Unmasking Deepfake News Clips

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!

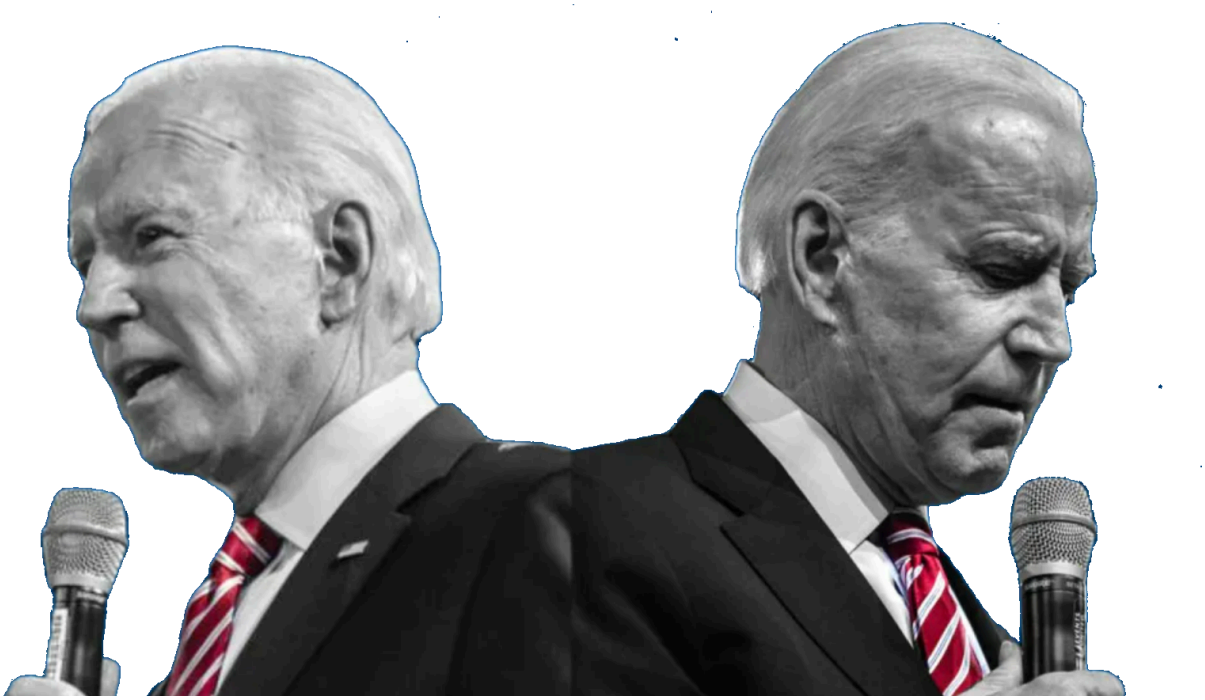
This caselet was developed by student Kavin Manivannan as a resource for Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD), a project funded by the National Science Foundation Award #2346473 and **supervised by** Dr. Vandana P. Janeja, Dr. Christine Mallinson **and Dr. Karen 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 prestigious news network to design a system to detect deepfake audio clips automatically. This news channel prides itself on delivering accurate and reliable news to millions of people. In recent years, however, the rise of audio deepfakes has posed a significant challenge to news networks. These artificial audio recordings, which convincingly mimic the voices of world leaders, CEOs, and other public figures, are becoming increasingly difficult to distinguish from authentic clips.

Amidst a 2024 election cycle, a robocall containing an audio deepfake of President Joe Biden was sent to multiple voters in New Hampshire. The spoofed audio told voters not to vote in an upcoming primary election. “Voting this Tuesday only enables the Republicans in their quest to elect Donald Trump again,” the voice mimicking Biden says. “Your vote makes a difference in November, not this Tuesday.” The spread of deepfake audio clips threatens to sway public opinion with misinformation and undermines public trust and the integrity of the news.

Curious to hear that 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

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.

Caselet Questions

1. **Question 1.** 1 point

The executives mentioned they haven't allocated enough funds for feature engineering for deepfake detection. How do you plan to overcome this challenge?

Mark only one oval.

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

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.

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.

