IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Voice Based Services Using Audio Classification

[1]Jay Prakash Malviya , [2]Dr. Shamshekhar S. Patil

[1]Post Graduate Student, [2]Associate Professor

[1,2]Department of Computer Science and Engineering,

[1,2]Dr.Ambedkar Institute of Technology,Bangalore,India

Abstract: The perfect, accurate and robust detection of the audio has been widely grown as the speech technology in the area of audio researches, audio forensics, speech recognition, and so on. However, in real time, it is a challenge to deal with the massive data arriving from distributed sources. Thus, the study introduces a method that effectively deals with the data from the different sources using the audio classification method. This method uses anywhere for our general purpose projects, Voice services are rapidly gaining popularity, with the use of virtual assistants such as Siri, Alexa, and Google Assistant becoming increasingly common. These systems rely on audio classification to understand and respond to user requests. Audio classification is the process of using machine learning algorithms to classify audio data into different categories or classes. This is a crucial component of voice services, as it allows the system to understand and interpret the user's spoken words.

Index terms: **Audio classification, Voice-based services, Voice recognition, Speaker identification, Natural language processing, Speech recognition, Language translation, Machine learning, Artificial intelligence, Human-computer interaction**

I. INTRODUCTION

Voice services are rapidly gaining popularity, with the use of virtual assistants such as Siri, Alexa, and Google Assistant becoming increasingly common. These systems rely on audio classification to understand and respond to user requests. Audio classification is the process of using machine learning algorithms to classify audio data into different categories or classes. This is a crucial component of voice services, as it allows the system to understand and interpret the user's spoken words.

To perform audio classification, the raw audio data is first preprocessed and transformed into a suitable representation for the machine learning algorithm. This might involve extracting features such as Mel-frequency cepstral coefficients (MFCCs) or spectrograms from the audio signal.

The extracted features are then fed into a machine learning model, which is trained on labeled data to learn how to classify audio data into different categories. The performance of the model is evaluated on a held-out test set to determine how well it can generalize to new, unseen data.

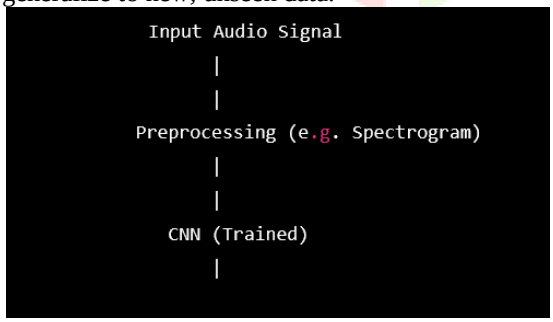

Fig – Working

There are several different approaches to training machine learning models for audio classification, including supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning involves training the model on a labeled dataset, where the correct classification for each input is provided. Unsupervised learning, on the other hand, involves training the model on an unlabeled dataset, where the correct classification is not provided. Semi-supervised learning falls somewhere in between, using a combination of labeled and unlabeled data for training.

The choice of machine learning algorithm for audio classification depends on the specific task at hand. For example, deep learning approaches such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been shown to perform well on tasks such as speech recognition and language modeling. Other algorithms, such as support vector machines (SVMs) and random forests have also been used for audio classification tasks.

The use of audio classification in voice services allows for more natural and intuitive human-computer interactions, making it a valuable tool for a wide range of applications. By using machine learning algorithms to classify and understand spoken words,

voice services are able to provide users with quick and accurate responses to their requests. As the technology continues to evolve, it is likely that the use of audio classification in voice services will become even more widespread and sophisticated.

II. LITERATURE SURVEY

Some computer software applications are already using voice commands and instructions, which are becoming increasingly important in providing informative services and enhancing user experience. The use of voice-based human-machine dialogs allows for the successful and efficient realization of the goals of telecommunication services. The emphasis is on enabling communication at any time and any place.

