Technical Report

Kemal Şahin

Student Number: 2200765021

Course Name: AIN413

Contents

1 Abstract

This project aimed to enhance the diagnosis of medical symptoms using a hierarchical machine learning framework that integrates speech-to-text conversion with advanced text classification. Initially proposed to utilize the waw2vec model for audio processing, technical limitations necessitated a pivot to Google's Speech Recognition API, which proved more feasible for real-time audio to text transcription. The transcribed data was then classified using a pre-trained BERT model, selected for its robust performance in text analysis. This integrated approach demonstrated high accuracy in classifying a variety of medical symptoms, suggesting significant potential for improving patient care through AI-driven diagnostics. The project highlighted the adaptability required in AI implementations and underscored the importance of choosing appropriate technologies based on practical constraints and performance considerations. Recommendations for future work include exploring alternative speech-to-text models to enhance transcription accuracy and expanding the classification model to cover a broader spectrum of medical conditions.

2 Introduction

2.1 Background

The integration of Artificial Intelligence (AI) into healthcare has transformed numerous aspects of patient care and medical diagnostics. AI's capability to analyze complex datasets far exceeds human ability, making it a critical tool in diagnosing diseases, predicting outcomes, and personalizing patient treatment plans. Specifically, AI's application in processing and classifying audio data has opened new ways for advancements in remote monitoring and diagnostic systems, allowing for more accessible and immediate patient care.

One of the most promising applications of AI in healthcare is the classification of medical symptoms from audio inputs. This approach leverages AI's auditory data processing strengths to interpret and classify verbal symptom descriptions, which are often the first indicators of medical issues. By automating symptom classification, healthcare providers can quickly identify patient needs and prioritize care accordingly, potentially saving lives in urgent care scenarios.

2.2 Project Objectives

The primary objective of this project was to develop an integrated system that leverages both speech-to-text technology and text classification to automate the process of medical symptom classification from audio data. This system aimed to address specific challenges in healthcare diagnostics, particularly the need for rapid and accurate interpretation of patient-reported symptoms in audio format. The goals of the project were multi-faceted and included:

- **Implementing Robust Speech Recognition:** To utilize advanced speech recognition technology for converting spoken language into text. The project initially planned to employ the waw2vec framework, but due to technical constraints, it was necessary to pivot to Google's Speech Recognition API. This change was intended to ensure higher reliability and accessibility in real-time audio processing.
- **Enhancing Text Classification with BERT:** To classify the transcribed text into predefined medical symptom categories using the BERT (Bidirectional Encoder Representations from Transformers) model, known for its effectiveness in understanding context within text.

- **Integrating Speech-to-Text with Text Classification:** To seamlessly integrate these technologies into a single workflow that could process audio inputs and provide immediate classification outputs, thereby enhancing the diagnostic process.
- **Developing a User-Friendly Interface:** To create a Streamlit-based application that allows users to upload audio files, view the transcriptions, and receive symptom classifications in real-time. This interface was designed to be intuitive, facilitating ease of use for both healthcare providers and patients.

The successful integration of these technologies was anticipated to significantly improve the efficiency of medical diagnostic processes, particularly in telehealth contexts where accurate and quick symptom assessment is crucial. The project sought to demonstrate the practicality and benefits of AI in enhancing patient care and operational efficiencies within healthcare systems.

3 Project Proposal Summary

3.1 Methodology Overview

The methodology proposed for this project was designed to leverage the power of machine learning to develop a hierarchical model that could efficiently process and classify medical symptoms from audio data. The proposed approach was structured as follows:

1. **Data Acquisition:** The project utilized an 8.5-hour dataset of audio recordings from Appen (formerly Figure Eight), available on Kaggle. These recordings included a diverse range of medical symptoms described by various contributors, accompanied by textual annotations.
2. **Speech-to-Text Conversion:** The initial plan was to deploy the waw2vec model to convert spoken words into text. This cutting-edge model was chosen for its ability to capture the nuances of speech, crucial for accurate transcription in medical contexts.
3. **Text Classification:** Following transcription, a BERT model was to be employed for classifying the textual data into predefined medical symptom categories. BERT's deep learning capabilities make it ideal for understanding the context and nuances within the text.
4. **Integration of Technologies:** A seamless integration of the speech-to-text and text classification technologies was planned to automate the entire process from audio input to symptom categorization.

