Real-Time Detection of AI-Generated Speech For Deepfake Voice Conversion

Created by	Trí Trần
Created time	@May 1, 2024 12:09 PM

Số	Nội dung	Liên kết
1	Link paper	https://arxiv.org/pdf/2308.12734v1
2	Nguồn data	https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition
3	Victory Speech by Joe Biden	https://www.youtube.com/watch?v=1AfNYztas2c (Last accessed: 07/23)
4	Golden Globes Speech by Ryan Gosling	https://www.youtube.com/watch?v=K8JLyUW_MSw (Last accessed: 07/23)
5	Commencement Speech by Elon Musk	https://www.youtube.com/watch?v=MxZpaJK74Y4 (Last accessed: 07/23)
6	Victory Speech by Barack Obama	https://www.youtube.com/watch?v=leCY-jKpoZ0 (Last accessed: 07/23)
7	BAFTAs Speech by Margot Robbie	https://www.youtube.com/watch?v=-JA3_QBfjG8 (Last accessed: 07/23)
8	Stepping Down Monologue by Linus Sebastian	https://www.youtube.com/watch?v=0vuzqunync8 (Last accessed: 07/23)
9	Women in Music Speech by Taylor Swift	https://www.youtube.com/watch?v=ZVpkFb9-fts (Last accessed: 07/23)
10	Victory Speech by Donald Trump	https://www.youtube.com/watch?v=Qsvy10D5rtc (Last accessed: 07/23)
11	Implementation	https://github.com/RVC-Project/Retrieval-based-

	of RVC	Voice-Conversion-WebUI
12	Information on Huggingface Model Hub	https://huggingface.co/models
13	Information on the AI Hub	https://discord.me/aihub

Abstract

Nghiên cứu này tạo ra tập dữ liệu DEEP-VOICE, bao gồm giọng nói thực và giọng nói được tạo ra bằng trí tuệ nhân tạo từ tám nhân vật nổi tiếng. Sử dụng phân tích thống kê, họ phát hiện ra sự khác biệt đáng kể giữa các đặc trưng âm thanh thời gian của giọng nói thực và giọng nói được tạo ra bằng trí tuệ nhân tạo. Họ cũng tối ưu hóa các mô hình học máy và đưa ra một mô hình Extreme Gradient Boosting có khả năng phân loại giọng nói với độ chính xác trung bình là 99.3% trong thời gian thực. Dữ liệu nghiên cứu này được công khai để hỗ trợ nghiên cứu tiếp theo về phát hiện giọng nói được tạo ra bằng trí tuệ nhân tạo.

1. Introduction

Công nghệ Trí tuệ nhân tạo (AI) hiện đang phát triển mạnh mẽ, cho phép chuyển đổi giọng nói thời gian thực. Mặc dù có thể hấp dẫn về mặt giải trí, nhưng nó đặt ra nguy cơ an ninh lớn khi có thể bị lạm dụng để xâm phạm quyền riêng tư và gây ra trộm danh tính. Cần có các giải pháp khoa học ngay lập tức để giải quyết vấn đề này.

Công trình này đóng góp ba điều quan trọng:

1. Tạo ra một tập dữ liệu phân loại âm thanh mới, bao gồm 8 nhân vật nổi tiếng, với âm thanh thực và giọng nói AI được tạo ra bằng phương pháp Retrieval-based Voice Conversion (RVC).
2. Phân tích thống kê các đặc trưng âm thanh để xác định những đặc trưng quan trọng trong việc phân loại giọng nói của con người và giọng nói AI.
3. Tối ưu hóa siêu tham số của các mô hình Học máy thống kê để cải thiện độ chính xác và thời gian suy luận, đảm bảo khả năng nhận diện giọng nói AI trong thời gian thực. Các mô hình này có thể được sử dụng trong các hệ

thống cảnh báo để ngăn chặn việc sử dụng giọng nói tổng hợp với mục đích đen tối trong các cuộc gọi điện thoại hoặc hội nghị.

Bên cạnh đó, bài báo nghiên cứu này cũng đóng góp bộ dữ liệu DEEP-VOICE1 để cho phép phân tích giọng nói được tạo ra bằng trí tuệ nhân tạo. Các tập dữ liệu được thu thập và tạo ra trong nghiên cứu này được công khai cho cộng đồng nghiên cứu, nhằm hỗ trợ công việc đa ngành về phân tích và nhận diện mẫu hình giọng nói được tạo ra bằng trí tuệ nhân tạo. Nghiên cứu này cũng khám phá các biện pháp phòng ngừa dựa trên Học máy chống lại sự giả mạo giọng nói tổng hợp.

Phần còn lại của nghiên cứu này được tổ chức như sau: Phần 2 trước hết khám phá tài liệu khoa học liên quan đến nghiên cứu này trước khi Phần 3 trình bày phương pháp nghiên cứu cũng như các thí nghiệm trong công việc này. Kết quả thí nghiệm sau đó được trình bày và thảo luận trong Phần 4. Cuối cùng, công việc này được kết luận cùng với đề xuất cho công việc tương lai dựa trên các kết quả của các nghiên cứu trong Phần 5.

2. Background

Phần này khám phá về deepfakes và phương pháp phát hiện chúng, đồng thời thảo luận về ảnh hưởng của tài liệu khoa học đối với phương pháp thí nghiệm trong nghiên cứu. Deepfakes là thuật ngữ chỉ một loại thuật toán có khả năng tạo ra phương tiện tổng hợp với mục đích thay thế hình ảnh của một cá nhân bằng một cá nhân khác, gây ra nhiều vấn đề đạo đức và pháp lý. Nghiên cứu này tập trung vào sao chép giọng nói và cách phát hiện chúng.

Nghiên cứu này bắt đầu từ ý tưởng của Beard (2001) về việc sao chép âm thanh và hình ảnh của một diễn viên sau khi họ qua đời. Buzzfeed (2018) sử dụng FakeApp để thay đổi giọng nói của Jordan Peele thành Barack Obama, thu hút sự chú ý đến tính chân thực của phương tiện tổng hợp. Google giới thiệu Tacotron vào năm 2017, và nghiên cứu tiếp theo tiếp tục cải tiến phương pháp nhận diện giọng nói tổng hợp.

Nghiên cứu của Lim, Suk-Young và Lee (2022) chỉ ra rằng CNNs và LSTM có thể đạt được độ chính xác từ 97 đến 99% trong việc nhận diện giọng nói tổng hợp. Cần lưu ý rằng các phương pháp tích chập thời gian tương đối tốn kém về mặt tính toán.

