This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Stopping Scam Calls with Deepfake Detection (Interdisciplinary Collaboration)

In this caselet, we'll explore audio deepfakes through a real world challenge! Read the background information carefully, and use it to answer the multiple choice questions that follow.

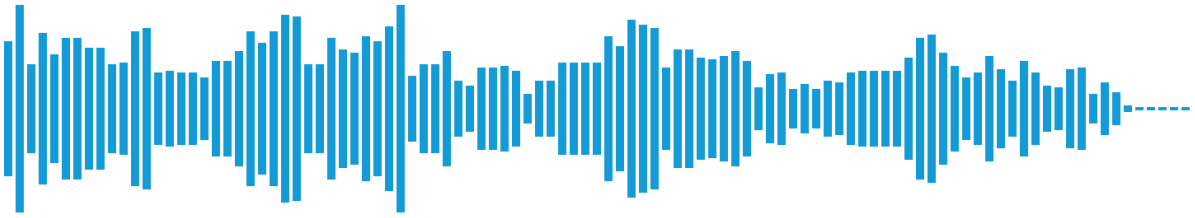
By the end of this caselet, you will craft your own solution to combat the growing threat of audio deepfakes. Good luck!

* Indicates required question

What is an audio deepfake?

An audio deepfake is an artificial audio recording that mimics a person's voice using AI and machine learning techniques. These technologies can be used for malicious activities like impersonation and spreading misinformation.

Curious to hear an audio deepfake? Click Here!



This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Problem Context

You have been hired as a data scientist by a top banking firm. The bank has been experiencing a rise in fraudulent activities using deepfake calls. These calls have used AI to mimic the voices of high-profile clients, often bypassing traditional security measures. To tackle this, the bank wants to develop a new system to detect deepfake voice calls automatically before they make unauthorized transactions.

This bank receives thousands of calls, often with poor audio quality or background noise, so your system will need to navigate these noisy environments.

Curious to hear an audio deepfake? Check out this clip: Click Here!

Data Summary

Checkout this dataset! The link below contains a collection of authentic and deepfake audio clips.

PDF Summary: data_profile

Caselet Questions

1. **Question 1.** * 0 points

   The executives mentioned they want a more explainable model to understand the logic behind your deepfake detection system. How do you plan to address this?

   Mark only one oval.

   - ◯ Build a classical machine-learning model
   - ◯ Build a deep-learning model

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Question 2.

Look at the chart below

Model	Definition
Logistic Regression	The statistical model used for binary classification that estimates the probability an instance belongs to a particular class using a logistic function.
CNN (Convolutional Neural Network)	Processes structured grid data, like audio signals, using convolutional layers to learn hierarchies of features. *A lighter-scale version that uses fewer resources and is less complex is called an LCNN.*
Decision Tree	Makes decisions by splitting the data into subsets based on the value of input features, creating a tree-like structure where each node represents a feature, each branch represents a decision rule, and each leaf represents an outcome.
GAN (Generative Adversarial Network)	Two neural networks, a generator, and a discriminator, compete against each other to generate realistic synthetic data. *A lighter-scale version that uses fewer resources and is less complex is called an LGAN.*
Random Forest	Constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of the individual trees.

2. What kind of models will you consider using for deepfake detection? * 0 points

(Check all that apply)

Check all that apply.

- [ ] Logistic Regression
- [ ] CNN
- [ ] Random Forest
- [ ] GAN
- [ ] Decision Tree

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

3. **Question 3:** * 0 points

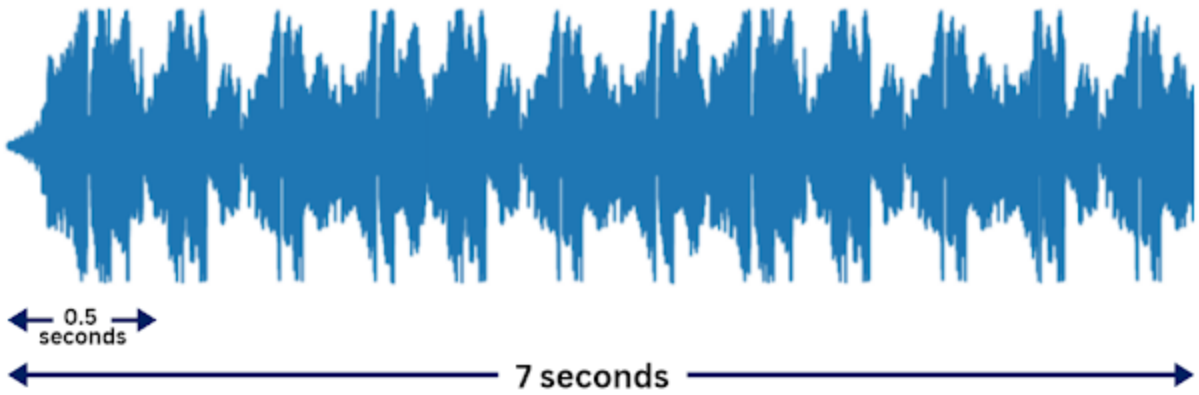
You wonder if a Logistic Regression model is appropriate. But your colleague disagrees. They argue for using a Random Forest model, claiming it's better suited for "*capturing the complex, non-linear relationships found in audio data.*" Which of the following statements is correct?

Mark only one oval.

- ◯ Logistic Regression is better because it handles large amounts of noisy audio data more effectively.
- ◯ Random Forest is better because analyzing non-linear trends is necessary to find differences in deepfake audio
- ◯ Logistic Regression is better because it is more interpretable
- ◯ Random Forest is better because it automatically processes complex audio data in real-time, making it ideal for live applications.

Question 4:

In audio analysis, preprocessing is a critical step that ensures the audio data is prepared for effective analysis and model training. You are preprocessing a 7-second audio clip shown below:

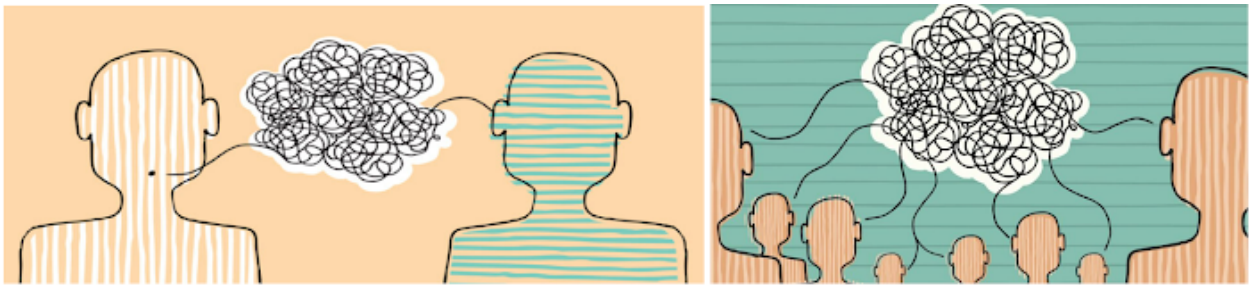


This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

4. You choose a window size of 0.5 seconds. What will the dimensions of the resulting data table be, assuming that each window captures 5 features? * 1 point

Mark only one oval.

- ◯ Rows: 12, Columns: 5
- ◯ Rows: 14, Columns: 5
- ◯ Rows: 10, Columns: 7
- ◯ Rows: 7, Columns: 5



This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Read Carefully

As you dive into your research, you stumble upon insights from the world of *variationist sociolinguistics*—a branch of linguistics that explores how language shifts across different social contexts. Imagine how we speak, not as a rigid structure, but as a dynamic, ever-changing expression, influenced by who we're talking to, where we are, and even how we feel. Human language is full of these subtle variations; we tweak our tone, pitch, and pronunciation in ways we often don't even notice. These phonetic nuances are a complex dance of sounds, both deliberate and intuitive, that reflect the diversity of our communication.

You contact a sociolinguistic expert at a local university, hoping they can offer insights on your dataset to help you develop a new method for detecting deepfake speech. They're excited to help!

Here's what they found:

All genuine clips have *linguistic features* that help differentiate them from deepfake clips.

There are five of them:

Pitch – Higher or lower tone of a speech sample.

Pause – Break in the middle of speech production.

Initial/Final Bursts – Burst of air produced when beginning to pronounce a consonant, or at the end of pronouncing a consonant. If genuine, the bursts create the sounds */p/, /b/, /t/, /d/, /k/,* and */g/*.

Intake/Outtake of Breath – Breathing in or out.

Audio Quality – Overall audio quality of a sample. This includes any disturbances or distortion in the audio.

Let's put these features to use! Listen to this clip you heard previously → here. This time, rate whether you hear an anomalous feature for any of the five features, such as an unusually high pitch where there shouldn't be.

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

5. **Question 5:**

* 0 points

Part A

How would you score this clip on its linguistic features? Give it a **1** if you hear an anomaly, or **0** if you don't.

Mark only one oval.

- ◯ Pitch: 0 Pause: 0 Burst: 1 Breath: 0 Audio Quality: 0
- ◯ Pitch: 0 Pause: 1 Burst: 0 Breath: 0 Audio Quality: 1
- ◯ Pitch: 1 Pause: 1 Burst: 0 Breath: 0 Audio Quality: 1
- ◯ Pitch: 1 Pause: 0 Burst: 0 Breath: 1 Audio Quality: 0

6. **Part B** *

0 points

Based on your rating, is this clip genuine or an audio deepfake?

Mark only one oval.

- ◯ Genuine
- ◯ Deepfake

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

7. **Question 6:** *

Let's try one more! Listen to this clip → here.

Part A

How would you score this clip on its linguistic features? Give it a **1** if you hear an anomaly, or **0** if you don't.

Mark only one oval.

- ◯ Pitch: 1 Pause: 0 Burst: 1 Breath: 0 Audio Quality: 0
- ◯ Pitch: 0 Pause: 1 Burst: 1 Breath: 0 Audio Quality: 1
- ◯ Pitch: 0 Pause: 1 Burst: 1 Breath: 0 Audio Quality: 0
- ◯ Pitch: 0 Pause: 1 Burst: 0 Breath: 0 Audio Quality: 0

Answer Key

Caselet 2 Answer Key

8. **Part B** * 0 points

Is this clip genuine or an audio deepfake?

Mark only one oval.

- ◯ Genuine
- ◯ Deepfake

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

9. **Question 7:** * 0 points

The features we just used make up a method called *Expert Defined Linguistic Features*, or ***EDLFs***. And you just used them to extract features! However, one of the executives is confused. There are other well-known methods, such as **LFCCs** (*Linear Frequency Cepstral Coefficients*), which analyze many unique features of an audio clip over time and use changes in their assigned values to detect anomalies.

"LFCCs can analyze thousands of clips quickly, without us having to manually inspect each one—something like a black box method that can identify anomalies and flag potential issues instantly."

Which feature set do you choose?

Mark only one oval.

- LFCCs because they analyze clips quicker
- EDLFs because they are more explainable
- LFCCs because they are more explainable
- EDLFs because they use less resources

Question 8:

Finally, the moment we've been waiting for—it's time to train and test! You get the training and test error below:

Model	Training Error	Test Error
A	0.10	0.60

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

10. **Part A** * 0 points

What issue is your model facing?

Mark only one oval.

- overfitting
- underfitting
- feature scaling imbalance
- model bias

11. **Part B** * 0 points

What may be causing this issue?

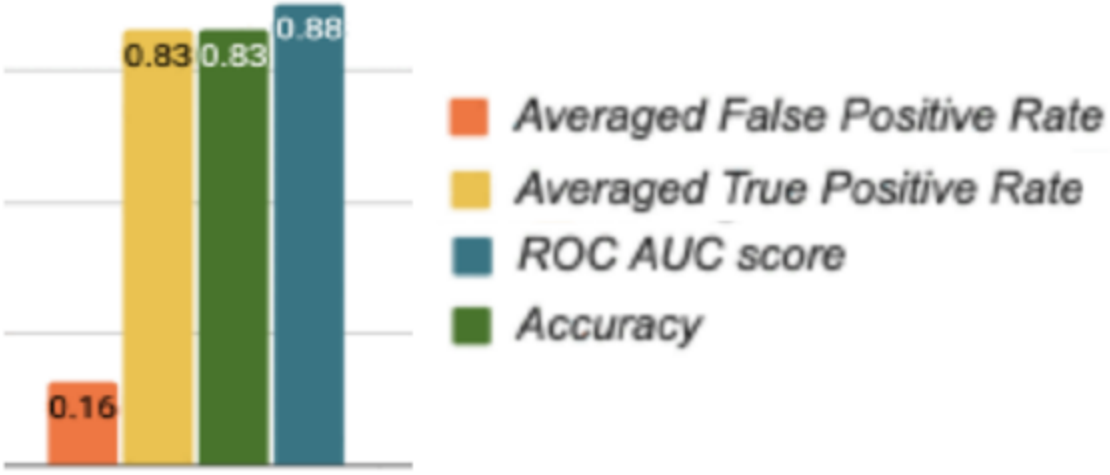
Mark only one oval.

- Insufficient feature engineering
- Too many data points
- Complex model architecture
- Imbalanced dataset

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

Question 9:

After figuring out what was causing our issue, we added some new performance measures in addition to accuracy! And now we're ready to test! The graph below details the metrics below:



12. Assume identifying an audio clip as a deepfake is positive. * 0 points

What is a false negative in this context?

Mark only one oval.

- When the system correctly identifies a deepfake as a fake.
- When the system classifies a deepfake as genuine.
- When the system incorrectly identifies a genuine audio clip as a deepfake.
- When the system classifies an audio clip as genuine.

This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen Chen** at the University of Maryland, Baltimore County (UMBC). For more information, visit cisaad.umbc.edu or email cisaad@umbc.edu.

13. **Question 10:** * 0 points

Assume identifying an audio clip as a deepfake is positive.

What is the cost of error if a false negative occurs?

Mark only one oval.

- ◯ The system correctly identifies a deepfake call but mistakenly blocks a legitimate call, leading to customer dissatisfaction.
- ◯ Damage to relationships with legitimate customers by blocking their transactions.
- ◯ Allowing malicious users to access sensitive information through audio deepfakes.
- ◯ None of the above

Credits

This caselet was created by Kavin Manivannan

The data used in this caselet is adapted from

Master_Data_SampleSet3.xlsx*, under the IEEE License*

Citations:

Z. Khanjani, L. Davis, A. Tuz, K. Nwosu, C. Mallinson and V. P. Janeja, "Learning to Listen and Listening to Learn: Spoofed Audio Detection Through Linguistic Data Augmentation," 2023 IEEE International Conference on Intelligence and Security Informatics (ISI), Charlotte, NC, USA, 2023, pp. 01-06, doi: 10.1109/ISI58743.2023.10297267. keywords: {Training;Annotations;Linguistics;Phonetics;Media;Feature extraction;Security;audio deepfake;spoofed audio detection;Artificial Intelligence;linguistics;sociolinguistics;linguistic perception},

Image Credits:

Image 1 (Header): Rozette Rago

Image 2: WaveVisual

Image 3: Shutterstock