The integration of these technologies aimed to create a robust system capable of enhancing diagnostic accuracy and efficiency in healthcare settings, particularly benefiting telemedicine and remote healthcare services.

3.2 Expected Outcomes

The project anticipated several significant outcomes that would demonstrate the effectiveness of integrating AI into healthcare diagnostics:

- **High Accuracy in Symptom Classification:** The primary expected result was a significant improvement in the accuracy of symptom classification, aiming to reduce errors and misdiagnoses in clinical settings.
- **Enhanced Diagnostic Efficiency:** By automating the transcription and classification processes, the project expected to speed up the diagnostic workflow, enabling quicker patient throughput and timely interventions.
- **Scalable and Adaptable System:** The system was designed to be scalable, capable of handling larger datasets and adaptable to incorporate additional symptom categories or different languages in the future.
- **Contribution to Telehealth:** A major outcome was the expected enhancement of telehealth capabilities, providing reliable diagnostic tools that could be employed remotely, thereby making healthcare more accessible.

These outcomes were intended not only to validate the project's technical feasibility but also to showcase its practical benefits in real-world healthcare applications, thereby supporting the broader adoption of AI technologies in patient care.

4 Data Description

4.1 Source and Nature of the Data

The dataset utilized for this project was sourced from Appen, formerly known as Figure Eight. It comprises 8.5 hours of audio recordings, which were collected to represent a wide array of medical symptoms described by various individuals. Each audio recording is accompanied by a corresponding textual transcription, providing a dual-layer of data: audio and text. This dataset is publicly available on Kaggle and is widely used for training models that require natural language processing and speech recognition capabilities in the healthcare context.

4.2 Features Available

The dataset includes several key features that are critical for both training the machine learning model and evaluating its performance. These features include:

- **Audio Clipping:** Indicates if any part of the audio exhibits clipping, which may suggest distortion or overly high input volume.
- **Background Noise Audible:** Denotes the presence of background noise within the recording.
- **Overall Quality of the Audio:** A numerical score denoting the perceived overall quality of the audio.
- **Quiet Speaker:** Identifies whether the recording's speaker volume is considered low.
- **File Download:** A URL from which the corresponding audio file can be downloaded.
- **Phrase:** The transcribed text of what is spoken in the audio file.

- **Prompt:** The context or topic that the spoken phrase is meant to address.

These features are used to assess the usability of the audio recordings for accurate transcription and subsequent symptom classification.

4.3 Data Preprocessing

Data preprocessing was a crucial initial phase aimed at enhancing the quality and usability of the dataset for subsequent modeling processes. The following key preprocessing steps were undertaken:

1. **Exploratory Data Analysis (EDA):** An initial exploration was conducted to understand the dataset's characteristics, focusing on identifying and addressing audio quality issues. The EDA helped in recognizing the presence of 'bad' sounds or recordings with undesirable features such as clipping, excessive background noise, or low speaker volume.

2. **Feature Engineering and Scaling:** Critical features relevant to audio quality assessment were extracted and encoded. This included one-hot encoding categorical variables like audio clipping, background noise visibility, and speaker volume presence. The features used for clustering included:

   - Audio clipping confidence
   - Background noise audible confidence
   - Overall quality of the audio
   - Quiet speaker confidence

   These features were scaled using a `StandardScaler` to normalize their distribution, facilitating more effective clustering.

3. **Clustering for Audio Quality Assessment:** Multiple clustering techniques were tested to segment the audio files based on quality metrics. The methods included K-means, K-means with PCA, Agglomerative Clustering, DBSCAN, Spectral Clustering, and Gaussian Mixture Models. Initially, K-means with PCA showed a silhouette score of 0.349, but it was DBSCAN that eventually provided the best segmentation with a silhouette score of 0.703, indicating a strong structure found in the data.

4. **Optimization of Clustering Parameters:** DBSCAN parameters were fine-tuned using a range of min_samples and eps values calculated based on the nearest neighbors technique. The optimal parameters found were eps = 0.0796 and min_samples = 6, achieving an improved silhouette score of 0.837.

5. **Labeling and Filtering:** The optimized DBSCAN model was applied to label the data, identifying and segregating low-quality recordings which were characterized by high noise levels and poor clarity. The clustering resulted in 36 clusters with a significant number of outliers (470 noisy points), indicating a variety of audio quality within the dataset. From each cluster, 2 sample was listened but there were no significant differences between clusters as expected.

Despite the sophisticated clustering and filtering strategies, the challenge of removing all noisy data wasn't end up well, leading to the decision to proceed with the original sound files for model training.

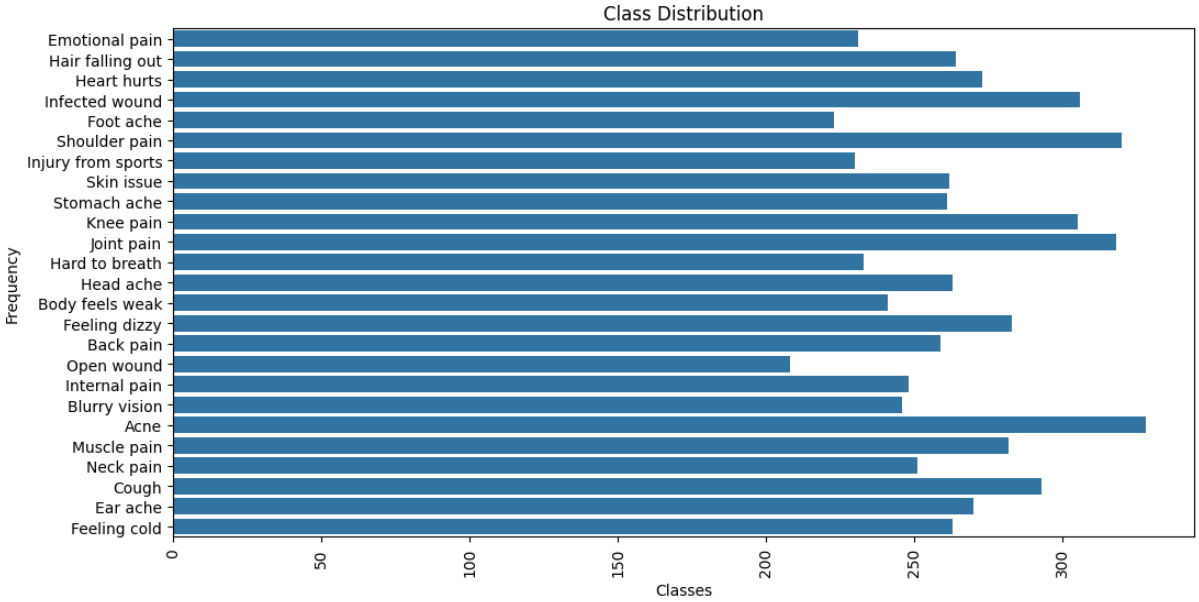

Figure 1: Distribution of Audio Quality Metrics

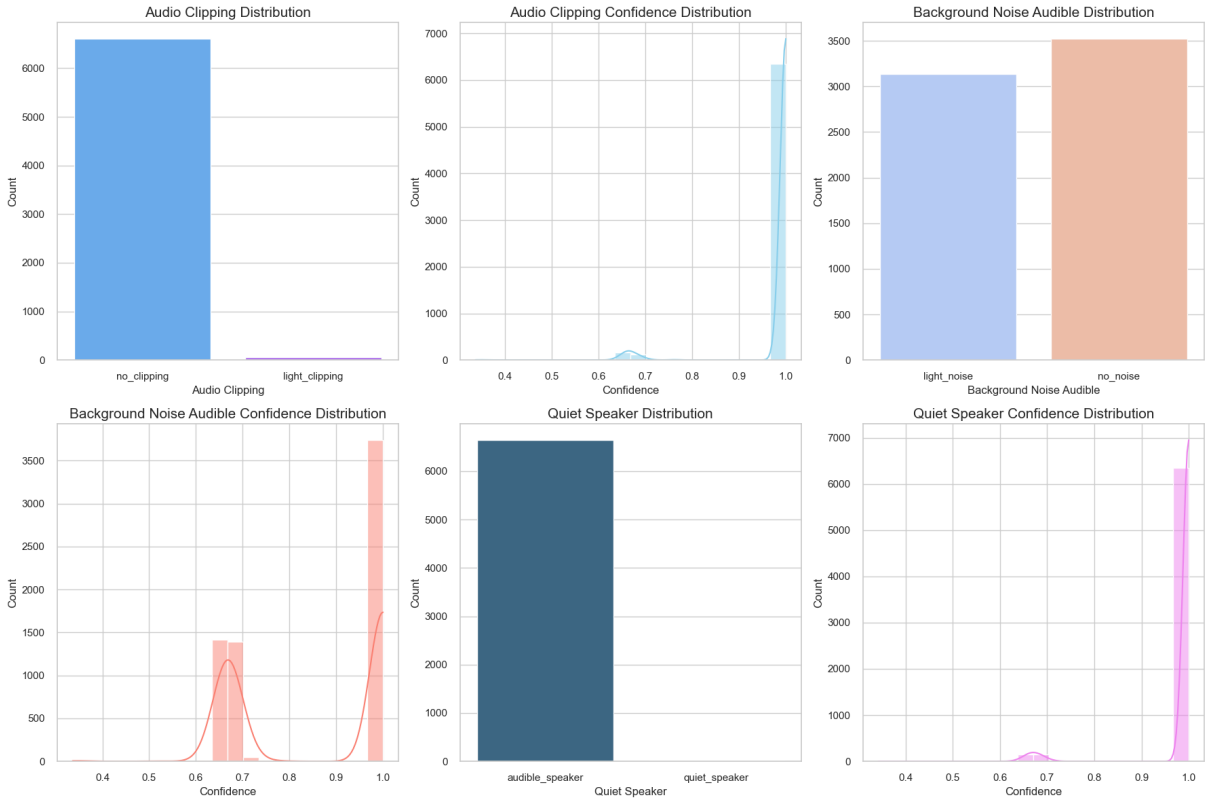

Figure 2: Class Distribution among Transcribed Phrases

5 Project Execution

5.1 Model Development and Integration

The development of the machine learning model underwent a significant shift due to practical constraints encountered during the initial phase of the project. Initially, the waw2vec model was selected for its advanced capabilities in speech-to-text conversion. However, due to unexpected technical limitations, particularly related to system resource requirements and model scalability under our dataset constraints, it became necessary to pivot to a more feasible solution.

Google's Speech Recognition API was adopted as an alternative due to its robustness, wide support, and ease of integration. This API provided reliable speech-to-text conversion that was crucial for the subsequent stages of the project. Following the successful transcription of audio data, the project utilized the BERT model, a state-of-the-art text classification framework. BERT was particularly chosen for its deep learning capabilities to understand and process natural language effectively, making it ideal for classifying medical symptoms from transcribed texts.

The integration involved:

- Configuring the Google Speech Recognition API to process audio files and convert spoken words into text.
- Employing the pre-trained BERT model to classify these texts into predefined medical symptom categories.
- Fine-tuning the BERT model to accommodate the specific nuances and terminology found in medical dialogues.

This dual approach of speech-to-text and text classification formed the core of our hierarchical machine learning model, ensuring that each component functioned synergistically to enhance overall diagnostic accuracy.

5.2 System Implementation

The implementation of the system was realized through the development of a Streamlit application, designed to allow users to interact with the model in real-time. The application serves as a user-friendly interface for uploading audio files, displaying the transcription results, and presenting the final symptom classification.

Key features of the Streamlit application include:

- **File Upload Capability:** Users can upload audio files directly through the interface, which are then automatically processed by the backend model.
- **Real-time Processing:** The application provides on-the-fly transcription and classification, showcasing the model's response almost instantaneously.
- **Display of Results:** After processing, the transcription and its corresponding classification are displayed to the user, allowing for immediate review and analysis.
- **Interactive User Experience:** Designed with the end-user in mind, the application features a minimalistic and intuitive layout that simplifies interactions without compromising on functionality.

This system not only demonstrates the practical application of the project's machine learning model but also highlights the potential for real-world deployment in clinical settings, where quick and accurate symptom assessment is crucial.

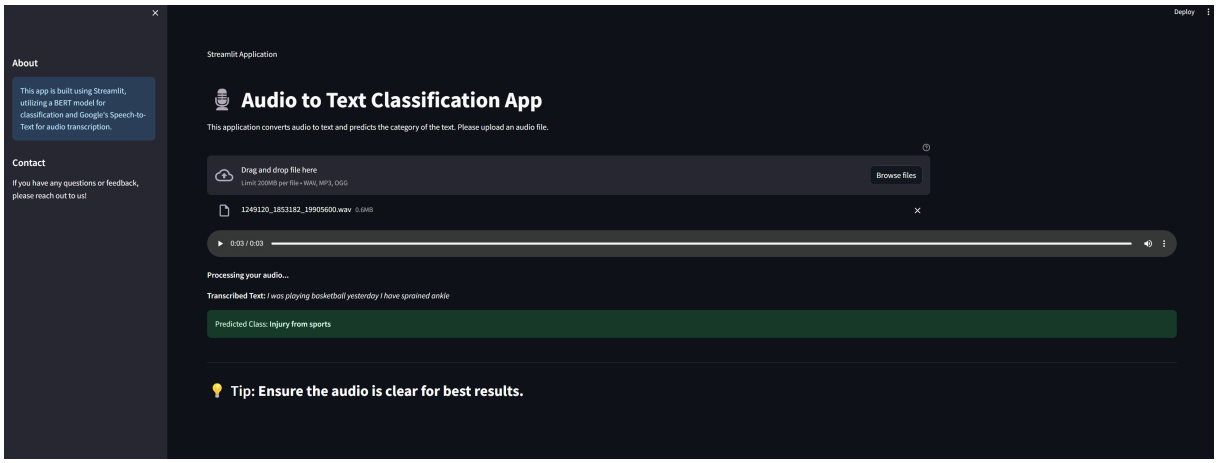

Figure 3: Screenshot of the Streamlit Application Interface

6 Results and Evaluation

6.1 Model Performance

The project utilized two main models: Google's Speech Recognition for audio-to-text conversion and BERT for text classification. Google's Speech Recognition provided a robust foundation for accurate transcription despite the lack of fine-tuning capability, which was critical given the shift from waw2vec due to memory constraints.

BERT's performance was exceptional, demonstrating its capability in handling natural language processing tasks within the medical domain:

- Epoch 1: Avg Loss: 2.0503, Test Accuracy: 94.49%
- Epoch 2: Avg Loss: 0.3913, Test Accuracy: 99.18%
- Epoch 3: Avg Loss: 0.1177, Test Accuracy: 99.33%

These results highlight BERT's rapid learning curve and its effectiveness in accurately classifying medical symptoms from text, confirming the model's suitability for deployment in real-world healthcare applications.

7 Comparison: Proposal vs. Execution

The project experienced several deviations from its initial proposal, primarily due to technical challenges and practical constraints encountered during execution. Originally, the project was designed to utilize the waw2vec model for speech-to-text conversion. However, due to memory limitations that led to system crashes, it was necessary to pivot to Google's Speech Recognition API. This change, although unforeseen, allowed the project to continue with a more stable and scalable solution.

Additionally, the initial proposal underestimated the complexity of accurately clustering audio data based on quality. The exploratory data analysis revealed significant challenges in distinguishing between good and poor-quality audio recordings. Despite the various algorithms and hours, no significant clusters were found to remove.

These changes had a profound impact on the project's direction and outcomes, demonstrating the need for flexibility in research and development projects and underscoring the importance of contingency planning in project management.

8 Conclusions and Future Work

8.1 Conclusions

The project successfully developed a hierarchical machine learning model capable of classifying medical symptoms from audio data with high accuracy. Key achievements of the project include:

- Successful integration of Google's Speech Recognition and BERT model for robust symptom classification.
- High accuracy rates in classification tasks, demonstrating the potential of AI in enhancing diagnostic processes in healthcare.
- Development of a user-friendly Streamlit application that allows for real-time audio processing and classification, enhancing user engagement and accessibility.

Lessons learned from the project include the importance of adaptable project planning and the challenges of working with audio data, particularly in terms of data quality and preprocessing.

8.2 Future Directions

Future research and development can extend this project in several promising directions:

- Exploring alternative speech-to-text models that might offer better accuracy or efficiency, especially those that can be fine-tuned for specific medical jargon.
- Enhancing the model's ability to handle a wider variety of audio qualities and accents to improve its applicability in diverse real-world environments.
- Expanding the system to include more symptom categories and possibly other languages, increasing the system's usability across different geographic and demographic contexts.
- Integrating the system with existing healthcare IT infrastructure to provide a seamless diagnostic tool for medical professionals.

Continued innovation in this area promises to significantly enhance the capabilities of AI in healthcare diagnostics, contributing to better patient outcomes and more efficient healthcare delivery.