The "anytime-anywhere" principle can often only be achieved using mobile and portable devices, which have limited keyboard and screen space. In these cases, a voice-based interface has advantages over traditional keyboard and screen interfaces. The paper presents a model for a multimodal interface, with a focus on recognizing voice commands for use in Lithuanian medical and social security enterprises.[1]

The paper presents the design and implementation of a voice-based mobile prescription application (vbmopa) that aims to improve healthcare services accessed via mobile phone. The system could lead to cost and time savings in healthcare centers, particularly in developing countries where treatment processes are often cumbersome and paper-based. The system uses voice XML, VUI, and UML.[2]

The paper presents a deep learning-based classification model that automates the quality assessment process of patient-doctor voice-based conversations in a telehealth service. The data consists of audio recordings from Altibbi, a digital health platform that provides telemedicine and telehealth services in the Middle East and North Africa (MENA). The goal of the model is to assist Altibbi's operations team in evaluating the quality of consultations in an automated manner. The model uses three sets of features: signal-level features, transcript-level features, and features that combine both signal and transcript levels. At the signal level, various statistical and spectral information is calculated to characterize the spectral envelope of the speech recordings.[3]

The paper describes a deep learning-based classification model that can automate the process of assessing the quality of voice-based conversations between patients and doctors in a telehealth service. The model uses audio recordings from Altibbi, a digital health platform that provides telemedicine and telehealth services in the Middle East and North Africa (MENA). The goal is to help Altibbi's operations team evaluate the quality of consultations automatically. The model uses three sets of features: features extracted from the audio signal, features extracted from the transcript, and features that combine both the signal and the transcript. To extract features from the audio signal, the model calculates various statistical and spectral properties of the speech recordings to characterize their spectral envelope.[4]

The paper describes the development of an interface for controlling a lightweight robot using cloud-based speech recognition. The main contribution of the work is the design and implementation of a software interface that can recognize voice commands, process them using cloud-based speech recognition technology, and convert them into machine-readable code. The paper also identifies requirements for evaluating different cloud services for controlling robots.[5]

The paper proposes an attention framework for extending a recent few-shot learning method that uses graph neural networks in audio classification. The goal of the proposed framework is to introduce a flexible way to selectively focus on support examples for each query process. The paper also presents an empirical study on confidence measures for few-shot learning by combining posterior probability with the normalized entropy of the network's probability output.[6]

The paper presents a study on classifying recordings of young children reading isolated words aloud in a classroom setting. The goal is to detect which recordings are noisy or speechless. The study explores two different neural network architectures: recurrent neural networks and fully connected neural networks. To train the classifiers, the authors introduce a transfer learning-based feature extraction approach that uses a pre-trained model called VGGish as a feature extractor. Due to the specific nature of the task, the different possible misclassifications have different consequences. Therefore, the authors propose an alternative metric to the F1 score that takes these consequences into account. The results show that networks trained on transfer learning-based features outperform networks trained on Mel Frequency Cepstral Coefficients, a feature representation commonly used in speech recognition tasks. The best-performing models showed a relative improvement of 25% over the use of MFCCs according to the proposed metric.[7]

This paper presents a video retrieval tool for the 2020 Video Browser Showdown (VBS), which aims to enhance the user's video browsing experience by making full use of a pre-constructed video analysis database. The tool employs deep learning-based techniques for object detection, scene text detection, scene color detection, audio classification, and relation detection, which are used to generate a scene graph. The data includes visual, textual, and auditory information, allowing users to search for videos based on a wide range of criteria beyond just visual information. The tool also provides a simple and user-friendly interface that allows novice users to quickly learn how to use it.[8]

This paper introduces a method using deep learning and convolutional neural networks (CNNs) to classify TV programs into one of five categories: advertisements, cartoons, news, songs, or sports, based on the analysis of the audio content. The goal of the work is to develop a CNN architecture that can accurately classify audio segments from TV broadcasts. To create the required dataset, the authors used a TV tuner card to capture audio from different channels, and also downloaded audio from YouTube channels. The proposed CNN model achieved an accuracy of 95% on the task of TV broadcast audio classification.[9]

This paper presents an approach to using a convolutional neural network (CNN) model to classify the sentiment of sentences in Myanmar text. The CNN model is built on top of a word embedding model (e.g., Word2Vec), which converts words into numerical vectors. The model can classify sentences and label them with sentiments such as positive, negative, neutral, unrelated, or unreadable. The model was trained on 1,152 sentences taken from customer reviews of products provided by a telecommunication company. When tested on 495 unseen sentences, the model achieved an accuracy of 86.26% and an average f-measure of 82.58% in sentiment prediction. The model was compared with traditional machine learning classifiers (e.g., support vector machines, Naive Bayes, and logistic regression), and outperformed these classifiers, with SVM achieving 64.44% accuracy, NB achieving 60.20% accuracy, and LR achieving 55.15% accuracy[10]

III. MATERIALS AND METHODS

A. DATASET

The Speech Commands dataset is a collection of audio recordings for use in training machine learning models for audio classification. It was created by researchers at Google and contains 65,000 one-second-long audio recordings of 30 different English words. The words include common commands, such as "yes," "no," and "stop," as well as digits from "zero" to "nine." The recordings in the dataset were collected from a diverse set of 2,618 different people, who spoke the words in a variety of accents and environments. This diversity makes the dataset more challenging to work with, but also more realistic and representative of real-world audio signals. The dataset is split into a training set, a validation set, and a test set, with 55,000, 5,000, and 5,000 recordings, respectively. The splits are disjoint, meaning that the recordings in each split are different and do not overlap.

This allows researchers to use the training set to train a machine learning model, the validation set to evaluate the model and tune its hyper parameters, and the test set to evaluate the final performance of the model. The Speech Commands dataset is widely used in research and development of audio classification models. It has been used to train and evaluate various machine learning algorithms, including convolutional neural networks, recurrent neural networks, and transfer learning models. The Speech Commands dataset is a valuable resource for researchers and developers working on audio classification tasks. It provides a large and diverse collection of audio recordings, which can be used to train and evaluate machine learning models for a variety of applications.

Keyword	No. of wave files
Yes	1000
No	1000
Up	1000
Down	1000
Left	1000
Right	1000
Stop	1000
Go	1000

These keywords from the speech command datasets used here to train, validate and test in ratio of 80:10:10 to the model

B. METHODOLOGY

1)Sequential

Sequential data refers to data that has a natural ordering or sequence, such as time series data or text data. This type of data is commonly found in applications such as speech recognition and natural language processing, where the order of the data is important for making predictions.

2)Resizing

Resizing (down sampling) is a common operation applied to sequential data, especially in the context of machine learning. It involves changing the length of the sequence by adding or removing elements. This can be useful for ensuring that all the sequences in a dataset have the same length, which is often required by machine learning models. There are various methods for resizing sequential data, such as padding or truncating the sequences to a fixed length, or interpolating or decimating the sequences to a different sampling rate.

3)Normalization

Normalization is another common operation applied to sequential data, especially in the context of machine learning. It involves scaling the data to have a mean of zero and a standard deviation of one, or some other specified values. This can be useful for ensuring that all the features in a dataset have the same scale, which can improve the performance of some machine learning models. There are various methods for normalizing sequential data, such as min-max scaling, z-score normalization, or robust scaling.

Sequential data often requires preprocessing operations such as resizing and normalization to make it suitable for use with machine learning models. These operations can help ensure that the data has the correct format and scale for making predictions.

Implementing voice-based services using audio classification with CNNs and spectrograms are as follows:

1. The service receives an audio signal as input, such as a recording of a person speaking or making a sound.
2. The audio signal is preprocessed to extract features that can be used as input to the CNN. This can involve converting the signal into a spectrogram, which represents the frequency components of the signal over time.
3. The extracted features are fed into a trained CNN, which makes a prediction about the class of the input audio. The CNN can be trained using a dataset of preprocessed audio recordings, with corresponding labels for each class.
4. The predicted label is used to determine the appropriate response or action for the input audio. This can involve mapping the predicted label to a specific response or action, such as providing information, executing a command, or generating an output signal.
5. The response or action is output by the service, allowing the user to interact with the service using voice commands or sounds.

C. DATA COLLECTION AND PRE PROCESSING

1. Collect a dataset of audio recordings that represent the different classes of sounds or words that the voice-based service should be able to recognize. For example, the dataset could include recordings of different words spoken in different languages, or different sounds produced by different instruments.
2. Preprocess the audio recordings to extract features that can be used as input to a CNN. This can involve converting the audio signal into a spectrogram, which represents the frequency components of the signal over time.

3.Train a CNN on the dataset of preprocessed audio recordings, using the extracted features as input and the corresponding labels as output. This can involve defining the CNN architecture, choosing an optimization algorithm, and adjusting hyper parameters to improve performance.

4. Evaluate the trained CNN on a separate test set to measure its accuracy and other metrics, such as precision, recall, and F1 score. This can involve making predictions on the test set, comparing the predicted labels to the true labels, and calculating evaluation metrics.

5.Use the trained CNN to classify new audio recordings that are encountered by the voice-based service. This can involve applying the same preprocessing steps to the new recordings, feeding the resulting features into the trained CNN, and using the predicted labels to determine the appropriate response or action.

IV RESULTS AND DISCUSSION

```
Training set size 6400
Validation set size 800
Test set size 800
```

Figure 2 - Train, Validate and Test samples

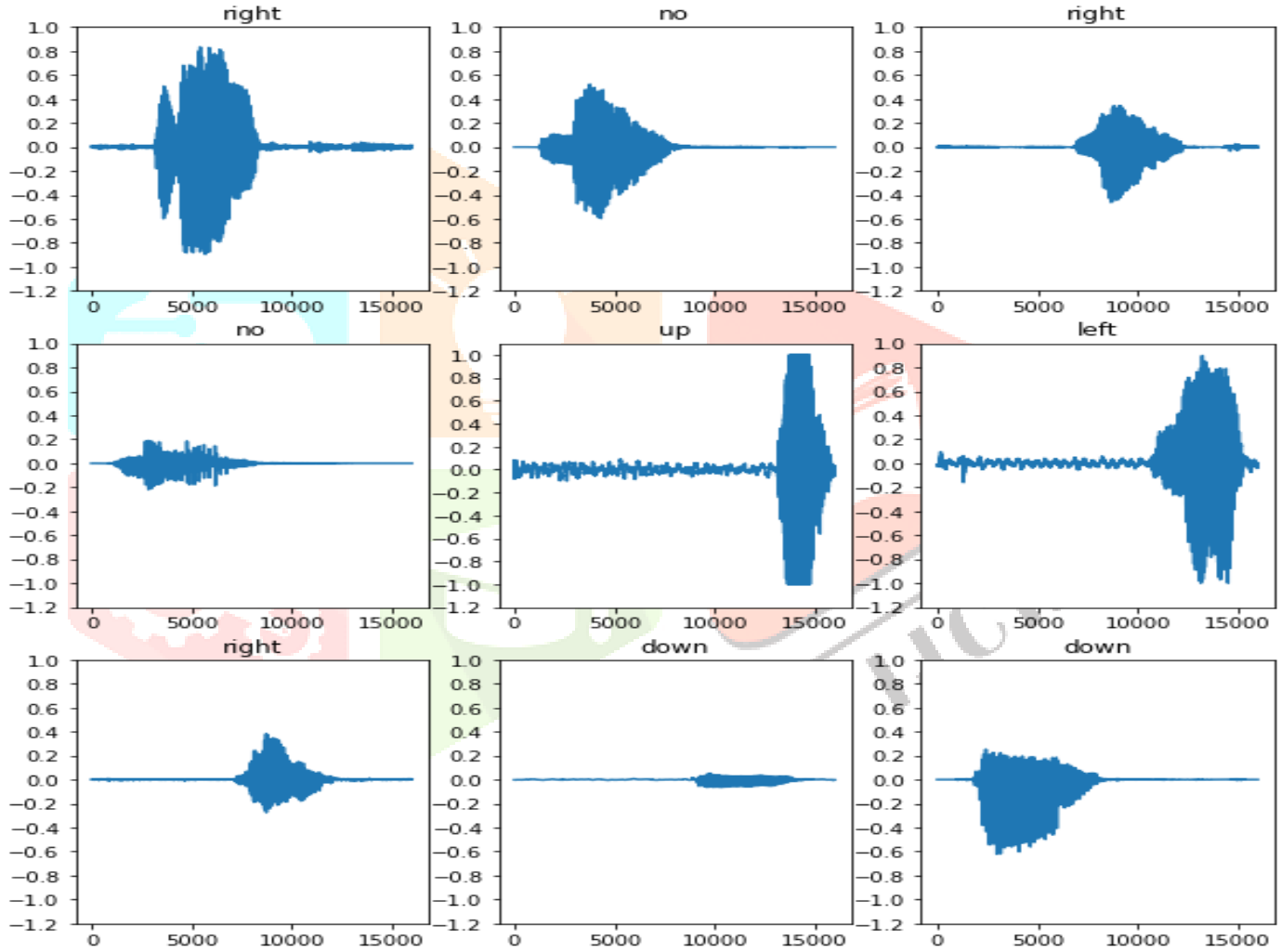

Figure 3 - Plot a few audio waveforms

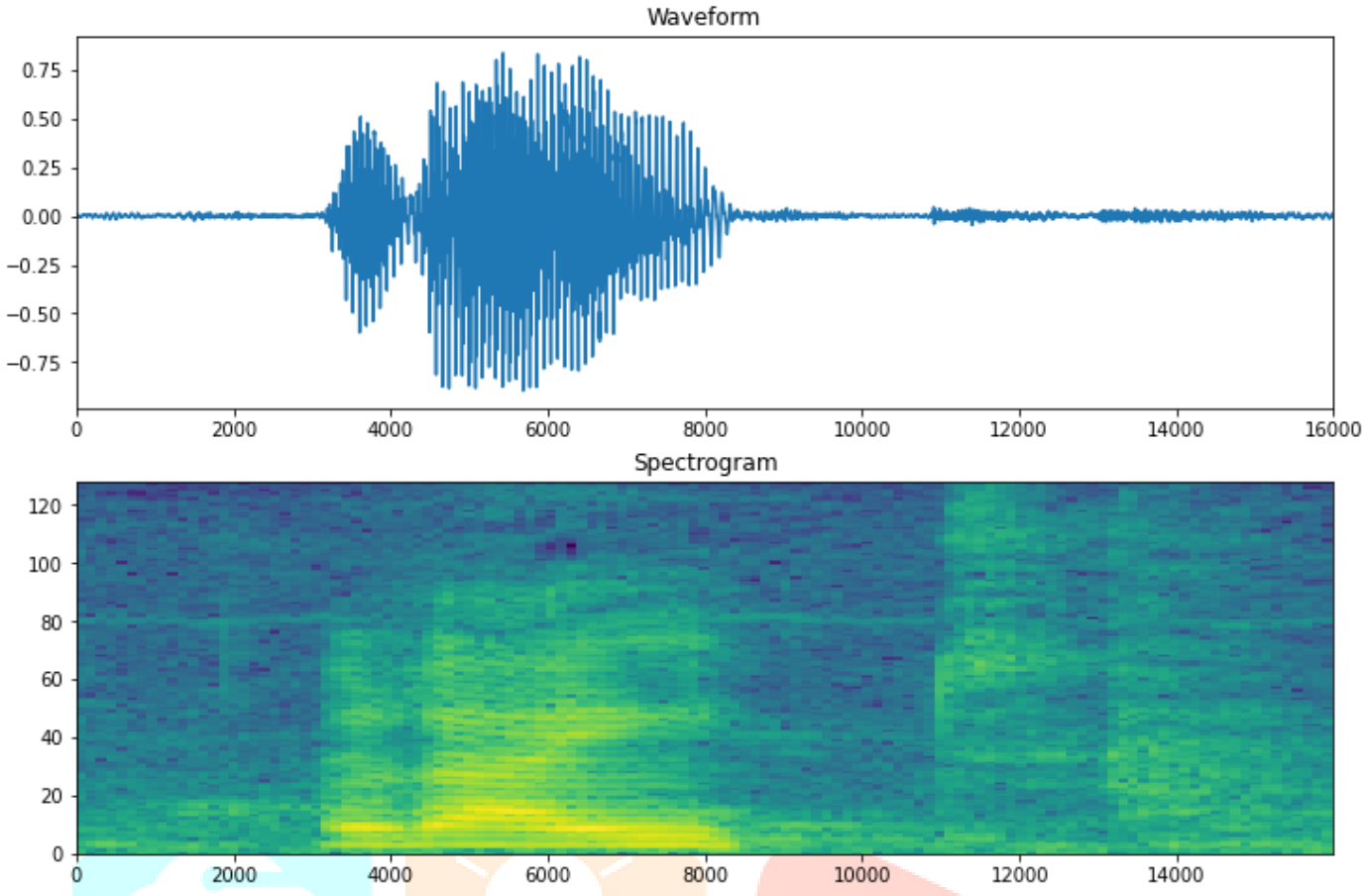

Figure 4- Waveform over time and the corresponding spectrogram (frequencies over time)

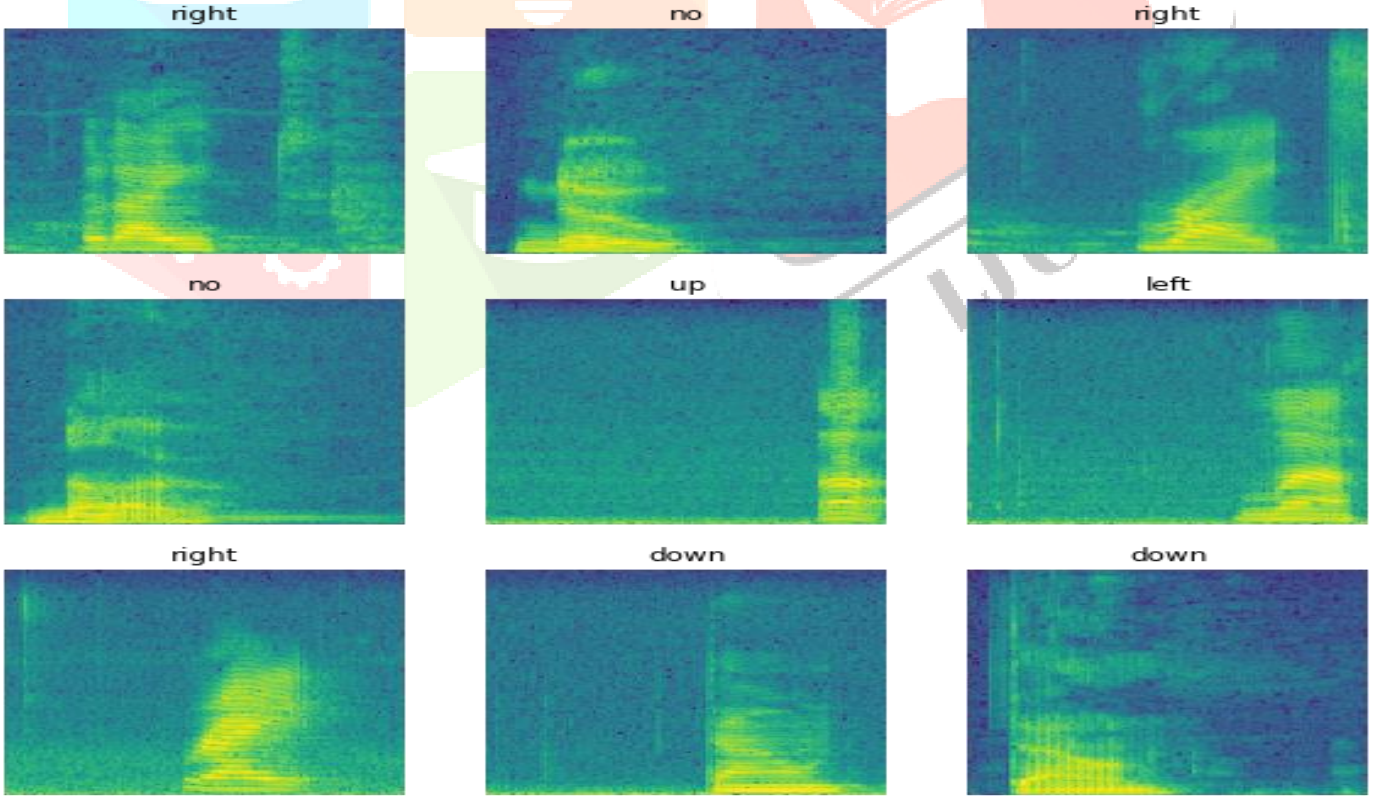

Figure 5 - The spectrograms for different examples of the dataset

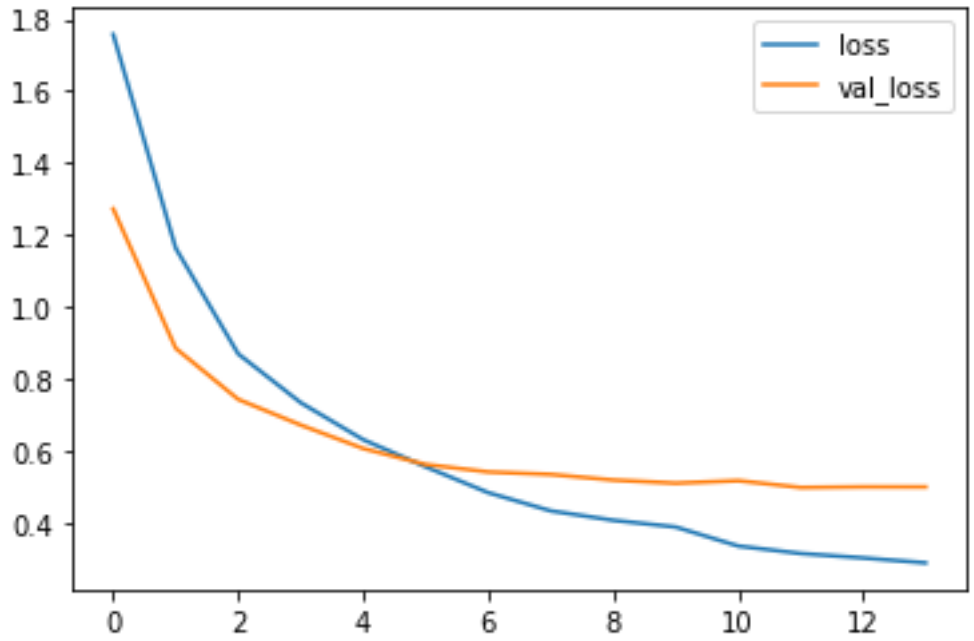

Figure 6- Plot the training and validation loss curves to check how your model has improved during training

```
Model Performance
Test set accuracy: 84%
```

Figure 7- Evaluate the model performance

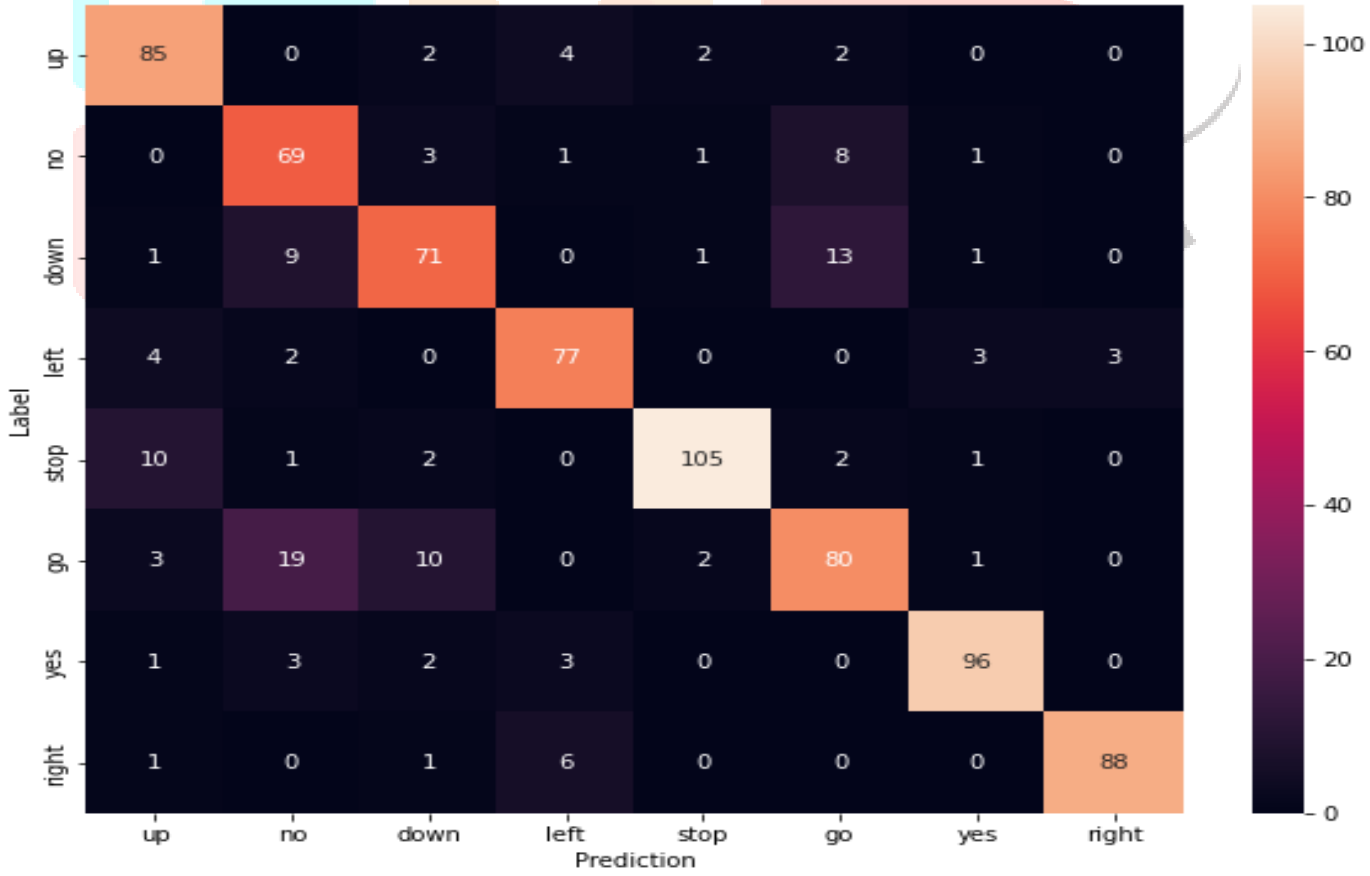

Figure 8- Display a confusion matrix

Results

Figure 9- Run inference on an audio file

The novelty of voice-based services using audio classification is that they allow for the identification and interpretation of sounds and speech in order to perform a wide range of tasks, such as voice recognition, speaker identification, and language translation. This technology has the potential to improve the way we interact with devices and machines by allowing us to use our voices as a natural and intuitive means of input. Additionally, because audio classification algorithms can be trained on large datasets, they can be highly accurate and can be used in a variety of settings, including in consumer devices, business applications, and even in medical and scientific research. Overall, the use of audio classification in voice-based services represents a significant advancement in the field of natural language processing and has the potential to greatly enhance our ability to communicate and interact with technology.

References

[1] Voice-based Human-Machine Interaction Modeling for Automated Information Services "Maskeliūnas, R., Ratkevicius, K., & Rudzionis, V. (2011). Voice-based Human-Machine Interaction Modeling for Automated Information Services. Elektronika Ir Elektrotechnika, 110, 109-112."

[2] A Voice-based Mobile Prescription Application for Healthcare Services (VBMOPA) "Ikhu-Omoregbe (2010). A Voice-based Mobile Prescription Application for Healthcare Services (VBMOPA)."

[3] Toward an Automatic Quality Assessment of Voice-Based Telemedicine Consultations: A Deep Learning Approac h“Habib, M., Faris, M., Qaddoura, R., Alomari, M., Alomari, A., & Faris, H. (2021). Toward an Automatic Quality Assessment of Voice-Based Telemedicine Consultations: A Deep Learning Approach. Sensors (Basel, Switzerland), 21.”

[4] Conversational AI - A Retrieval Based Chatbot "Surendran, A., Murali, R., & Babu, R. (2020). Conversational AI - A Retrieval Based Chatbot."

[5] Human-robot-interaction using cloud-based speech recognition systems "Deuerlein, C., Langer, M., Sessner, J., Heß, P., & Franke, J. (2021). Human-robot-interaction using cloud-based speech recognition systems. Procedia CIRP, 97, 130-135."

[6] Few-Shot Audio Classification with Attentional Graph Neural Networks “Zhang, S., Qin, Y., Sun, K., & Lin, Y. (2019). Few-Shot Audio Classification with Attentional Graph Neural Networks. INTERSPEECH.”

[7] Transfer Learning based Audio Classification for a noisy and speechless recordings detection task, in a classroom context “Hajji, M.E., Daniel, M., & Gelin, L. (2019). Transfer Learning based Audio Classification for a noisy and speechless recordings detection task, in a classroom context. SLaTE.”

[8] Deep Learning-Based Video Retrieval Using Object Relationships and Associated Audio Classes “Kim, B., Shim, J., Park, M., & Ro, Y.M. (2020). Deep Learning-Based Video Retrieval Using Object Relationships and Associated Audio Classes. MMM.”

[9] A Deep Learning CNN Model for TV Broadcast Audio Classification “Kamatchy, B., & Dhanalakshmi, P. (2020). A Deep Learning CNN Model for TV Broadcast Audio Classification. International journal of engineering research and technology, 9.”

[10] Sentence Sentiment Classification Using Convolutional Neural Network in Myanmar Texts Oo, S.H., Theeramunkong, T., & Hung, N.D. (2020). Sentence Sentiment Classification Using Convolutional Neural Network in Myanmar Texts. Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing.