible Alternatives for System Implementation

5.3.1 Mapping and analysis of alternatives for system implementation will refer to and be derived from the high-level content presented and detailed in the block diagrams and process diagrams.

5.3.2 Mapping and analysis of relevant alternative components for the system's implementation configuration will be presented in a table, describing the components of each alternative, their advantages and disadvantages, and defining metrics for comparison between alternatives.

5.3.3 The weight of each examined metric will be indicated, and each alternative will be given a final score according to its contribution to the system or project. The selected components will be those that received the highest scores.

5.3.4 The alternatives will be presented and analyzed based on the scope, configuration, and block diagrams of the planned system.

5.3.5 The mapping and analysis of alternatives and their components within the scope of the feasibility document will be at a higher level, not at a detailed engineering technological level, which is addressed in the engineering content document (the subsequent deliverable)

Component	Alternative	advantage	disadvantage	Comparison scale [between alternatives]	Weight [1-low , 10-high]	Score[up to weight]
Data - DB	A) Create a DB by ourselves B) Using a model of reinforcement learning	A.1) Database is tailor made to our problem and algorithm A.2) No need to worry about losing access at a later point A.3) Certainty in lack of issues B.1) No need to acquire and verify databases B.2) A highly flexible model type	A.1) Creating a database takes a lot of time and effort A.2) We risk having a low sample size, or low variance as our sources are limited A.3) We will need to acquire high quality recording equipment as well as storage space B.1) Model train time is significantly longer B.2) Model error rate is significantly higher B.3) Because of the nature of our project we need to guarantee which samples are real and which are fake, something Reinforcement cannot provide	I. Time- how much time does using this alternative cost. II. Reliability- How much can we trust the alternative won't pollute or harm the system.	I. 3 II. 7	A.I) 1 A.II) 7 B.I) 2 B.II) 3

Pre processing - feature preparation and selection	A) Usage of deep-learning algorithms that don’t require feature selection B) Not performing feature selection	A.1) Saves time, no need to analyze features and choose which ones to keep. Or in some cases even prepare them A.2) These types of algorithms can usually detect patterns other models and algorithms cannot B.1) Saves time, by passing the model all the features we do not spent time on writing feature selection algorithms or choosing them manually B.2) With a wider berth of data, human error in selecting wrong features is cut out of the equation and the model has access to all the data in the file.	A.1) Deep learning isn’t an explainable AI, making mistakes not be able to be traced to their source. A.2.) Requires significantly more computing resources, making training or even just running the model heavy and expensive B.1) The model might detect patterns which will be harmful to its prediction accuracy. A wolfs on snow case(a case where a image processing model design to detect wolves couldn’t find them on any other terrain other than snow, as the training photos were all snowy) B.2) Processing time, while initializing and getting the model to run will be faster, having to iterate over more data during training will increase the training time.	I. Model accuracy- How accurate the model is with predictions on voice samples II. Time and Resources- The amount of time training and running the model takes, as well as how expensive on computing resources it is.	I. 8 II. 2	A.I) 6 A.II) 1 B.I) 5 B.II) 2

Training - AI algorithm	A) Choosing a architecture that doesn’t need training, a “learn as you run” one B) Making predictions based on the features without training, hardcoded guidelines	A.1) Saves time as no training is needed A.2) no need to secure a database B.1) high manual adjustability of decision parameters B.2) reduce the resources and time needed to train the model	A.1) Bad initial results for a while until model picks up speed A.2) no way to verify correctness of predictions reliably, which could lead to a spiral in the wrong direction B.1) Unreliable results, our preset parameters can’t cover everything B.2) Spoof makers might change their methods to go around our hardcoded rules.	I. Model accuracy- How accurate the model is with predictions on voice samples II. Time and Resources- The amount of time training and running the model takes, as well as how expensive on computing resources it is.	I. 8 II. 2	A.I) 4 A.II) 2 B.I) 3 B.II) 2

5.3.6 Sample Contents of Alternatives (to be presented for each module, hereafter referred to as a block) planned for system implementation

5.3.6.1 Alternative of Preliminary Analysis using Different Algorithms - The use of algorithms that bypass the feature extraction and filtering process by performing tests on the entire input during system execution. This allows the system to predict what may be a genuine audio file in the future and what may not be, without knowing which feature carries more weight compared to others.

Alternative of Real-time Learning Architecture - Real-time learning architectures cannot reliably determine whether their predictions were accurate or not. Therefore, the learning process is not always reliable over time and requires broader supervision. This may even result in higher resource costs, especially in the security field we are dealing with, compared to early architecture selection and training.

Alternative of Guessing Prediction - We do not aim to reach this stage at all, as it is essentially the essence of the system. If the system cannot predict the results with reasonable probabilities, it will not be worthwhile. A "yes" or "no" guess is not 50%, and a reliable system requires at least 80% accuracy.

5.3.6.2 Use of Public Python Libraries for Machine Learning, Computer Vision, and Statistics such as OpenCV, plotly, numpy, torchvision, scipy.

5.3.6.3 We do not have integration with different ready-made commercial factors and packages.

5.3.7 For each alternative, the relevant components and their contribution weight to the system's goals, the level of difficulty in implementing the alternative, its advantages, and disadvantages will be specified.

6. Literature Review

This chapter consists of two parts: literature review and market analysis.

6.1 Literature Review and Articles

In this section, we examine relevant literary sources, research papers, and publications related to the scientific and engineering knowledge in the field of the project and the planned system. The project team will compare and present a comprehensive review of similar works conducted worldwide, including relevant citations and quotes. Additionally:

6.1.1 As part of the literature review, students are expected to explore and survey scientific sources relevant to the project and the planned system, including scientific articles, books, and official or professional websites.

6.1.2 The literature review should provide detailed information in the project proposal document, clearly indicating the source of information, its relevance to the surveyed or researched topic, and the relevant findings for this project.

6.1.3 The literature review will include sources from both domestic and international origins. Each citation in the text should follow the writing and presentation guidelines specified in the project procedures.

"Deepfake Audio Detection by Speaker Verification"

Authors: Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, Luisa Verdoliva

Summary:

This article discusses the challenges faced by voice recognition software in detecting deep fake audio in real-world scenarios. The authors highlight the presence of background noise and other factors that can disrupt the algorithm's performance. They propose a method that combines voice vectors with feature vectors to determine whether the voice belongs to a human using supervised learning. The authors conduct experiments using the CB (Centroid-Based) and MS (Maximum Similarity) methods on different datasets. The results show that the POI-F method achieves the most effective performance, although variations are observed across different datasets.

"Fake Audio Detection Using Hierarchical Representations Learning and Spectrogram Features"

Authors: Mohsan Ali, Alina Sabir, Mehdi Hassan

Summary:

This article presents a method for detecting fake audio by utilizing hierarchical representations, learning and spectrogram features. The authors convert audio files into images using the ASVSpoof2019 dataset, achieving an accuracy rate of 98%. They employ MFCC (Mel-Frequency Cepstral Coefficients) and CNN (Convolutional Neural Network) to decompose the audio files and extract features. The resulting images are then used to visualize the outcomes. The authors employ a neural network architecture that combines the extracted features from both the voice file and the image. The output is passed through a trained model to determine the authenticity of the audio file. The experiments show high success probabilities without evidence of overfitting or underfitting. The dataset used comprises three different attack possibilities, primarily focusing on the Replay attack. The authors utilize the LSMN (Long Short-Term Memory with Short-Term Fourier Transform) architecture for training and testing.

"Analysis of Amplitude and Frequency Perturbation in the Voice for Fake Audio Detection"

Authors: Kai Li∗, Yao Wang∗, Minh Le Nguyen, Masato Akagi, Masashi Unoki†

Summary:

In this article, the authors explore various methods for extracting features from voice, including both dynamic and static approaches, in addition to spectrogram analysis. They also investigate the effects of "Shimmer" and "Glitter." The error is measured using Binary Cross Entropy (BCE). The researchers use the ADD2022 dataset, which they claim contains more cases similar to real-world scenarios compared to the ASV dataset. They preprocess the audio files to create a diverse dataset. The authors evaluate the Equal Error Rate (EER) using CS3 with various features.

Detecting Deep Fake Voice Using Explainable Deep Learning Techniques

Authors: Suk-young Lim, Dong-Kyu Chae, Sang-Chul Lee

Summary:
The paper discusses the implementation of explainable artificial intelligence (XAI) methods for deep fake voice detection and the interpretation of the models at a human perception level. The study highlights the limitations of current deepfake detection methods, such as the less-than-perfect accuracy of detection and the lack of interpretability. The authors propose the use of XAI methods to provide interpretable explanations for deep fake voice detection. They focus on simplicity and interpretability, using simple model structures and spectrogram-based feature extraction. The experiments are conducted using A spoof 2019 Logical Access and LJSpeech datasets, and XAI methods such as Deep Taylor, integrated gradients, and layer-wise relevance propagation are used for interpretation. The study aims to provide multi-modal interpretability for deep face detection, considering both visual and audio interpretation.

The paper also discusses related works on explainable deep fake image detection and explainable speech recognition. It mentions the limited application of XAI methods to speech recognition and the need for qualitative evaluation of interpretation. The authors introduce the concept of interpreting deepfake detection models at a human cognitive level and address the limitations of existing XAI methods. They propose using the attribution scores obtained from XAI methods for qualitative interpretation and compare them with traditional audio analysis methods. The paper provides an overview of deep fakes, speech analysis techniques, and the datasets used for the experiments. It concludes by summarizing the major contributions of the study and outlining the organization of the remaining sections, including related works, preliminaries and datasets, methods, and the interpretation based on the attribution scores.

On the Generalizability of Two-dimensional Convolutional Neural Networks for Fake Speech Detection

Authors: Christoforos Papastergiopoulos, Anastasios Vafeiadis, Ioannis Papadimitriou, Konstantinos Votis, Dimitrios Tzovaras

Summary:
The article discusses the importance of feature extraction methods in the pre-processing phase of deep neural network (DNN) models for audio data. The authors compare two approaches: using raw audio data in the time domain and computing time-frequency representations such as Short-time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (MFCCs), and Mel spectrograms. They evaluate the performance of these features using the VGG16 architecture for a fake speech detection task. The results show that Mel spectrograms and Mel magnitude spectrograms perform well in distinguishing

between fake and real audio, while STFT spectrograms struggle to differentiate between the two. The authors also augment the TIMIT audio dataset with synthetic speech samples and evaluate the trained models on this augmented dataset, finding a significant drop in performance, indicating a difference in the distributions of the training and test data.

In summary, the study highlights the impact of feature extraction methodologies on the performance of DNN models for audio data. The use of Mel spectrograms and Mel magnitude spectrograms proves effective in identifying fake speech, while STFT spectrograms exhibit limitations in distinguishing between fake and real audio. The evaluation on the augmented dataset reveals a discrepancy between the training and test data distributions, indicating the need for more diverse and representative training data to enhance model generalization.

DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voice

Authors: Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, Yang Liu

Summary:

DeepSonar is a novel framework designed to distinguish real human speeches from AI-synthesized fake voices effectively. The key lies in monitoring layer-wise neuron behaviors in deep neural networks (DNNs), enabling the detection of subtle differences between inputs.

The process involves collecting a diverse set of real human speeches and AI-synthesized fake voices. A DNN-based speaker recognition system captures the layer-wise neuron behaviors, and two neuron coverage criteria, average count neuron (ACN) and top-k activated neuron (TKAN), are used to identify crucial neurons for discriminating between real and fake voices. These neuron behaviors are then fed into a binary classifier, which learns to predict whether a voice clip belongs to a human or is an AI-synthesized fake.

By utilizing the layer-wise neuron behaviors, DeepSonar achieves robustness against voice conversion and noise addition, making it a powerful tool for voice authentication and detecting synthesized voices with high accuracy.

Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction

Authors: Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O’Dell, Kevin Butler, and Patrick Traynor, University of Florida

Summary:

The hypothesis of the paper is that human-created speech is limited by the anatomical structures involved in its production, while synthetic audio generated by deepfake algorithms is not bound by these physical constraints. The paper proposes a methodology for detecting deep fake audio samples by modeling the acoustic behavior of a speaker's vocal tract.

The security model presented in the paper involves an adversary, a victim, and a defender. The adversary's goal is to create a deep fake audio sample of the victim uttering a specific phrase. The defender's task is to determine whether the provided sample is a deepfake or an organic audio sample. The defender does not have access to the victim's audio data or the attacker's audio generation algorithm. The paper assumes a scenario where the media acts as the defender, deciding whether to publish a potentially synthetic audio sample of a politician.

The methodology consists of two steps: the Vocal Tract Feature Estimator and the Deep Fake Audio Detector. The Vocal Tract Feature Estimator constructs a mathematical model of the speaker's vocal tract based on the frequencies present in their voice during a specific pair of adjacent phonemes. The Deep Fake Audio Detector uses the values obtained from the organic dataset to determine if an audio sample can be realistically produced by an organic speaker. The paper also discusses the limitations and assumptions of the model, such as ignoring energy losses and limitations to vowel phonemes.

Overall, the paper aims to provide a technique for detecting deep fake audio samples by analyzing the acoustic properties of a speaker's vocal tract. The methodology involves modeling and estimating vocal tract features, and it is applied in a scenario where the defender needs to assess the authenticity of audio samples provided by an adversary.

Unused articles:

A Machine Learning Model to Detect Fake Voice

Summary:

The article titled "A Machine Learning Model to Detect Fake Voice" explores the use of machine learning (ML) and logistic regression (LG) models for detecting fake voice. However, the article does not present any novel or groundbreaking information that could contribute to our project. It provides an overview of ML and LG models and describes the construction of a regular model without offering any new insights or relevant findings. As a result, this article does not provide additional information essential for our research efforts.

6.2 Market Analysis

This section analyzes the current market situation and its relevance to similar systems, methods, and existing products that address the problem the project aims to solve:

6.2.1 The analysis will focus on similar systems and potential competitors.

6.2.2 For each finding, system, or competitor documented in the market analysis, a brief explanation and a link to further information should be provided, highlighting their advantages and disadvantages compared to the proposed solution within the project framework.

1) Dessa- a company working on an AI voice detection product to combat the rise of deepfakes and synthetic media. The company aims to raise public awareness about the risks of AI-powered speech synthesis and has developed a proprietary speech synthesis model called RealTalk. They recreated the voice of popular podcaster Joe Rogan to showcase the technology's capabilities and highlight the issue of audio deepfakes. Now, Dessa is sharing the next step in their work—a detector system that can distinguish between real and fake audio examples.

   The detector system utilizes spectrograms, visual representations of sound, to discern between real and fake audio clips. By analyzing the differences in spectrograms, which are not easily detectable by the human ear, the detector model can accurately identify fake audio. The model is a deep neural network that uses Temporal convolution and is trained on Google's AVSSpoof dataset, which contains both real and fake audio clips. Dessa's baseline model achieved high accuracy rates of 99% on the train set, 95% on the validation set, and 85% on the test set. Overall, the detector model can currently predict over 90% of the fake audio clips it is shown.

   Our project’s advantage over Dessa is the fact we utilize both spectrograms and audio file analysis in order to extract our features and train our model, giving us a wider berth of features information to work with. Both of our projects will utilize the AVSSpoof database. An advantage Dessa has over us is their ability to create reliable fake audio files on their own, as exemplified with the Joe Rogan example.

2) Reality defender- Winners of 2023 SXSW Pitch Competition in the Artificial Intelligence, Voice, & Robotics Technologies category, Reality Defender are a company that deals with detecting deep fakes, spoofs, and other AI created media, they Produces alerts, report cards and other ways to visualize and take action against fake content. Reality Defender's advanced deep face detection technology helped NATO STRATCOM identify deepfakes and misinformation spread by pro-Kremlin actors on VKontakte, providing them with the information they needed to make informed decisions and take appropriate actions to stop the spread of disinformation and propaganda.

   Not much direct information on how Reality defender’s model works, and as such we don’t know exactly of any advantages we hold over them with our current goals and design, however their track record with impressive case files, and their wide array of products suggest a powerful model or architecture they have developed.

3) Aivoicedetector - aivoicedetector.com is a website that offers a convenient solution for verifying whether an audio file has been generated by artificial intelligence (AI) or a human voice. By uploading the audio file to the website and clicking on the "detect now" button, users can leverage the advanced algorithms of aivoicedetector.com to analyze the audio and receive a comprehensive report on its origin.

   The functioning of aivoicedetector.com involves the analysis of different characteristics within the audio file. Factors such as pitch, tone, inflection, and subtle cues are taken into account by sophisticated algorithms to determine whether the audio was generated by AI or a human voice. Upon completion of the analysis, users are presented with a detailed report containing the probability of AI generation and other relevant data points. This information empowers users to make informed decisions about the authenticity of the audio file in question.

   Our advantage over devicedetector is our use of pictorgarms as well as audio file analysis, this gives us more features to work with and increases our chances of identifying successfully AI-generated voices

6.2.3 A comparison table should be included, outlining the features, services, and costs of each competitor (the proposed system will be presented in the last column of the table).

ID	Features	Services	Costs	System
1	Deepfake audio detection using spectrograms Preprocessing of raw audio into mel-frequency spectrograms Temporal convolution neural network architecture Masked pooling to prevent overfitting Dense layer and sigmoid activation function for probability output	Open-source code and resources provided for building deep face detection models Access to pre-processed data, training code, and inference code Compatibility with the free Community Edition of Foundations Atlas, an ML development platform	The services mentioned, including access to code, data, and tutorials, are provided for free.	The deep fake detection system is based on a deep neural network model. Raw audio is preprocessed into mel-frequency spectrograms, which serve as input to the model. The model performs convolutions over the time dimension of the spectrogram, uses masked pooling, and includes a dense layer with a sigmoid activation function. The model predicts the probability of audio being real or fake, with a threshold of 0.5 for classification. The system achieved high accuracy on train, validation, and test sets, with the ability to predict over 90% of fake audio clips.
2	Reality Defender uses advanced artificial intelligence algorithms to identify and mitigate harmful deepfake and generative content created by bad actors. The platform offers thorough scanning and detection of fraudulent audio and visual content,	Reality Defender specializes in deep face detection services, helping individuals and organizations identify and mitigate the risks associated with deepfake technology. The platform offers ongoing protection against deep fake threats, ensuring that users have access to the	The text does not provide specific information about the costs associated with Reality Defender's services.	Reality Defender's system utilizes AI-driven algorithms and models developed by a leadership team with expertise in data science and cybersecurity. The system is designed to defend against present and future fabrication techniques, indicating that it is adaptable to evolving deep fake threats.

	ensuring a comprehensive approach to deepfake detection. Reality Defender provides actionable results, allowing users to take appropriate measures based on the detected deep fake content.	latest detection technologies and techniques. Reality Defender collaborates with partner clients such as Microsoft, Visa, and Intel, leveraging their expertise and technologies to enhance deepfake detection services.		
3	Analysis of audio files to determine if they are generated by AI or human voice. Ability to upload audio and video files in any format for analysis. Protection against fraud and audio manipulation.	Verification of audio authenticity by distinguishing between AI-generated and human voices. Flexible file format support for convenient analysis. Security measures to protect against fraud and audio manipulation.	Monthly subscription fee: $15.80.	Users start by uploading their audio file to the website. The system accepts audio files in various formats, ensuring convenience and compatibility. Once the audio file is uploaded, aivoicedetector.com employs advanced algorithms to thoroughly analyze the audio. The system considers various features, such as pitch, tone, inflection, and subtle cues specific to AI-generated speech. The algorithms scrutinize these features to assess the likelihood of the audio being produced by AI or a human voice. aivoicedetector.com generates a detailed report presenting the findings. The report provides information about the probability of the audio being AI-generated or human-generated, along with other relevant data points. Users can rely on this report to make informed decisions about the authenticity of the audio file.

7. Discussion

These following topics encompass crucial considerations for the project's implementation and overall effectiveness. We will explore the challenges, potential solutions, and ethical implications associated with building such a system.

- Data Collection and Labeling: The foundation of any AI system lies in the quality and diversity of the data it is trained on. In order to develop an effective voice authentication model, we need to curate a comprehensive dataset comprising both authentic and AI-generated voices. This dataset should be carefully labeled to provide ground truth information for training and evaluation purposes. Ensuring a balanced representation of different voice types, accents, and languages will be vital to ensure the system's robustness and generalizability.

- Feature Extraction and Representation: Extracting meaningful features from voice signals is a critical step in developing a reliable classification model. Techniques such as Mel Frequency Cepstral Coefficients (MFCCs), spectrograms, or deep learning-based methods like convolutional neural networks (CNNs) can be explored to capture the relevant characteristics of human and AI-generated voices. The chosen approach should be able to distinguish between subtle nuances and artifacts introduced by AI synthesis, enabling accurate discrimination between the two.

- Model Selection and Training: The selection of an appropriate machine learning model will significantly impact the performance of our system. Supervised learning techniques, such as support vector machines (SVMs), recurrent neural networks (RNNs), or transformer-based architectures, could be explored for voice classification. The training process should involve optimizing the model's hyperparameters and evaluating its performance using appropriate metrics, such as accuracy, precision, recall, and F1 score. Regularization techniques and data augmentation methods should be employed to mitigate overfitting and enhance the model's generalization capabilities.

- Adversarial Attacks and Countermeasures: As AI-generated voices become more sophisticated, adversarial attacks aimed at deceiving our system will likely emerge. It is crucial to anticipate and address potential vulnerabilities, ensuring that the system remains robust and reliable. Exploring techniques such as adversarial training, anomaly detection, or incorporating domain knowledge can help fortify the system against adversarial attempts, enhancing its resistance to manipulation.

- Ethical Implications and User Privacy: The development and deployment of a voice authentication system carry ethical considerations and raise concerns about user privacy. It is essential to design the system with transparency, accountability, and privacy protection in mind. Clear guidelines should be established to ensure responsible use and prevent misuse of the technology. User consent and data anonymization must be prioritized, and mechanisms for redress and recourse should be in place to address any potential violations or biases.

8. Conclusions

To conclude, we understand the risks, challenges, and crucial factors that are important for the success of developing an AI system to differentiate between real human speech and machine-generated voice files to counter deepfake technology. Our main goal is to create a dependable and adaptable model that offers strong defense against spoof attacks and fraud. However, there are challenges to overcome. Detecting machine-generated speech among background noise and accents is a big hurdle. We need thorough training on diverse datasets and continuous updates to tackle these challenges.

Attaining high accuracy, reliability, and user-friendliness is vital for the success of our voice authentication system. This requires extensive testing on different datasets, covering various languages, accents, and audio qualities. Additionally, we must keep an eye on emerging threats by monitoring and updating the system regularly. Reliability, availability, and seamless integration with existing systems are also important for widespread use.

Furthermore, we must prioritize ethical considerations, compliance with regulations, and user privacy to ensure the system's ethical and legal use. Scalability, performance optimization, and robust security measures are also essential for the system's long-term success. By addressing these risks, challenges, and important factors, we can maximize the impact of our voice authentication system and effectively protect communication channels.

9. Appendices

9.1 This section should include any material, technical or functional specifications, diagrams, and additional relevant content related to the project in general and the proposal document in particular.

9.4 Any other information that assists in describing and understanding the project can be included.

Date: 21.05.23

Participants: Guy Ben Ari, Ynon Friedman, Revital Elgrabli Merom, Professor Yitzhak Lapidot

Subject: Creating AI for Voice Recognition, with emphasis on detecting fake voices compared to genuine voices.

Professor Yitzhak Lapidot, an esteemed professor in the Afeka Speech and Language Center, joined the discussion to provide his insights on the project.

Professor Yitzhak Lapidot, a highly respected member of the Afeka Speech and Language Center, shared his expertise on the matter. He emphasized the distinction between the issues of deep fake and anti-spoofing. Deep fake primarily concerns manipulated recordings with added noises and elements that need to be identified. On the other hand, anti-spoofing focuses on verifying the authenticity of the voice signal, often in scenarios like connecting to a bank account where poor signal quality can lead to rejection.

Regarding voice conversion, Professor Lapidot mentioned its applications in various fields such as criminal or security aspects, gaming, and multimedia. While there are software programs available that can change voices, more research is needed in this area. He suggested expanding the topic of voice conversion to explore its potential further.

In terms of resources, the Afeka Language Processing Center has a vast collection of speech data, but access to it may require coordination with the Computing Center's technical support. It was suggested to reach out to Kfir and Anton for assistance. The Language Processing Center, headed by Vered, may also have relevant individuals who could contribute to the project, although further confirmation is needed.

To gain insights into ASV spoofing, the ASV spoofing 2019 competition and its associated report were recommended for review. The competition is based on a 512-dimensional GMM model, which is lightweight and can be run on personal computers. The competition report mentions several names, including Nick Evans, Maximiliano, a person with a complicated Tomi name, and a Japanese individual. Exploring their work and contributions could be valuable for the project.