2. What kind of models will you consider using for deepfake detection? * 1 point

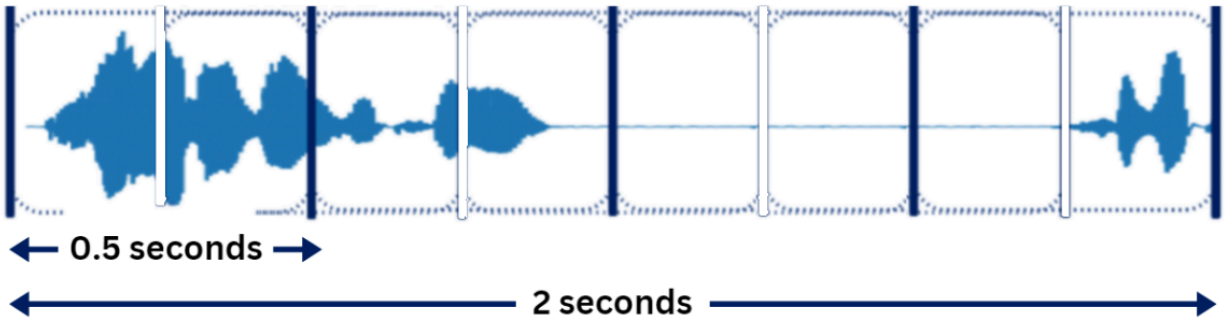
(Check all that apply)

Check all that apply.

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

Question 3:

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 2-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.

3. You choose a window size of 0.5 seconds. What will the dimensions of the resulting data table be, assuming that each window captures 13 features? * 1 point

Mark only one oval.

- ◯ Rows: 2, Columns: 13
- ◯ Rows: 5, Columns: 10
- ◯ Rows: 3, Columns: 13
- ◯ Rows: 4, Columns: 13

There are several methods to extract features from audio data. But you've narrowed it down to two:

Mel Frequency Cepstral Coefficients (MFCCs): Imagine an audio signal as a necklace. Each bead (*coefficient*) on this necklace represents a specific piece of info from an audio clip. To understand the necklace better, we analyze the beads' colors and patterns using a scale that reflects human hearing (*the Mel scale*). We then take these patterns and transform them into a simplified form. We arrange these beads into a new pattern, letting us capture the characteristics of an audio signal.

Linear Frequency Cepstral Coefficients (LFCCs): Similar to MFCCs, except in our necklace, we use *the Linear scale,* where each bead's position directly corresponds to the actual frequencies in the audio signal. This means that each bead's color and pattern represent the raw frequency components of the audio without any perceptual weighting used on the Mel scale.

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. **Question 4:** * 1 point

Each method uses a *filter bank* to extract coefficients. A filter bank acts like a net, capturing frequencies in an audio sample to identify discrepancies that may indicate an audio deepfake. On the Mel scale, a filter bank has more filters to capture data concentrated at lower frequencies and fewer at higher frequencies, mimicking human auditory perception.

Which feature set should you choose and why?

Mark only one oval.

- ◯ MFCCs because LFCCs require more computational resources to process audio data as each filterbank covers a broad frequency range
- ◯ LFCCs because MFCCs may have difficulty detecting higher-frequency voices
- ◯ MFCCs, because they provide a more accurate representation of frequency ranges by basing their filterbank on biological systems like the human ear
- ◯ LFCCs because they are less sensitive to background noise compared to MFCCs.

Question 5:

Our dataset features a variety of audio deepfakes, each with unique characteristics.

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.

Attack Type	Definition
Text-to-Speech	Synthesizing speech from written text using machine learning models.
Voice Conversion	Synthetically transforms one person's voice into another's using machine learning models.
Mimicry	Impersonates another person's voice. Unlike other types, it's a two-channel audio, instead of a monochannel one.
Replay Attacks	Playback of a real person's voice.

Listen to each type below:

Text-to-Speech
Voice Conversion
Mimicry
Replay Attacks

5. Which attack types may be more difficult to detect with your selected features? 1 point

(Check all that apply)

Check all that apply.

- [ ] Text-to-Speech
- [ ] Voice Conversion
- [ ] Mimicry
- [ ] Replay Attacks

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 6:

You train and test your feature set on these deepfake types with three more models:

Model	Training Error	Test Error
A	0.15	0.45
B	0.50	0.60
C	0.10	0.50

6. What is each model exhibiting? 1 point

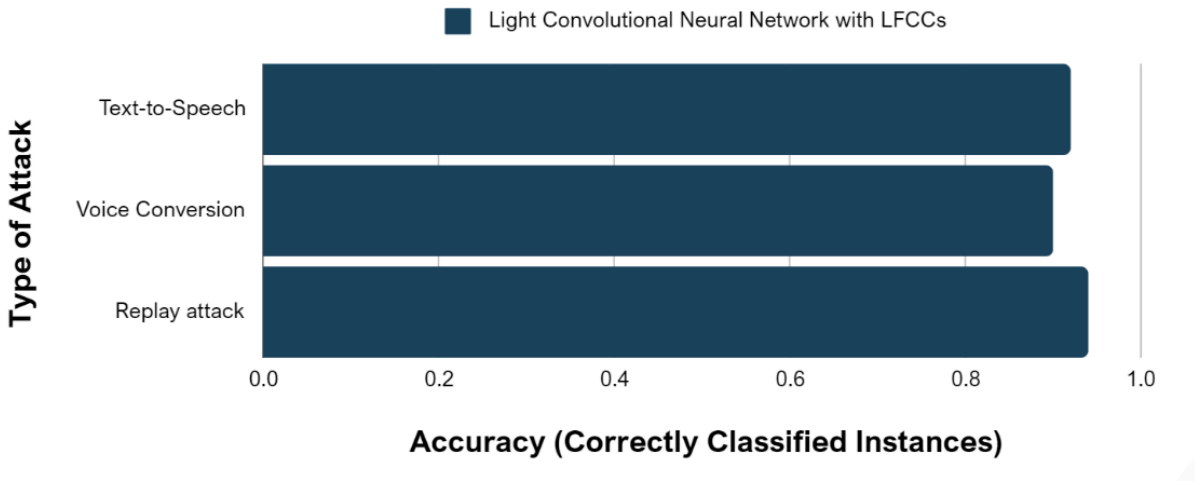
Mark only one oval.

- A exhibits overfitting, B exhibits a well-generalized model, and C exhibits underfitting
- A exhibits a well-generalized model, B exhibits overfitting, and C exhibits underfitting
- A exhibits overfitting, B exhibits underfitting, and C exhibits overfitting
- A exhibits underfitting, B exhibits overfitting, and C also exhibits overfitting

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 7:

Rather than spending more time debugging the current models, you switch gears and train the model you originally chose. You obtain the following results for training:



7. The news executives are astounded—your model achieved 90% accuracy during training across multiple attack types! However, you hold off the celebration. You believe this is a result of overfitting. 1 point

How do you address this issue?

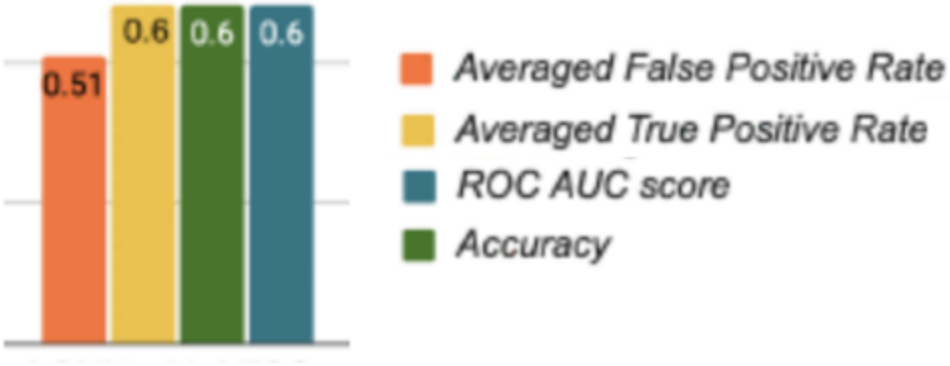
Mark only one oval.

- ◯ Increase the size of the training data
- ◯ Use k-fold cross-validation
- ◯ Simplify the model architecture
- ◯ Select a different 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 8:

Now that we've made changes to our validation methods, we're ready to test! We measure performance by four metrics. The graph below details the metrics below:



8. Assume identifying an audio clip as a deepfake is positive. 1 point

What is a false positive in this context?

Mark only one oval.

- When the system correctly identifies a deepfake as a fake.
- When the system fails to detect a deepfake, it classifies it as genuine.
- When the system finds a genuine audio clip, and classifies it as genuine.
- When the system incorrectly identifies a genuine audio clip as a 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 9:** 1 point

Assume identifying an audio clip as a deepfake is positive.

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

Mark only one oval.

- ◯ Loss of revenue due to blocked legitimate content.
- ◯ Damage to reputation due to falsely accusing individuals or organizations.
- ◯ Increased vulnerability to deepfake attacks due to system compromise.
- ◯ None of the above

Answer Key

Caselet 1 Answer Key

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.

Credits

This caselet was created by Kavin Manivannan

The data used in this caselet is adapted from Baseline_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},

S. P. Dewi, A. L. Prasasti and B. Irawan, "The Study of Baby Crying Analysis Using MFCC and LFCC in Different Classification Methods," 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, 2019, pp. 18-23, doi: 10.1109/ICSIGSYS.2019.8811070. keywords: {Pediatrics;Mel frequency cepstral coefficient;Feature extraction;Maximum likelihood detection;Nonlinear filters;Band-pass filters;DBL;MFCC;LFCC;KNN;VQ},

Swenson, Ali, and Will Weissert. “New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary.” *Associated Press*, 24 Jan. 2024, https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5.

Image Credits:

Image 1 (Header): Shutterstock

Image 2: WaveVisual

Image 3: The Washington Post