Một nghiên cứu (17) chỉ ra rằng một mạng nơ-ron tích chập dư (residual CNN) đạt được tỉ lệ lỗi thấp nhất cho thách thức ASVspoof 2019, với tỷ lệ lỗi 4.04%, sau đó giảm xuống còn 1.26%. Một nghiên cứu khác (18) sử dụng các mạng

nơ-ron tích chập để học từ hình ảnh được tạo ra từ các dạng âm thanh khác nhau và chỉ ra rằng một mạng nơ-ron tích chập VGG-16 đạt được khoảng 85.91% độ chính xác trong việc phát hiện deepfake. Nghiên cứu này cũng chỉ ra rằng các tiến bộ gần đây trong giọng nói deepfake làm cho việc phát hiện chúng trở nên khó khăn. Một nghiên cứu khác (19) đề xuất rằng các đặc trưng từ Nhận dạng Cảm xúc Giọng nói (SER) có thể được sử dụng để phát hiện giọng nói tạo ra từ trí tuệ nhân tạo, và thí nghiệm trên tập dữ liệu ASVSpoof2019 cho thấy rằng việc học chuyển giao từ mô hình SER có thể cải thiện việc phân loại giọng nói tạo ra từ trí tuệ nhân tạo.

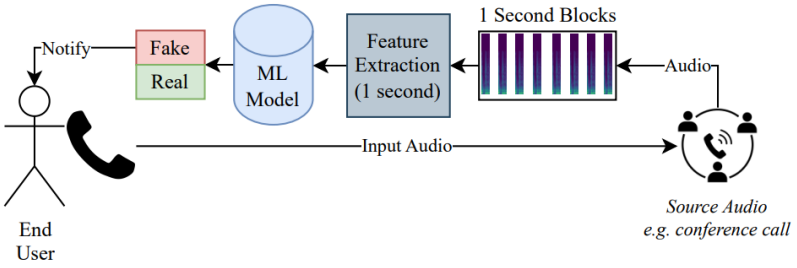


Figure 1: Usage of the real-time system. The end user is notified when the machine learning model has processed the speech audio (e.g. a phone or conference call) and predicted that audio chunks contain AI-generated speech.

Surname	Speech	Length (MM:SS)
Biden	Victory Speech	10:00
Gosling	Golden Globes Speech	1:33
Musk	Commencement Speech	10:00
Obama	Victory Speech	10:00
Robbie	BAFTAs Speech	1:19
Sebastian	Stepping Down Monologue	9:30
Swift	Women in Music Speech	10:00
Trump	Victory Speech	10:00
Total		**62:22**

Table 1: Data collected for training, validation, and unseen testing for the experiments in this work (sorted alphabetically by surname). Audio segments are cropped to a maximum of ten minutes.

Nghiên cứu đề xuất sử dụng các đặc trưng âm thanh như chromagrams, spectrograms, mel-spectrum và mel-frequency cepstral coefficients, lấy cảm hứng từ nghiên cứu trước, và xem xét về tính phức tạp tính toán. Mặc dù CNN

và LSTM hiệu quả, nhưng chúng có thể không đảm bảo dự đoán thời gian thực cho người dùng. Do đó, nghiên cứu tập trung vào tối ưu hóa các thuật toán thống kê về cả thời gian dự đoán và độ chính xác.

3. Method

Phần này cung cấp một tổng quan về phương pháp được sử dụng trong nghiên cứu này. Một tổng quan về việc thu thập và tiền xử lý dữ liệu được cung cấp trước khi đi vào chi tiết về phương pháp chuyển đổi giọng nói DeepFake được áp dụng. Tiếp theo, chi tiết về các mô hình máy học (ML) và quá trình tối ưu hóa của chúng được cung cấp. Cuối cùng, chi tiết về phần cứng và phần mềm được sử dụng cho các thí nghiệm được mô tả với mục đích tái tạo.

Câu hỏi nghiên cứu chính của nghiên cứu này là làm thế nào để phát hiện Giọng nói Tổng hợp trong thời gian thực và thông báo cho người dùng cuối một cách phù hợp. Biểu đồ trong Hình 1 minh họa một trường hợp sử dụng mà hệ thống được đề xuất có thể được triển khai. Âm thanh nguồn, chẳng hạn như một cuộc gọi điện thoại hoặc cuộc họp trực tuyến, được xử lý và phân loại. Nếu dự đoán rằng âm thanh chứa giọng nói được tạo ra bởi trí tuệ nhân tạo, người dùng cuối sẽ nhận được thông báo.

3.1 Data Collection and Preprocessing

Dữ liệu gốc ban đầu: Tám cá nhân được chọn với nguồn âm thanh thực và dữ liệu để chuyển đổi thành giọng nói được tạo ra bởi trí tuệ nhân tạo. Tổng cộng, thu thập 62 phút và 22 giây âm thanh từ tám cá nhân, với mỗi đoạn âm thanh được giới hạn tối đa là mười phút. Đa dạng chất lượng âm thanh từ chất lượng sản xuất đến chất lượng thấp được lựa chọn để tạo sự đa dạng trong tập dữ liệu.

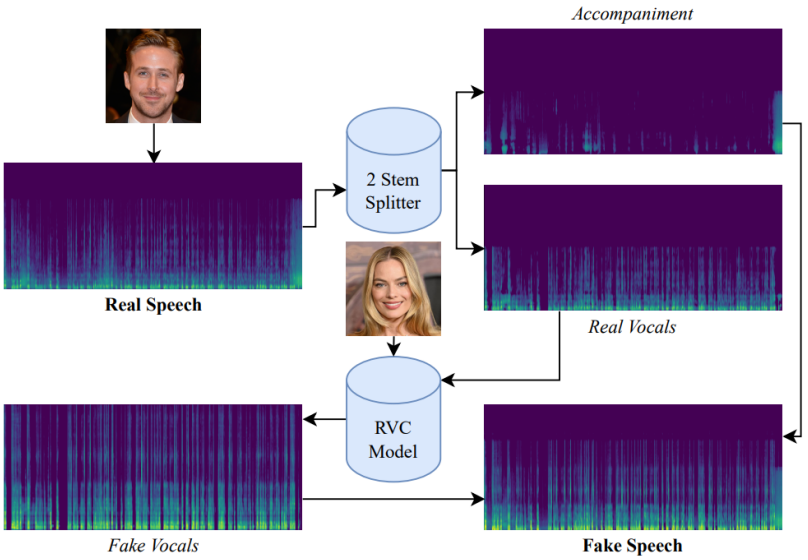


Figure 2: Overview of the Retrieval-based Voice Conversion process to generate DeepFake speech with Ryan Gosling's speech converted to Margot Robbie. Conversion is run on the extracted vocals before being layered on the original background ambience.

Một ví dụ về chuyển đổi giọng nói từ thật sang giả có thể được tìm thấy trong Hình 2. Đầu tiên, giọng nói thực được phân tách thông qua mô hình hai hạt [21] từ Spleeter, đây là một mạng nơ-ron tích chập (CNN) mã hóa-giải mã trong kiến trúc U-Net. Mô hình bao gồm 12 lớp, với 6 lớp cho mỗi mạng mã hóa và giải mã. Sau khi phân tách các bản thu thực và bản đệm, các bản thu được chuyển đổi bằng một mô hình Chuyển đổi Giọng nói Dựa trên Truy xuất (RVC) sang một cá nhân khác. Cuối cùng, bản đệm gốc và các bản thu RVC được kết hợp để tạo thành một bản thu giả. Lý do phân tách các bản thu là để đảm bảo rằng phong cách của giọng nói giả không bị chuyển đổi sang bất kỳ tiếng ồn nền nào, chẳng hạn như tiếng cổ vũ hoặc tiếng cười của khán giả. Nói cách khác, mục tiêu của phương pháp này là bảo tồn âm thanh môi trường trong khi chỉ chuyển đổi giọng nói của người nói.

Trong quá trình chuyển đổi phong cách được mô tả, mỗi đoạn âm thanh thu thập và cắt từ Bảng 1 được sử dụng và các đặc trưng được trích xuất cho mỗi 1 giây của tín hiệu âm thanh. Tổng cộng, 26 đặc trưng được trích xuất, bao gồm Chromagram, Spectral Centroid (SC), Spectral Bandwidth (SB), Spectral Rolloff (SR), Zero Crossing Rate (ZCR), Root Mean Square (RMS) và 20 hệ số Cepstral tần số Mel (MFCCs).

- **Chromagram: Được tính từ Biến đổi Fourier ngắn hạn của tín hiệu âm thanh và được chuẩn hóa thành các dải chroma.**
- **Spectral Centroid (SC): Vị trí trọng tâm trong phổ.**

- **Spectral Bandwidth (SB): Độ chênh lệch trong tần số xung quanh trọng tâm.**
- **Spectral Rolloff (SR): Tần số dưới 85% tổng năng lượng phổ.**
- **Zero Crossing Rate (ZCR): Tần số tín hiệu thay đổi dấu.**
- **Root Mean Square (RMS): Căn bậc hai của trung bình bình phương của tín hiệu âm thanh.**
- **20 hệ số Cepstral tần số Mel (MFCCs): Được tính từ Biến đổi Fourier ngắn hạn của tín hiệu âm thanh, sau đó được chuyển đổi sang tỉ lệ Mel và áp dụng Biến đổi Cosin rời rạc (DCT).**

Đúng, các đặc trưng được trích xuất từ tín hiệu âm thanh như Chromagram, Spectral Centroid, Spectral Bandwidth, Spectral Rolloff, Zero Crossing Rate, Root Mean Square và MFCCs thường được sử dụng để phân tích và nhận biết âm thanh, bao gồm cả âm thanh được tạo ra bởi trí tuệ nhân tạo (AI). Các đặc trưng này có thể cung cấp thông tin quan trọng về các đặc điểm của âm thanh, bao gồm cả thông tin về âm sắc, tần số, và biên độ, giúp phân biệt giữa âm thanh thực và âm thanh được tạo ra bởi AI.

Mỗi cá nhân được sử dụng để tạo ra bảy bản ghi âm giả mạo. Sự mất cân bằng này có thể ảnh hưởng đến hiệu suất của mô hình phân loại, do đó cần phải cân bằng lại dữ liệu. Để làm điều này, dữ liệu trong lớp giả mạo được giảm mẫu (**undersampling**) để có tỷ lệ 1:1 giữa giọng nói thực và giả mạo. Một mẫu có chiều dài bằng với dữ liệu thực được chọn ngẫu nhiên và sử dụng cho quá trình phân tích và phân loại tiếp theo.

3.2 DeepFake Voice Conversion

Phương pháp chuyển đổi giọng nói được sử dụng là mô hình RVC, dựa trên kiến trúc VITS. Mô hình này được chọn vì khả năng chuyển đổi nhanh chóng và có thể sử dụng trong thời gian thực. Các mô hình cho các cá nhân được nghiên cứu được lấy từ Huggingface Model Hub, bao gồm các nhân vật nổi tiếng như Tổng thống Joe Biden, Donald Trump, Barack Obama, Elon Musk, Linus Sebastian, Taylor Swift, Ryan Gosling và Margot Robbie.

- Joe Biden: 500 epochs

- Donald Trump: 600 epochs
- Barack Obama: 300 epochs
- Elon Musk: 350 epochs
- Linus Sebastian: 300 epochs
- Taylor Swift: 300 epochs
- Ryan Gosling: 350 epochs
- Margot Robbie: 350 epochs

3.3 Machine Learning Model

Sau khi trích xuất đặc trưng từ mỗi khối âm thanh có độ dài 1 giây, nghiên cứu này triển khai một loạt các mô hình học máy khác nhau. Mục tiêu của các mô hình là thực hiện phân loại nhị phân của âm thanh, học xem âm thanh có phải là giọng nói tự nhiên của một con người hay đã được thay đổi bởi việc chuyển đổi giọng nói dựa trên việc tìm kiếm. Các mô hình được lựa chọn từ một loạt các phương pháp thống kê khác nhau để so sánh.

Các mô hình được sử dụng:

1. Extreme Gradient Boosting (XGBoost) [25]
2. Random Forests [26]
3. Quadratic and Linear Discriminant analyses [27]
4. Ridge Regression [28] (linear regression with L2 regularization)
5. Gaussian and Bernoulli Naive Bayes [29]
6. K-Nearest Neighbors [30]
7. Support Vector Machines [31]
8. Stochastic Gradient Descent [32]

Ngoài ra, còn có một mô hình tùy chỉnh được gọi là "Bird & Lotfi: Real-time Detection of AI-Generated Speech.”

Ngoài ra, còn có Mô hình Quy trình Gaussian [33]. Bên cạnh các chỉ số phân loại, thời gian suy luận cũng được xem xét, vì phát hiện giọng nói được tạo ra bởi trí tuệ nhân tạo có thể hữu ích trong thời gian thực trong cuộc họp hoặc cuộc gọi điện thoại. Thời gian suy luận trung bình được tính toán bằng cách đo lường và lấy trung bình thời gian suy luận của 1000 đối tượng dữ liệu ngẫu nhiên trong tập dữ liệu. Đối với tối ưu siêu tham số, XGBoost và Random Forests được tối ưu thông qua tìm kiếm tuyến tính của {10, 20, 30, ..., 500} vòng tăng cường và kích thước rừng, tương ứng. Không gian vấn đề K-Nearest Neighbor được tìm kiếm theo cách tương tự, với kích thước tập láng giềng là {1, 2, 3, ..., 100}.

Điều này cho thấy cách tiếp cận để phát hiện giọng nói được tạo ra bởi trí tuệ nhân tạo trong thời gian thực. Bằng cách sử dụng một loạt các đặc trưng âm thanh và mô hình học máy, nghiên cứu này đề xuất một phương pháp để nhận biết giọng nói tự nhiên và giọng nói được tạo ra bởi AI. Các phương pháp đánh giá và so sánh hiệu suất của các mô hình cũng được mô tả, bao gồm cả các chỉ số đánh giá như **precision, recall, F1 score, và Matthews Correlation Coefficient (MCC)**. Điều này giúp đánh giá hiệu suất của các mô hình không chỉ dựa trên độ chính xác cổ điển mà còn trên các khía cạnh khác của việc phân loại âm thanh.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}. \tag{6}$$

Precision is important since a high precision would minimise false accusations of AI-generated speech when the audio is, in fact, natural voice. Similarly, recall which is a measure of how many positive cases are correctly predicted, which enables analysis of false-negative predictions:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}. \tag{7}$$

Higher recall suggests that the model is not falsely classifying AI-generated speech as human speech. These results are then combined to compute the F-1 score:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

The Matthews Correlation Coefficient (MCC) is then considered, which is a metric that considers all potential correct and incorrect predictions, calculated as:

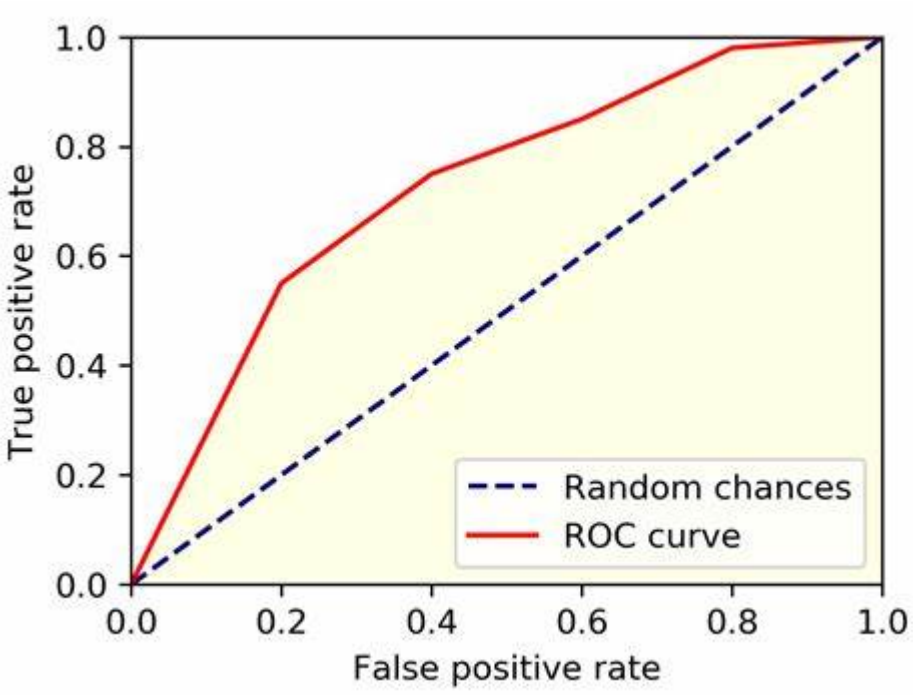
$$\text{MCC} = \frac{\text{T} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}. \tag{9}$$

MCC is measured in the range of -1 to 1. -1 is the complete disagreement between predictions and labels, +1 is a perfect classifier, and an MCC of 0 suggests random predictions.

Cuối cùng, Diện tích dưới đường cong ROC (ROC AUC) được tính toán. ROC AUC quan trọng đối với các phương pháp dự đoán lớp dựa trên xác suất, vì nó

là một kiểm tra về khả năng dự đoán qua các ngưỡng xác suất. Đường cong ROC là biểu đồ của recall và tỉ lệ false positives dựa trên các ngưỡng này, và AUC là đo lường của diện tích dưới đường cong đã vẽ. Chỉ số này được đánh giá từ 0 đến 1, trong đó 0.5 là một bộ phân loại ngẫu nhiên và 1 là hoàn hảo. Mỗi mô hình được huấn luyện qua 10-fold cross validation, với các phân chia dữ liệu được thiết lập để có thể tái lập và so sánh trực tiếp thông qua một giá trị seed ngẫu nhiên là 42.

1. **Receiver Operating Characteristic (ROC) curve**: Đây là một biểu đồ biểu diễn hiệu suất của một mô hình phân loại ở nhiều ngưỡng khác nhau. Trên trục X, chúng ta có tỷ lệ dự đoán sai (False Positive Rate), và trên trục Y là tỷ lệ dự đoán đúng (True Positive Rate). Mỗi điểm trên đường cong ROC tương ứng với một ngưỡng dự đoán khác nhau.
2. **Area Under the Curve (AUC)**: Là diện tích nằm dưới đường cong ROC. Nó đo lường khả năng của một mô hình phân loại phân biệt giữa các lớp. Một mô hình với ROC AUC gần 1 được coi là một mô hình tốt, trong khi AUC gần 0.5 chỉ ra rằng mô hình không phân biệt được nhiều giữa các lớp hoặc dự đoán ngẫu nhiên.
3. **Cross-validation**: Là một kỹ thuật đánh giá hiệu suất của mô hình trên dữ liệu mà nó chưa từng thấy trước đó. Trong kỹ thuật 10-fold cross-validation, dữ liệu được chia thành 10 phần bằng nhau, mỗi phần được sử dụng lần lượt làm tập kiểm tra trong khi các phần còn lại dùng để huấn luyện. Quá trình này được lặp lại 10 lần để đảm bảo mọi phần dữ liệu đều được sử dụng để kiểm tra và huấn luyện.



3.4 Experimental Hardware and Software

Cấu hình máy tính và công cụ được sử dụng trong các thí nghiệm:

1. **CPU và GPU**: Các thí nghiệm được thực hiện trên một máy tính dùng chip Intel Core i7 với tốc độ xung nhịp là 3.7GHz. Việc chuyển đổi giọng nói được thực hiện trên một GPU với sự hỗ trợ của thuật toán CREPE, và một Nvidia RTX 2080Ti được sử dụng cho mục đích này với 4.352 lõi CUDA.
2. **Thư viện và framework**: Đặc trưng được trích xuất từ âm thanh bằng thư viện Librosa, một thư viện phổ biến cho xử lý âm thanh trong Python. Các mô hình máy học được triển khai bằng thư viện scikit-learn, một thư viện mã nguồn mở phổ biến cho học máy trong Python.

4. Result and Observations

Dưới đây là các quan sát được thực hiện và kết quả được tìm thấy trong các thí nghiệm.

4.1 Dataset Analysis

Trước khi triển khai Machine Learning, Bảng 2 thể hiện các thống kê quan sát được giữa tập dữ liệu thật và giả. Có sự khác biệt lớn về một số đặc điểm giữa giọng nói giả và giọng nói thật

Table 2: Observed statistics in the dataset between the two classes of data.

Attribute	Real			Fake		
	Mean	*Med.*	*Std.*	*Mean*	*Med.*	*Std.*
Chromagram	0.41	0.40	0.07	0.43	0.43	0.07
Root Mean Square	0.04	0.03	0.03	0.04	0.03	0.02
Spectral Centroid	2541.34	2353.97	1,253.73	2897.06	2,762.58	800.40
Spectral Bandwidth	2883.99	2806.94	1,080.81	3216.61	3193.74	545.97
Rolloff	4683.44	4094.33	2602.42	5271.79	5061.05	1572.49
Zero Crossing Rate	0.06	0.05	0.04	0.08	0.07	0.03
MFCC 1	-376.65	-354.28	84.12	-388.48	-378.02	74.31
MFCC 2	158.25	162.36	40.18	131.86	132.94	25.62
MFCC 3	-29.60	-24.75	31.77	-19.80	-16.72	21.92
MFCC 4	14.66	11.63	23.63	27.96	27.38	19.07
MFCC 5	-6.33	-11.72	23.88	-6.31	-5.15	15.61
MFCC 6	3.40	5.17	15.01	11.41	12.48	12.54
MFCC 7	-8.61	-7.29	12.31	-10.37	-11.01	10.49
MFCC 8	-7.40	-6.79	10.82	-4.73	-4.59	7.23
MFCC 9	-8.41	-8.28	11.28	-3.48	-4.31	8.04
MFCC 10	-10.98	-13.21	8.74	-7.26	-6.84	8.81
MFCC 11	-3.27	-3.20	7.53	-1.21	-1.72	7.78
MFCC 12	-5.88	-5.45	6.95	-3.00	-2.94	5.93
MFCC 13	-2.07	-1.99	4.52	-1.25	-0.90	5.63
MFCC 14	-1.67	-1.64	5.27	-2.55	-2.88	5.39
MFCC 15	-3.06	-2.88	4.80	-2.16	-1.93	4.97
MFCC 16	-2.29	-1.07	5.59	-0.99	-0.62	5.59
MFCC 17	-3.09	-2.95	4.53	-3.55	-3.48	4.65
MFCC 18	-4.85	-4.05	4.91	-1.38	-1.81	4.40
MFCC 19	-1.79	-1.99	4.40	-3.72	-3.41	5.29
MFCC 20	-4.08	-3.29	6.23	-4.77	-4.22	4.59

Centroid của dữ liệu giả là 2897.06 trong khi đối với dữ liệu thật thấp hơn nhiều, chỉ ở mức 2541.34. Giá trị trung bình của hai MFCC đầu tiên cũng khác biệt đáng kể giữa các lớp. Tuy nhiên, một số tập hợp đặc điểm có sự tương đồng, như MFCC thứ 5.

Table 3: Results of the unpaired t-test for each attribute between the real and fake classes of data.

Attribute	T-Statistic	P-Value	Significance?
Chromagram	-17.488	1.25E-67	Y
Root Mean Square	7.799	6.78E-15	Y
Spectral Centroid	-18.351	3.52E-74	Y
Spectral Bandwidth	-21.078	7.70E-97	Y
Rolloff	-14.848	2.02E-49	Y
Zero Crossing Rate	-17.173	2.65E-65	Y
MFCC 1	8.087	6.69E-16	Y
MFCC 2	42.5	0.00E+00	Y
MFCC 3	-19.467	4.24E-83	Y
MFCC 4	-33.626	8.95E-237	Y
MFCC 5	-0.05	9.61E-01	N
MFCC 6	-31.418	3.89E-208	Y
MFCC 7	8.349	7.65E-17	Y
MFCC 8	-15.774	1.74E-55	Y
MFCC 9	-27.323	2.01E-159	Y
MFCC 10	-23.012	1.17E-114	Y
MFCC 11	-14.627	4.98E-48	Y
MFCC 12	-24.232	1.25E-126	Y
MFCC 13	-8.665	5.08E-18	Y
MFCC 14	8.989	2.88E-19	Y
MFCC 15	-9.949	3.14E-23	Y
MFCC 16	-12.613	3.08E-36	Y
MFCC 17	5.345	9.22E-08	Y
MFCC 18	-40.388	0.00E+00	Y
MFCC 19	21.553	4.37E-101	Y
MFCC 20	6.894	5.69E-12	Y

Tiếp theo, Bảng 3 cho thấy kết quả của kiểm định t không đôi giữa các tập dữ liệu. Có thể thấy rằng **tất cả các đặc điểm, ngoại trừ MFCC thứ 5 như đã đề cập, đều có ý nghĩa thống kê giữa hai lớp dữ liệu.** Với hầu hết các đặc điểm có giá trị trung bình khác biệt đáng kể giữa hai lớp, chúng có thể là cơ sở quan trọng để phân biệt và do đó hữu ích để huấn luyện các mô hình máy học.

Trong kiểm định t, T-Statistic (còn được gọi là giá trị thống kê t) là một số đo mức độ khác biệt giữa các giá trị trung bình của hai mẫu dữ liệu. Nó được tính toán bằng cách chia sự khác biệt giữa giá trị trung bình của hai nhóm cho sai số tiêu chuẩn của sự khác biệt đó. T-Statistic càng cao, tức là sự khác biệt giữa hai nhóm càng lớn.

P-Value (giá trị p) là xác suất có điều kiện của một kết quả quan sát hoặc một kết quả mạch thí nghiệm, dựa trên giả định rằng giả thuyết không đúng. Nó cung cấp một cách để đánh giá xem liệu sự khác biệt giữa các mẫu dữ liệu có thực sự đáng kể hay không. Nếu giá trị p nhỏ hơn một ngưỡng xác định trước

(thường là 0.05), chúng ta có thể kết luận rằng có sự khác biệt đáng kể giữa hai nhóm. Ngược lại, nếu giá trị p lớn hơn ngưỡng này, chúng ta không có đủ bằng chứng để bác bỏ giả thuyết không đúng và coi sự khác biệt là ngẫu nhiên.

Table 4: Metrics when using a single rule-based classifier to split predictions via the 2^{nd} Mel Frequency Cepstral Coefficient. Overall, using this feature, a mean accuracy of 69.84% was observed over 10-fold cross validation.

Class	Metric				
	Precision	***Recall***	***F1-Score***	***MCC***	***ROC***
Real	0.708	0.677	0.692	0.397	0.698
Fake	0.690	0.720	0.705	0.397	0.698
Weighted Average	0.699	0.698	0.698	0.397	0.698

Bảng 4 cung cấp các chỉ số thu được bằng cách sử dụng đặc điểm có mức tương quan cao nhất, đó là hệ số **MFCC thứ hai với hệ** số tương quan Pearson là 0.36, để chia dữ liệu để dự đoán. Kết quả cho thấy rằng bằng cách chia dữ liệu dựa trên thuộc tính này một mình, đạt được độ chính xác trung bình là 69.84% ⇒ **Chỉ riêng một chỉ số MFCC thứ 2 này đã đạt được độ chính xác này rồi**

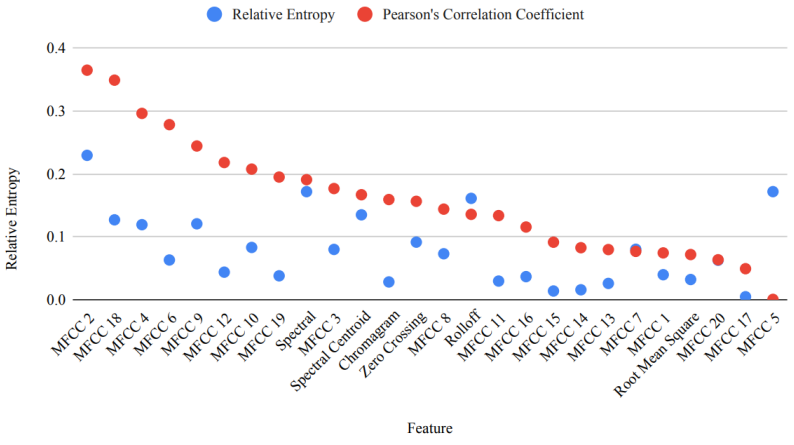


Figure 3: Pearson's Correlation Coefficient and Relative entropy for all of the extracted features when used for binary classification of real or AI-generated vocals (sorted by Pearson's).

Hình 3 trình bày các hệ số tương quan và entropy tương đối của các đặc trưng khi sử dụng cho việc phân loại âm thanh là thật hay được tạo ra bởi trí tuệ nhân tạo. **MFCC thứ 2 có hệ số tương quan cao nhất với lớp dữ liệu, tiếp theo là MFCC thứ 18. Những đặc trưng MFCC thứ 20, thứ 17 và thứ 5 có hệ số tương**

quan thấp nhất. MFCC thứ 2 cũng có entropy tương đối cao nhất. **Điều này cho thấy MFCC thứ 2 có ảnh hưởng lớn nhất đến việc phân loại âm thanh.**

4.2 Hyperparameter Optimisation

Trong phần tối ưu siêu tham số, các kết quả cho các mô hình phân cụm, **Random Forest và XGBoost** được mô tả chi tiết.

4.2.1 KNN:

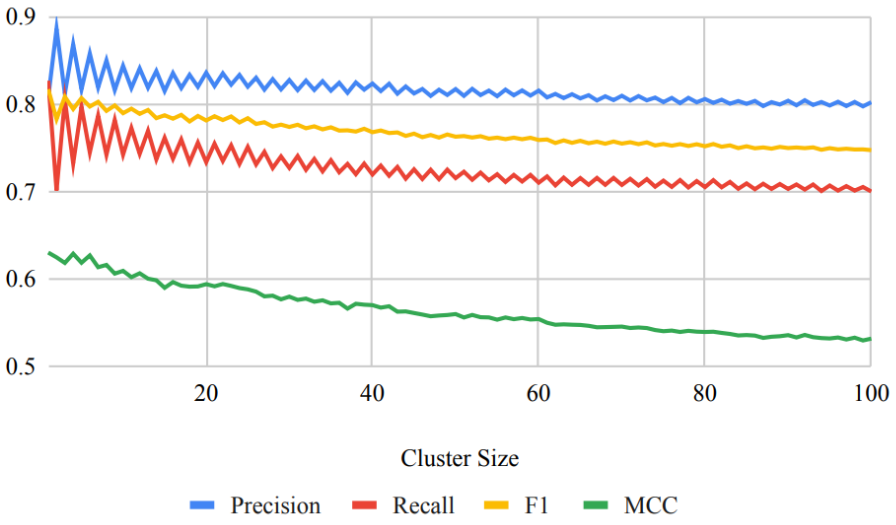


Figure 4: Results for the KNN model when searching for the most optimal cluster.

Hình 4 thể hiện kết quả cho các mô hình KNN. Đáng chú ý, **mô hình hiệu suất cao nhất là cụm nhỏ nhất có 1 láng giềng gần nhất**, đạt được **độ chính xác 81,48%, recall 0,827, F-Score 0,817, MCC là 0,63, và diện tích dưới đường cong ROC là 0,815**. Tuy nhiên, **giá trị precision cao nhất là 0,886 khi sử dụng hai láng giềng gần nhất làm dự đoán**.

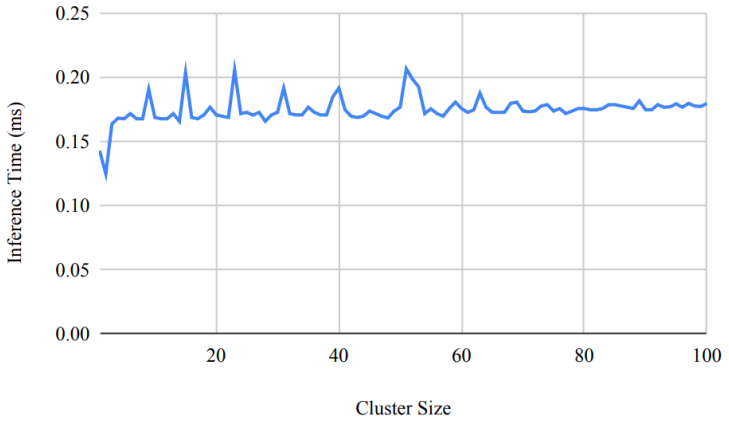


Figure 5: Observed average inference time for the KNN models to classify 1-second of audio data.

Như có thể thấy trong Hình 5, **mô hình này mất trung bình 0,143 miligiây để phân loại 1 giây dữ liệu âm thanh là thật hay giả.**

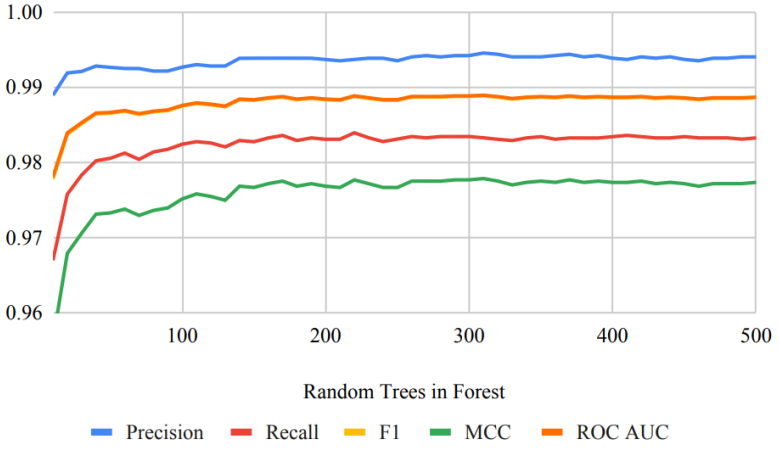


Figure 6: Results for the Random Forest model when searching for the most optimal ensemble size.

4.2.2 Random Forest:

Khác với KNN, **các kết quả của Random Forest hơi tương đối so với nhau dựa trên số cây trong rừng**. Hình 6 thể hiện kết quả phân loại cho mỗi kích thước tập hợp. Như có thể thấy, mô hình chứa **310** cây đã đạt được độ chính xác trung bình 98.89% trên 10 lần gấp dữ liệu. Tương tự, recall, precision, F-Score, MCC, và diện tích dưới đường cong ROC lần lượt là 0.995, 0.983, 0.989, và 0.989.

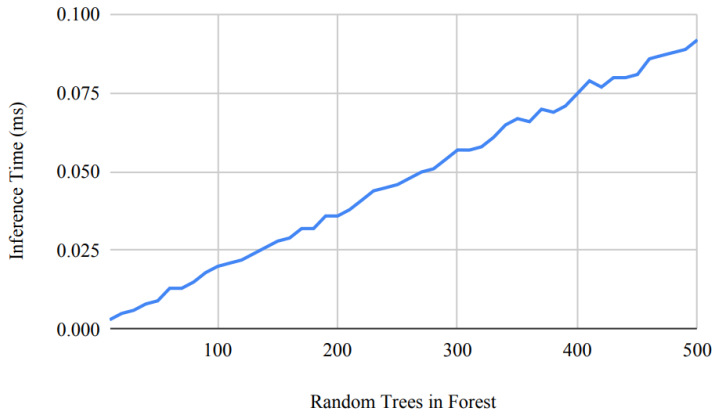


Figure 7: Observed average inference time for the Random Forest models to classify 1-second of audio data.

Hình 7 thể hiện thời gian suy luận dựa trên kích thước tập hợp. Như có thể dự đoán, thời gian suy luận tăng tương đối theo cách tuyến tính, được thể hiện trong Hình 7. Tập hợp trên đã đề cập của **310 cây ngẫu nhiên mất trung bình 0,057 miligiây để phân loại 1 giây dữ liệu âm thanh.**

4.2.3 XGBoost:

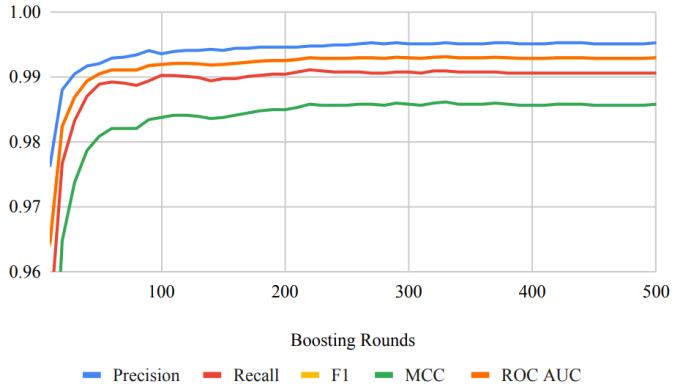


Figure 8: Results for the XGBoost model when searching for the most optimal number of boosting rounds.

Hình 8 minh họa tác động của số vòng boosting đối với hiệu suất của mô hình XGBoost trong việc phân loại dữ liệu. Một sự tăng đáng kể khoảng 3% trong độ chính xác được quan sát giữa 10 đến 50 vòng boosting, sau đó sự cải thiện ổn định. Phương pháp hiệu quả nhất tổng thể được đạt được với **330 vòng boosting**, dẫn đến một độ chính xác phân loại là **99.3%**. Mô hình này thể hiện **độ chính xác là 0.995, recall là 0.991, và F-Score là 0.993. Hệ số tương quan**

Matthews (MCC) được ghi nhận là 0.986, với diện tích dưới đường cong ROC là 0.993.

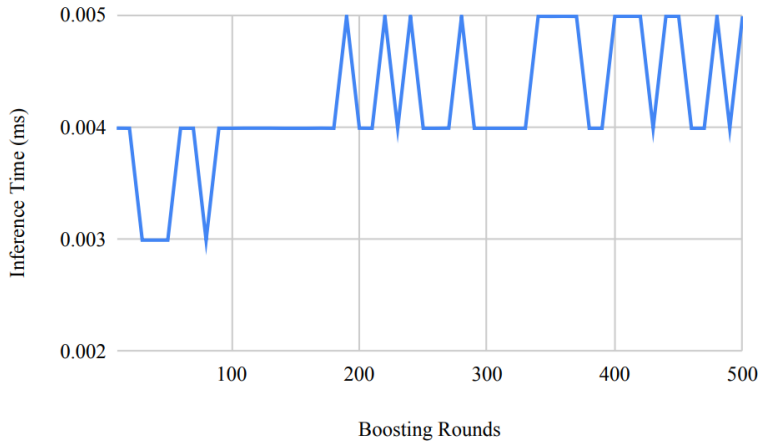


Figure 9: Observed average inference time for the XGBoost models to classify 1-second of audio data.

Hình 9 trình bày kết quả suy luận, mà trong một phạm vi 0.001 mili giây, đều đặn. Cụ thể, mô hình XGBoost sau **330 vòng mất trung bình 0.004 mili giây để dự đoán lớp của dữ liệu âm thanh 1 giây**.

4.3 Results Comparison

Phần này so sánh tất cả các kết quả cho các mô hình được huấn luyện trong nghiên cứu này. Như đã mô tả trước đó, sử dụng phương pháp kiểm tra chéo 10 lần và báo cáo các chỉ số trung bình cùng với phương sai.

Table 5: Comparison of averaged validation metrics over 10-fold cross validation for the machine learning models. Inference time denotes the average time taken for the model to predict the class of 1-second of audible speech.

Model	Mean Value over 10-fold Cross Validation						
	Acc.	*Prec.*	*Rec.*	*F1*	*MCC*	*ROC AUC*	*Inference Time (ms)*
XGBoost (330)	0.993	0.995	0.991	0.993	0.986	0.993	0.004
Random Forest (310)	0.989	0.995	0.983	0.989	0.978	0.989	0.057
Quadratic Discriminant Analysis	0.948	0.969	0.924	0.946	0.896	0.948	0.002
Linear Discriminant Analysis	0.889	0.886	0.893	0.889	0.778	0.889	0.001
Ridge	0.883	0.884	0.882	0.883	0.767	0.883	0.001
Naïve Bayes (Gaussian)	0.830	0.864	0.784	0.822	0.664	0.830	0.001
KNN (1)	0.815	0.808	0.827	0.817	0.630	0.815	0.143
SVM	0.723	0.815	0.576	0.675	0.465	0.723	0.605
Naïve Bayes (Bernoulli)	0.692	0.742	0.587	0.655	0.391	0.691	0.001
Stochastic Gradient Descent	0.668	0.732	0.760	0.681	0.407	0.673	0.001
Gaussian Process	0.614	0.997	0.229	0.372	0.358	0.614	0.561

Bảng 5 thể hiện sự so sánh tổng quan về các kết quả cho mỗi mô hình học máy. Như có thể thấy, mô hình hiệu quả nhất là mô hình Extreme Gradient Boosting, đạt được độ chính xác trung bình là 99.3% qua 10 lần gấp dữ liệu. Tuy nhiên, mô hình thứ ba hiệu quả nhất, Phân tích phân loại bậc hai, có thể đã đạt được 94.8% độ chính xác, nhưng có thể phân loại dữ liệu âm thanh trong nửa thời gian với 0.002 mili giây mỗi đối tượng. Hiệu suất của **XGBoost** và **Random Forest** so với các mô hình khác cho thấy một phương pháp tập hợp là hữu ích nhất cho vấn đề phân loại này.

Table 6: Comparison of standard deviation within the classification metrics over 10-fold cross validation for the machine learning models.

Model	Mean Value over 10-fold Cross Validation					
	Acc.	*Prec.*	*Rec.*	*F1*	*MCC*	*ROC AUC*
XGBoost (330)	0.002	0.002	0.005	0.002	0.005	0.002
Random Forest (310)	0.003	0.003	0.005	0.003	0.006	0.003
Quadratic Discriminant Analysis	0.004	0.006	0.008	0.005	0.009	0.005
Linear Discriminant Analysis	0.009	0.013	0.009	0.009	0.018	0.009
Ridge	0.008	0.011	0.008	0.008	0.016	0.008
Naïve Bayes (Gaussian)	0.007	0.013	0.014	0.011	0.015	0.008
KNN (1)	0.009	0.017	0.011	0.011	0.019	0.009
SVM	0.012	0.022	0.019	0.018	0.026	0.013
Naïve Bayes (Bernoulli)	0.012	0.014	0.021	0.018	0.023	0.012
Stochastic Gradient Descent	0.086	0.178	0.287	0.139	0.146	0.086
Gaussian Process	0.008	0.004	0.013	0.017	0.010	0.006

Phương sai của các chỉ số qua 10 lần gấp dữ liệu có thể được tìm thấy trong Bảng 6. **Hai mô hình có hiệu suất tốt nhất, XGBoost và Random Forest, cũng có độ lệch chuẩn thấp**. Các giá trị thấp này cho thấy các mô hình thực hiện tốt một cách đồng nhất trên các lần gấp khác nhau của dữ liệu trong quá trình kiểm tra chéo. Đáng chú ý là phương pháp QDA, mặc dù hoạt động kém hơn một chút, nhưng có thể được giải thích và có thể cung cấp khả năng giải thích tăng lên cùng với thời gian suy luận thấp hơn

5. Conclusion and Future Work

Nghiên cứu này đã giải quyết một số vấn đề an ninh đang gia tăng với trí tuệ nhân tạo sinh học, cụ thể là những vấn đề liên quan đến lừa đảo bằng giọng nói con người được tạo ra bằng trí tuệ nhân tạo. Với sự tăng trưởng nhanh chóng về chất lượng của các hệ thống này, quan trọng là các hệ thống phải cung cấp tính minh bạch trong thời gian thực về tính hợp pháp của một giọng nói con người trong một cuộc họp hoặc cuộc gọi điện thoại. Giọng nói được tạo ra bằng trí tuệ nhân tạo có thể được sử dụng cho các mục đích đen tối, như giả mạo trong các cuộc tấn công kỹ thuật xã hội. **Các đóng góp của công việc này bao gồm một bộ dữ liệu phân loại âm thanh ban đầu được phát hành cho công việc tương lai, phân tích toàn diện về ý nghĩa thống kê của các đặc điểm âm thanh được trích xuất từ giọng nói thật và giọng nói được tạo ra bằng trí tuệ nhân tạo, và cuối cùng là tối ưu hóa các mô hình máy học có thể dự đoán tính hợp pháp của giọng nói trong thời gian thực.**

Kết luận, khả năng đáng kinh ngạc của các mô hình XGBoost và Random Forest trong việc tổng hợp qua các lần gấp của cross-validation cho thấy rằng hoàn toàn có thể phát hiện ra giọng nói được tạo ra bằng trí tuệ nhân tạo ngay cả với những mô hình tiên tiến nhất vào thời điểm viết bài này. Các chỉ số đánh giá mô hình khác nhau cũng cho thấy tính mạnh mẽ của mô hình, và đã được quan sát thấy dữ liệu được suy luận trong khoảng thời gian từ 0,004 đến 0,057 mili giây, hiệu quả biến chúng thành bộ phân loại thời gian thực. Các phương pháp hiện đại nhất, như những phương pháp đã thảo luận trong phần nghiên cứu văn học, đòi hỏi chi phí tính toán cao hơn nhiều và do đó có thể không thể phân loại giọng nói trong thời gian thực. Điều này có nghĩa là phương pháp của chúng tôi có thể phát hiện các cuộc tấn công trong khi chúng đang diễn ra.

Trong tương lai, có thể cải thiện phương pháp bằng cách khám phá các bộ mô hình hoạt động tốt nhất. Điều này có thể cho phép tổng hợp tốt hơn thông qua việc nhận diện và sửa chữa các sai lầm của mô hình. Các phương pháp được khám phá cũng có thể được cải thiện **bằng cách bổ sung các biểu diễn đặc trưng âm thanh bổ sung, tăng số chiều đầu vào và cung cấp nhiều cơ hội hơn cho việc tạo ra quy tắc.** Cuối cùng, tập dữ liệu DEEP-VOICE có thể được mở rộng với nhiều người nói hơn trong tương lai để tăng đa dạng dữ liệu và tăng cường khả năng tổng quát hóa, cũng như sử dụng các phương pháp tạo **giọng nói neural khác nhau ngoài RVC.**

Để kết luận cuối cùng, nghiên cứu này đã đề xuất một phương pháp học máy mạnh mẽ để **phát hiện chuyển đổi giọng nói dựa trên RVC, cho phép nhận diện các cuộc tấn công kỹ thuật xã hội dựa trên giọng nói trong thời gian thực. Tập dữ liệu DEEP-VOICE được tạo ra cho nghiên cứu này được công bố công khai để thúc đẩy nghiên cứu đa ngành trong phân tích giọng nói được tạo ra bởi trí tuệ nhân tạo.** Khi lĩnh vực này tiếp tục phát triển với tốc độ nhanh chóng, các phương pháp chủ động là cần thiết để đảm bảo sự minh bạch và khuyến khích việc sử dụng đạo đức của trí tuệ nhân tạo sinh học.

6. Data Availability Statement

The datasets generated during and/or analysed during the current study are available in the DEEP-VOICE repository,

https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition