English Accent Identification from speech for six regions in the UK using a hybrid of deep learning techniques.

by

Ijezie Daniel Ekenedilichukwu

Abstract

Accents are feature rich tools that are used to differentiate people who speak the same language but hail from different regions, and so, a system that can classify accents would have so many applications. Deep learning and transfer learning models like the VGGish and the YamNet have proven to perform well on a variety of audio classification tasks, however, there research on the performance of these models in accent classification tasks, as well as determining the which is most effective, is a gap that has not yet been filled .In this paper, we compare the performance of 9 models in the task of classifying 6 accents in the UK including the Irish, southern, northern, midlands, Scottish, and Welsh accents. The models evaluated include a Convolutional Neural Network (CNN), a Long short-term memory Network (LSTM), and a Multilayer Perceptron (MLP) all trained on Mel Frequency Cepstral Coefficient (MFCC) features. Furthermore, we explored the performance of CNN, an MLP and an LSTM on the task using audio features extracted from the yamnet and the vggish audio feature extraction transfer learning models. After the evaluation, we were able to determine that the best models for the UK accent detection tasks were the CNN trained on MFCC features and an LSTM model trained on features extracted from the VGGish model (VGGish-LSTM), with accuracies of 95.10% and 86.65% respectively, F1-scores 94.18% and 86.38% respectively recalls of 96.78 and 87.70%, and precisions 92.05% and 85.70% respectively.

Table Of Contents

List Of Figures

List Of Tables

List Of Equation

1. INTRODUCTION

1.1. BACKGROUND

According to Parikh et al. (2020), accents are powerful tools in identifying a speaker's origin or ethnicity especially among people who speak the same language. This implies that a system that can classify accents would be applicable in several domains.

According to Walfisz (2023), researchers from the university of Essex discovered that accents in the UK are disappearing and Hassan (2023) claimed that this is due to factors like the shift to societal evolution, increased mobility, and universal education. At this rate, certain accents would be gone in the future, and this emphasizes the need for accent identification technologies especially for individuals interested in recognizing or learning such accents in the future.

Parikh et al. (2020) claims that accent identification systems are applicable in forensic investigations, because by identifying the ethnicity of criminals through their accents, the list of suspects for crimes can be narrowed down easily.

1.2. LITERATURE REVIEW

Several research have been done to classify accents using deep learning, for example, Kiran (2021) trained a simple neural network (NN) on normalized MFCC features to classify 6 English accents from the Speaker Accent Recognition (2020) dataset. There NN model obtained an accuracy of 82.68%, a precision of 83% and a recall of 82.88% after testing on the dataset.

Cetin (2022) claimed that the most appropriate way to classify accents is to convert waveforms to spectrograms and train a CNN their image representations rather than on the spectrogram features directly. They compared the performance of a CNN and AlexNet models using weights from ImageNet in classifying accents in the "Crowdsourced high-quality UK and Ireland English Dialect speech[1]" dataset. There CNN and the AlexNet models obtained test accuracies of 92.92% and 93.39% respectively and F1-scores of 92.67% and 93.19%.

Using the speech accent archive dataset[2], Parikh et al. (2020) built a hybrid model to classify English accents. They first used an LSTM to extract features from audio signals and then trained a Deep Neural Network (DNN) on long term features and a Recurrent Neural Network (RNN) on short term features, then they extracted MFCC features from the same audio signals to train a CNN model. After this, they fused the models which formed a DNN-RNN-CNN hybrid model trained on both LSTM extracted features and MFCC features. There hybrid model had a test accuracy of 68.67%, a recall of 65% and an F1-score of 67%.

Alam, Bhuiyan and Monir (2023) proposed a DNN trained on multiple extracted features which included Zero Crossing Rate (ZRC), MFCC, Root Mean Square (RMS) and Mel-spectrograms to classify Bangla accents using the gathered Bangla accent dataset by Islam et al. (2020). There model had a test accuracy of 94%, and they compared this accuracy with other neural networks which included an LSTM, stacked LSTM and Deep Convolutional Neural Network (DCNN) and they had had accuracies of 67%, 71% and 85% respectively.

[1] https://openslr.org/83/

[2] https://www.kaggle.com/datasets/rtatman/speech-accent-archive

1.3. RESEARCH AIMS

The aim of this research was to determine if transfer learning models like VGGish[3] and YAMNet[4] would have higher accuracies in comparison to the regular deep learning models and to propose a model that performs better than other models in comparison to the wider literature.

2. METHODOLOGY

Several steps were taken to complete this research which can be seen in Figure 1,but firstly, we discuss the datasets used for the research.

2.1. DATASETS

According to Ashiq (2022), there are approximately 40 different accents in the UK. However, to keep things simple, we grouped the accents based on the constituent countries of the UK. Also, considering that England is the largest country in the UK, we proceeded to further group its accents into 3 regions namely, Northern England, Southern England, and Midlands. This allowed us to classify accents across the UK while still maintaining simplicity and regional diversity of accents in the UK.

The primary dataset used in for this research is the "Crowdsourced high-quality UK and Ireland English Dialect speech" dataset[1] proposed by Demirsahin et al. (2020).

This dataset consists of 17,877 crowd sourced audio recordings from the University of Cardiff, representing 6 accents in the UK including Northern, Southern, Scottish, Welsh, and northern Irish accents. See Table 1 for more details on the distribution of the dataset.

Table 1. Table showing the number of male, female and total audio samples in the Crowdsourced high-quality UK and Ireland English Dialect speech dataset. This table shows that the dataset is not balanced but that was not an issue because it was handled by assigning class weights. The fact that there are no female samples for the Ireland accent was not much of an issue because the genders were merged, and the models learned the Irish accent from the male samples.

Region	**Male Samples**	**Female Samples**	**Total samples**
Ireland	450	0	450
Midland	450	246	696
Northern	2,097	750	2,847
Scotland	1,649	894	2,543
Southern	4,331	4,161	8,492
Wales	1,650	1,199	2,849
Total	10,627	7,250	17,877

In addition to the primary data set we used the LibriTTS-British-Accents[5] dataset to validate and test the ability of our model to generalize. The dataset contains recordings of only British speakers from the LibriTTS. The dataset has 4 classes which include English, Irish, Scottish, and Welsh but we removed the English class for the sake of this research because we observed that it was a combination of all British accents.

After removing the English class, we were left with 581 samples. See Table 2 for how the samples are distributed among the classes.

[3] https://github.com/tensorflow/models/tree/master/research/audioset/vggish
[4] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet
[5] https://www.kaggle.com/datasets/oscarvl/libritts-british-accents

Table 2. Table showing number of samples for the Irish, Scottish and Welsh accents classes in the LibriTTS-British-Accents[5] after removing the English class. The table shows that the total number of samples is very small, but this was not a problem because we only used the dataset for testing and exploring transfer learning and not training a model from scratch.

Region	**Total Samples**
Irish	215
Scotland	204
Welsh	162
Total	581

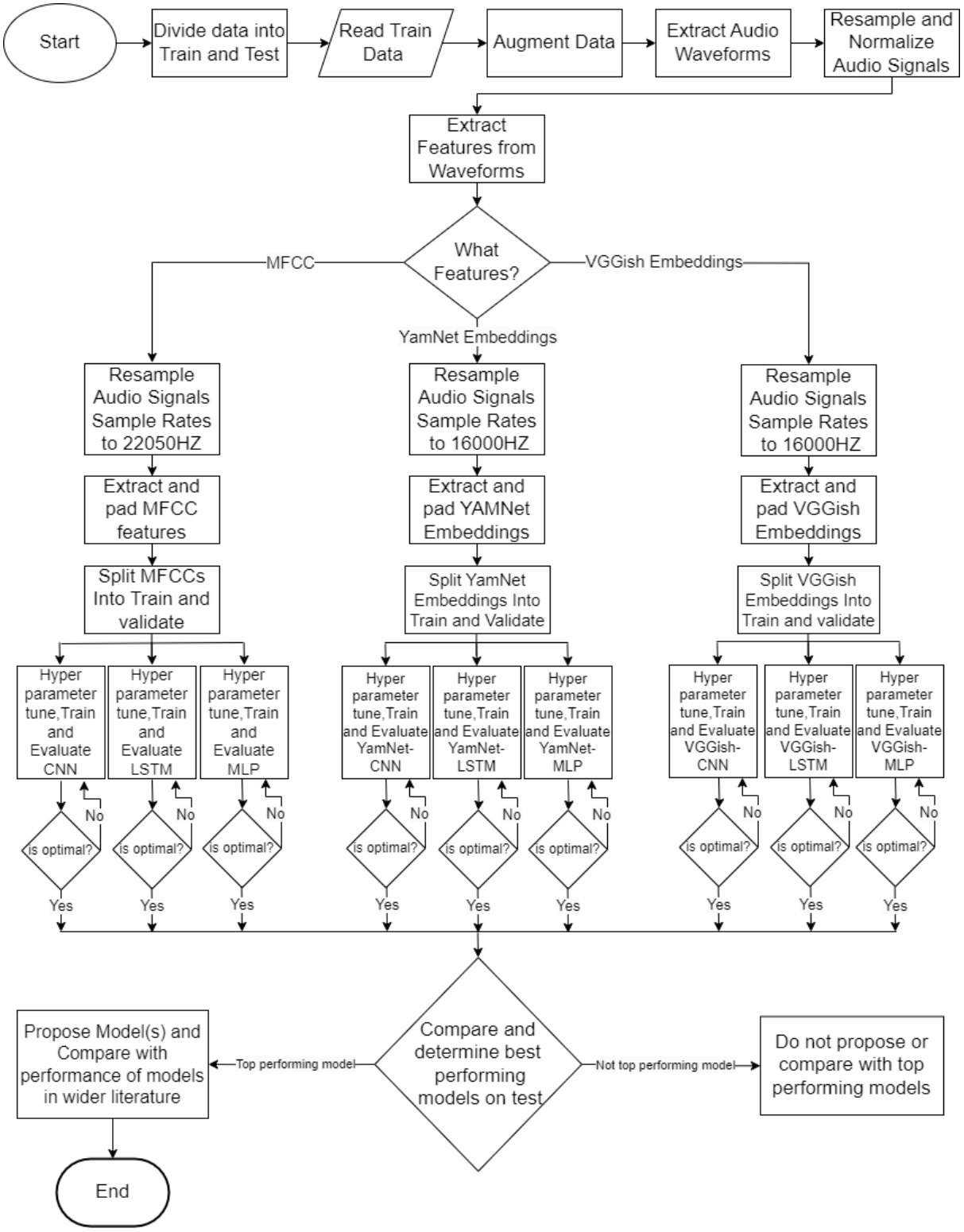

Figure 1. Flow chart showing the methodology and steps taken to complete the research project. The oval shape represents the start of the process, the rectangle shape represents processing steps, the diamond shape represents decision-making steps, the arrows dictate the directions of the flow, and the terminator shape is used to show the end of the flow diagram. It is worth noting that the Hertz (Hz) in the diagram does not represent frequency but sample rate. The reason that both frequency and sample rates share the Hz unit is shown by the Nyquist theorem (Por, Van Kooten and Sarkovic, 2019). It is also worth a mention that the audio signal sample rates were reshaped to 16000HZ due to specifications in their official documentations and the ones for MFCC features were resampled to 22050HZ because according to Kunchur (2007) the human hearing upper limit is 20 to 22KHz and we wanted the features to accurately represent how humans hear.

2.2. PREPROCESSING

Firstly, the male and female audio samples for each accent were combined to remove the need to handle them separately. Secondly, 20% of the data was removed from each accent for the purpose of testing after the models have been built. The remaining 80% of the data was then read in to properly begin the processing. This split was used because according to Gholamy, Kreinovich and Kosheleva (2018), it is an optimal split for machine and deep learning tasks.

2.2.1. DATA AUGMENTATION

The dataset used for training the models in this research is clean, but according to Flamme et al. (2012) there is a lot of noise in real life data. For this reason, we augmented the dataset.

The augmentation used for the dataset included pitch shifting, high pass filtering, adding gaussian noise and adding background noise and they were added to the dataset using probabilities of 0.05, 0.05, 0.18 and 0.81 respectively. The background noise that was used consists of 25 recordings of noise from the University of Hull library, cars passing on the road, rain falling, coffee shops and shopping malls. The augmentation was carried out using the audiomentations[6] library.

[6] https://iver56.github.io/audiomentations/

2.2.2. AUDIO SIGNAL EXTRACTION AND NORMALIZATION

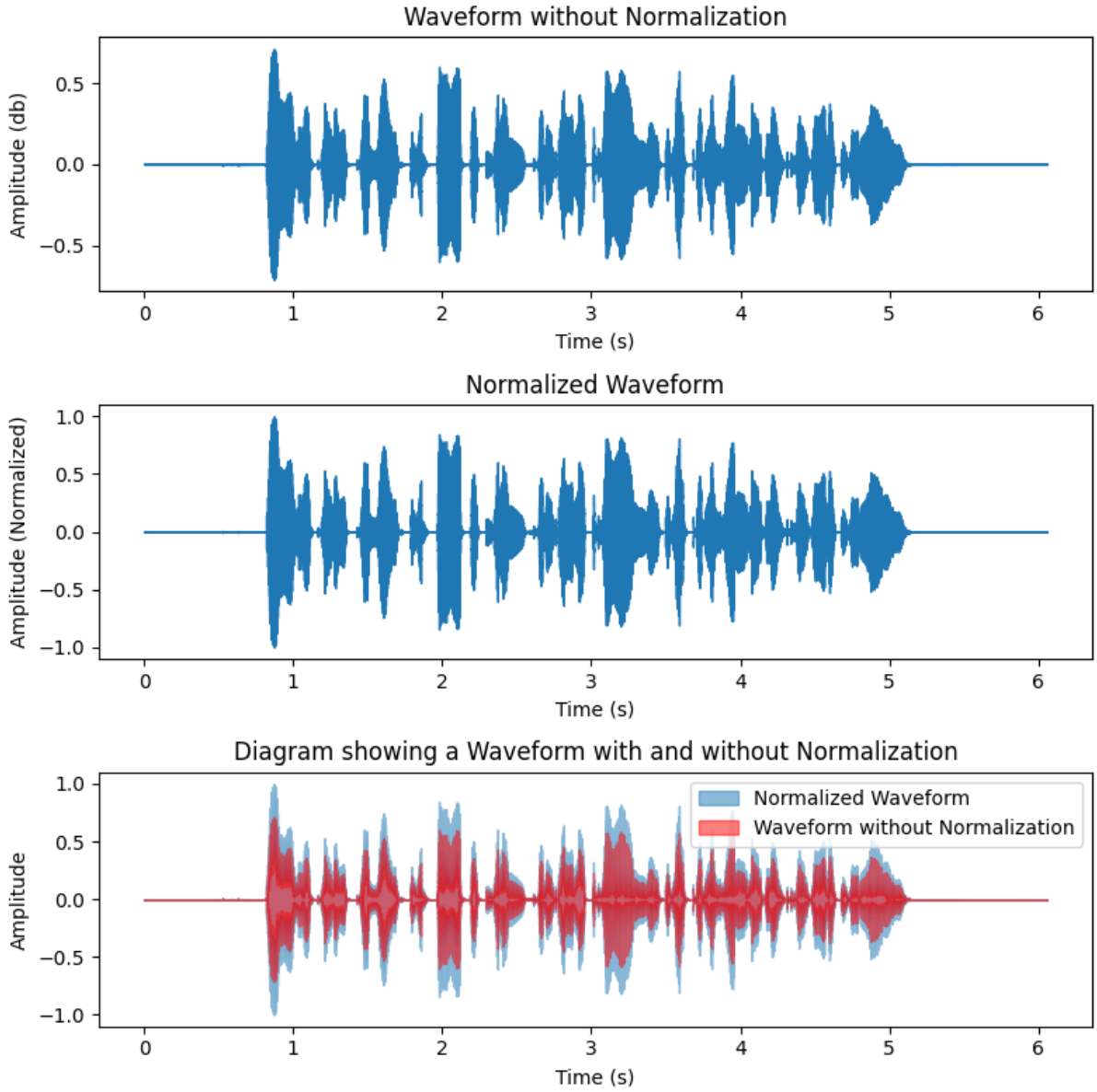

Figure 2. Subplots showing three plots of the same audio wave extracted from a sample in the dataset. The first plot shows the audio wave as it is without normalization, the second plot shows the normalized version of the audio wave, and the third plot shows both the audio wave with and without normalization overlapping each other to show the difference. In the third plot, the red wave represents the waveform without normalization while the blue wave represents the waveform with normalization.

According to Douglas et al. (2009), a waveform is a one-dimensional pattern of amplitude variation with respect to time (as seen in Figure 2), and they are important in audio processing because they directly represent mechanical sound that we hear. For more information on waveforms, see Douglas et al. (2009).

The audio files in the dataset were feature engineered to be in their respective normalized waveforms in the range of -1 to 1 (as seen in Figure 2), so that the amplitude of all the samples in the dataset would have the same scale.

2.2.3. FEATURE EXTRACTION

Although it was possible to train the models using the waveforms extracted, we decided not to because studies like the research by Zhang, Leitner and Thornton (2019) have shown that training models on time domain features like waveforms tend to have very low accuracies in comparison to time and frequency domains. With this in mind, we trained our models with 3 sets of time and frequency domain features namely, MFCCs, VGGish embeddings, and YamNet embeddings which were all padded to match the length of the longest audio.

2.2.3.1. MFCC Features

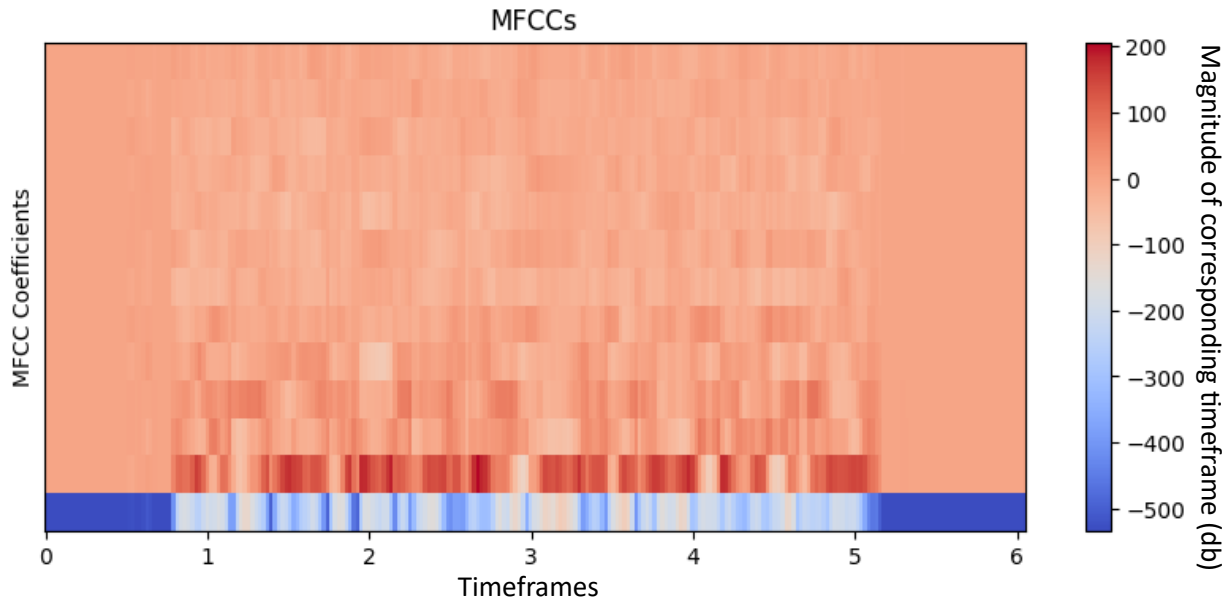

Figure 3. Figure showing the MFCC representation of a sample audio file in the Dataset. The x axis represents the specific time frames in the waveform, the y-axis represents the MFCC coefficients, and the colour bar represents how the magnitude of the MFCC coefficients changes over time and it is measured in decibels(db). The colour intensity at a particular point represents the magnitude of the corresponding time frame at a time frame.

According to Singh (2019), MFCCs are two dimensional features derived from the power spectrum of waveforms, that capture the key elements of the spectral envelopes. In order words, it gives a concise representation of both the time and frequency of an audio signal as seen in Figure 3. For more information on how MFCCs are derived, see the article on MFCCs by Singh (2019).

MFCC features were chosen for this research because firstly, they imitate and focus on the aspects of audio signals that represent the way we humans perceive sound (Nikhate, Chaudhari and Kulkarni, 2019). Secondly, the research by Tantisatirapong, Prasoproek and Phothisonothai (2018), proved that they are the best audio features for accent classification tasks.

Each waveform was resampled to a frequency of 22050Hz for the same reason explained in Figure 1, after which MFCC features with 40 MFCC coefficients were extracted because higher coefficients imply more detail which would help the models learn better according to Rabiner and Juang (1993). These features where then used to train a CNN, LSTM, and MLP models.

2.2.3.2. YAMNET MODEL

YamNet[4] is a transfer learning model that is built on the MobileNet v1 architecture and was trained on AudioSet Corpus[7] to extract high level audio features from audio waveforms. See Wang et al. (2020) for more information on the MobileNet v1.

The YamNet[4] model expects a normalized waveform in the range -1 to 1 with a sample rate of 16000Hz as input, and returns a 3-tuple (scores, embeddings, log Mel spectrograms). For the sake of this research, we focused on just the embeddings because those are high-level features. The embeddings are a 2-dimensional tensor that contains per frame embeddings and the embedding vector is an average-pooled output. For more information on the Yamnet Model, see the official YamNet Documentation[4]

2.2.3.3. VGGISH MODEL

The VGGish[3] model has similar input expectations as the YamNet[4] but unlike it, it is a variant of the VGG architecture and was trained on the YouTube-8M[8] dataset. The model returns a 2-dimensional tensor with the shape (N, 128), where N is the number of frames produced. For more information on the VGGish Model, see the official VGGish documentation[3].

2.3. MODELS

As seen in Figure 1, nine models were built in this research. While each model would be briefly discussed in this section, only the best performing model was discussed in detail later in the results section. Additional information on the models not covered in detail can be found in the accompanying Jupyter notebooks. Also, it should be noted that all the models were built using the TensorFlow module. See the book by Pattanayak (2023) for more information on tensorflow.

2.3.1. CONVOLUTIONAL NEURAL NETWORK (CNN)

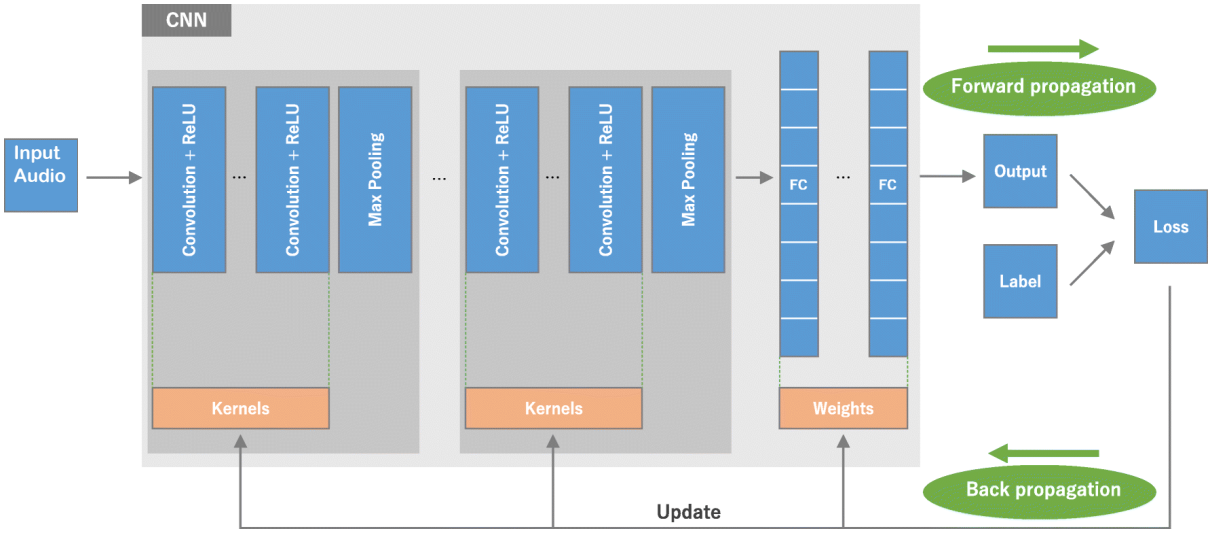

Figure 4. Diagram showing the architecture of a CNN model and its training process (Yamashita et al., 2018). It is worth noting that FC in the diagram stands for the fully connected layer.

As seen in Figure 4, a CNN is a stack of blocks including convolution layers, pooling layers, and fully connected layers. Additionally, the performance of a CNN models is calculated with a loss function through forward propagation, and the kernels and weights are updated based on the loss value via

[7] https://research.google.com/audioset/
[8] https://research.google.com/youtube8m/

backpropagation with the help of an optimizer. For more information on the CNN architecture and training process see the research by Yamashita et al. (2018).

We explored CNNs for this task because the article by IBM (2023) claims that CNNs have superior performances with audio classification. Also, Cetin (2022) achieved an accuracy of 92.67% on our primary dataset using a CNN which implies that they are good for accent classification tasks.

Three CNN models were built for the 3 set of features used in this research as seen in Figure 1

2.3.2. LONG SHORT TERM MEMORY NETWORKS (LSTM)

Before we can talk about an LSTM, we must first talk about Recurrent Neural Networks (RNN). According to Olah (2015), RNNs learn about future events from past ones. This makes RNNs a good choice for time series data like accent data because previous utterances inform future ones to form an accent. But the issue with regular RNNs is that they only keep a history of short-term dependencies which would be an issue for long audio samples. This is the reason we employed the LSTM, which is a variation of the RNN that can keep a history of short term as well as long term dependencies. For more information about RNNs and LSTMs, see Saxena (2021).

Similarly, to the CNN model, 3 LSTM models were trained on the 3 different features discussed as seen in Figure 1.

2.3.3. MULTILAYER PERCEPTON (MLP)

According to Brownlee (2016), MLP is a fully connected multi-layered feed forward neural network. Although MLPs can make predictions on their own, they are often combined with other neural networks to make predictions, for example the fully connected layer in a CNN as seen in Figure 4 is an MLP.

We decided to use an MLP because firstly, we wanted to see how the pretrained model would perform without combining them with complex neural networks. Secondly, there is no research on training an MLP on MFCC features available and so we decided to explore it.

2.4. EVALUATION METRICS

We used 5 metrics to evaluate the performance of the models, but before we delve in, we would look at important measures gotten a confusion matrix that are used to generate these metrics. These measures include the True Positives (TP), True Negatives (TN), False positives (FP), and False negatives (FN).

The True Positives (TP) represents the number of times that the models correctly predict that a sample belongs to a particular accent for each accent. The True Negatives (TN) represents the number of times that model correctly predicts that a sample does not belong to a particular accent for each accent, the False positives (FP) represents the number of times that the models incorrectly make a positive prediction for a particular accent, and the False negatives (FN) represents the number of times that the models incorrectly make a negative prediction for a particular accent across each accent.

The metrics used to evaluate the models include precision, recall, F1-score, Accuracy, and confusion matrix.

Accuracy measures the overall classification correctness of the models across all the classes in the dataset. The equation to calculate accuracy is seen in (1).

Recall measures the proportion of the number of times that the models make true positive predictions and among all the times that the accents truly belong to the accent. The equation to calculate the recall can be seen in equation (2).

The precision is how the accuracy of the positive predictions for each accent is measured. As seen equation (3), it is calculated as the ratio of the TP to the sum of the TP and FP.

The F1-score is important to our research because the dataset that we trained our models on was not balanced and unlike accuracy, it measures the performance of the model in the face of imbalance on each class by taking the harmonic means of the precision and recall. In summary, the F1 score measures the ability of the models to make accurate predictions for each class. The equation for the recall can be seen in equation (4).

Accuracy $$\frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Recall $$\frac{TP}{TP + FN} \quad (2)$$

Precision $$\frac{TP}{TP + FP} \quad (3)$$

F1-Score $$\frac{2\ x\ precision\ x\ recall}{Precision + recall} \quad (4)$$

In addition to these metrics, we would also make use of the confusion matrix to allow us to visualize the predictions that the models made throughout the set, and to prove that the approach of assigning weights each classes handled the imbalance issue with our dataset. For more information on the confusion matrix and the other metrics used, see chapter 15 of the book by Kulkarni, Chong and Batarseh (2020).

3. RESULTS

Table 3. Table showing the performance of the different models on the test dataset. It shows the Precision, recall, f1-score, and support of each of the models built in this research.

Model	Precision	Recall	F1-score	Accuracy
LSTM	79.34%	88.14%	82.33%	80.18%
CNN	92.05%	96.78%	94.18%	95.10%
MLP	74.09%	85.94%	79.58%	69.77%
YamNet-CNN	49.85%	63.25%	50.73%	52.11%
YamNet-LSTM	76.07%	78.37%	76.91%	79.35%
YamNet-MLP	55.82%	71.35%	59.29%	58.40%
VGGish-CNN	66.67%	77.44%	69.78%	69.91%
VGGish-LSTM	85.70%	87.70%	86.38%	86.65%
VGGish-MLP	78.01%	86.12%	81.14%	82.06%

From Table 3, we observe that CNN model trained on MFCCs outperforms the other models across all metrics with an accuracy of 95.10% and weighted precision, recall and F1-score of 92.05%, 96.78%, and 94.28% respectively. We also observe that the VGGish models are the second-best performing models with the VGGish-LSTM leading with an accuracy of 84%. We also observe that the YamNet and MLP models are the poorest which hints that they are not good for the task.

As discussed earlier, we propose and delve further into the architecture and the training of our CNN model since significantly outperformed the rest.

Table 4. Table showing the different set of hyperparameters for the CNN model which performs best, there issues, there losses and their accuracies. There were a lot more parameter sets but these were the ones that we felt stood out.

Set	Hyper Parameter set	Issues	Model Loss	model Accuracy
1	Epochs: 100 Number of Convolutional Layers: 4 Number of Dropout Layers: 3 Dropout Values: 0.5, 0.5, 0.5 Pooling type: Max Number of kernels: 32, 64, 128, 256 Number of Dense Layers: 2 Prediction activation function: SoftMax Optimizer: ADAM Learning Rate: 0.0005 Loss function: categorical Cross Entropy Class weights: Balanced Batch size: 64 Train Size: 64% of dataset Validation Size: 16% of dataset Test Size: 20% of dataset	None	Training: 0.0723 Validation: 0.1063	Training: 96.89% Validation: 96.26% Test: 95.10%
2	Epochs: 50 Number of Convolutional Layers: 3 Number of Dropout Layers: 2 Dropout Values: 0.5, 0.4, 0.5 Pooling type: Max Number of kernels: 32, 64, 128, 256 Number of Dense Layers: 3 Prediction activation function: SoftMax Optimizer: ADAM Learning Rate: 0.005 Loss function: categorical Cross Entropy Class weights: Balanced Batch size: 64 Train Size: 64% of dataset Validation Size: 16% of dataset Test Size: 20% of dataset	Model did not a converge	Training: 0.1814 Validation: 0.2104	Training: 89.52% Validation: 92.24% Test: 88.25%
3	Epochs: 30 Number of Convolutional Layers: 1 Number of Dropout Layers: 1 Dropout Values: 0.2 Pooling type: Average Number of kernels: 64 Number of Dense Layers: 1 Prediction activation function: SoftMax Optimizer: SGD Learning Rate: 0.01 Loss function: Sparse categorical Cross Entropy Class weights: Balanced Batch size: 32 Train Size: 64% of dataset Validation Size: 16% of dataset Test Size: 20% of dataset	Model Overfit	Training: 0.1086 Validation: 0.4688	Training: 92.13% Validation: 69.46% Test: 68.95%

For our proposed model, we selected the parameter set number 1 in Table 4 because they gave the best results without any issues. Figure 5 further proves that training the CNN with parameter set 3 does not lead to issues like overfitting because we see that the training loss is not significantly lower than the validation loss, and the training accuracy and the validation accuracy are practically the same. Additionally, from the shape of the plots in Figure 5 it is evident the CNN model trained with the set of hyperparameters 1 in Table 4 converges which implies that the model learned to its full potential based on the parameter set.

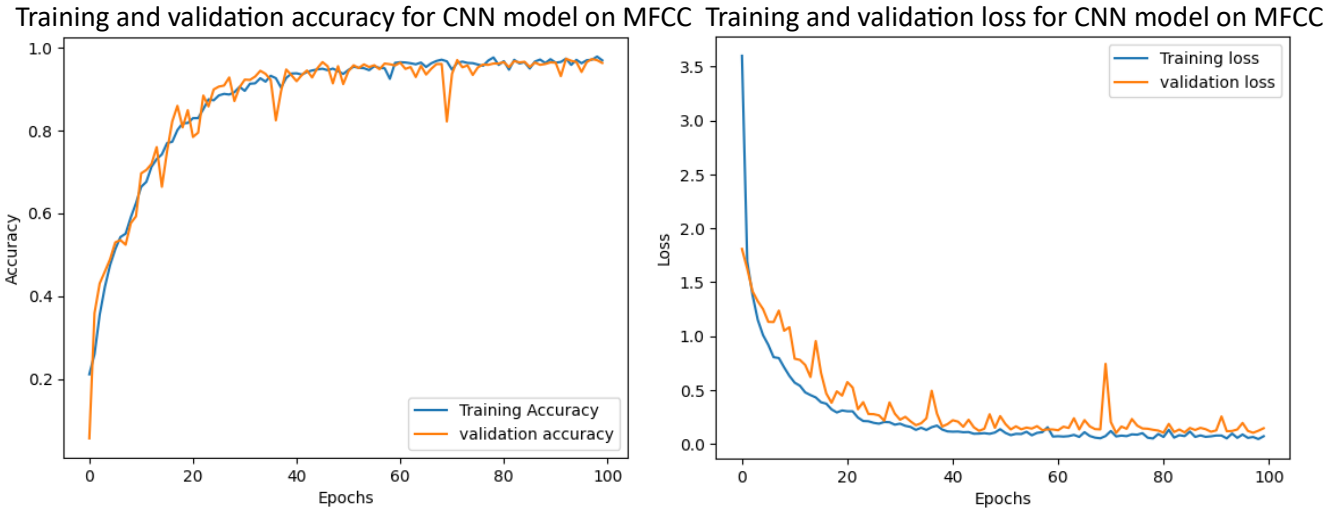

Figure 5. Diagrams showing the Training and validation accuracies and losses for the best performing CNN trained on MFCC features on the Crowdsourced high-quality UK and Ireland English Dialect speech data set dataset using the parameter set 1 in Table 3.

Table 5. Table classification report for the proposed best performing CNN model. The classification report shows the precision, recall and f1-score in percentages and the support (number of samples) for each class for the predictions made on the test split of the Crowdsourced high-quality UK and Ireland English Dialect speech data set[1].

Accent Classes	Precision	Recall	F1-Score	Support
Irish	93.68%	98.89%	96.22%	90
Midlands	78.41%	99.28%	87.62%	139
Northern	94.62%	95.75%	95.20%	569
Scottish	91.21%	96.06%	93.58%	508
Southern	98.02%	93.29%	95.59%	1698
Welsh	96.35%	97.36%	96.10%	569

The first thing that catches the eye in Table 5 is how the support across the classes show imbalance, which is why the model was trained with balanced class weights as seen in Table 4. Also, from Table 5 we observe that the southern class has the highest precision and the lowest recall with values of 98.02% and 93.29% and from equations (2) and (3) this means that the southern class is the class that the model has the highest accuracy in predicting but at the same time, the model fails to identify instances of the class the most. The reason for this is that the support for the southern class is 1,698 which is significantly higher than the support of the other classes. This is why the most suitable metric for comparing the results in this situation is the F1-Score because it balances the balances the recall and the precision, hence it is not affected by class imbalance. The F1-Scores for all the classes are relatively high and consistent and this proves that the model was robust, performs well across all the accents and was not affected by the class imbalance.

Confusion Matrix for the predictions of CNN model trained on MFCC Features from the Crowd Sourced Uk Data

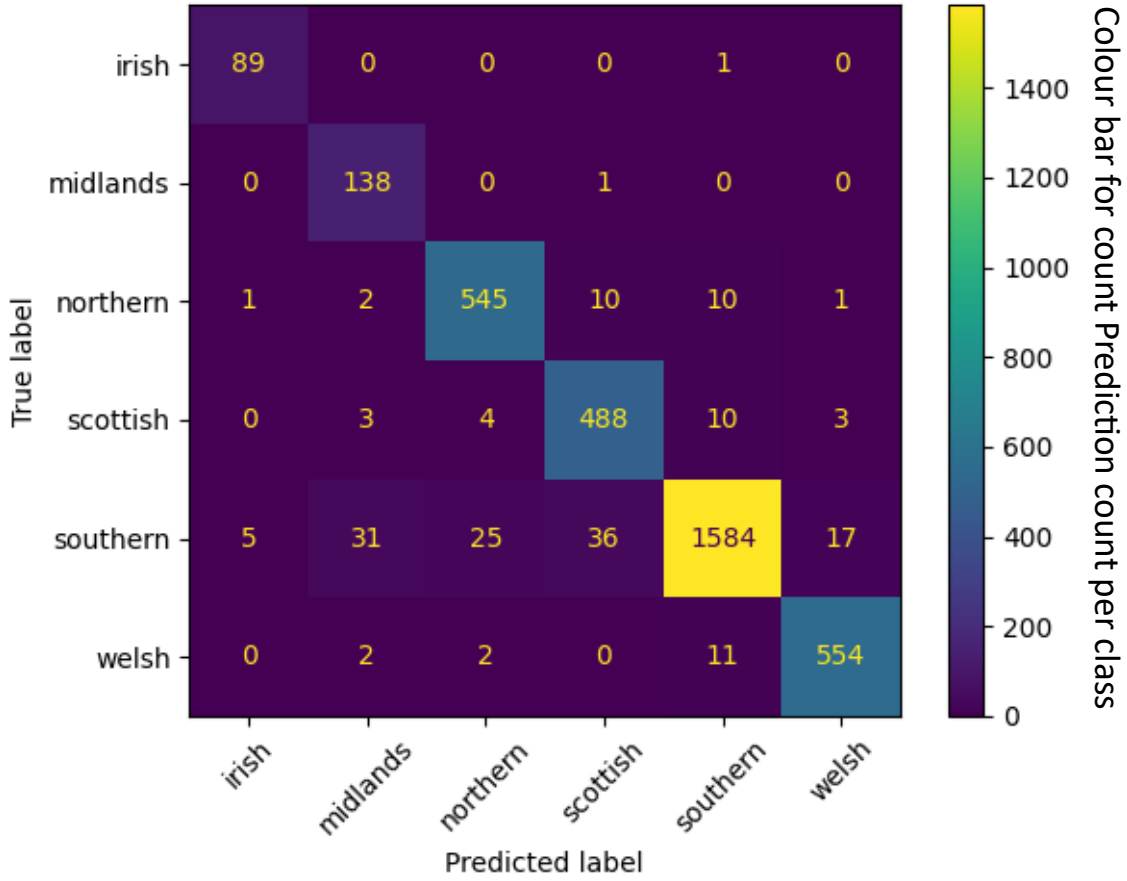

Figure 6. Confusion Matrix showing the predictions of our proposed CNN model trained on MFCC features on the Crowdsourced high-quality UK and Ireland English Dialect speech data set[1] across each accent class. The True label represents the correct label of the accent, the predicted label represents what the class was predicted as, and the colour bar just aids the visualization to see when there are high numbers where the brightness of the colour is directly proportional to the numbers.

Table 6. Table showing the classification report for our proposed model after being fine-tuned on the LibriTTS-British-Accents dataset[5]. The table shows the precision, recall, f1-score, accuracy, weighted average of the metrics and support (number of sampled) for each class in the dataset. In summary, this table shows the performance of our proposed CNN model trained on MFCC features in the task of transfer learning.

Accent Classes	**Precision**	**Recall**	**F1-Score**	**Support**
Irish	96.70%	95.35%	96.02%	215
Scottish	96.45%	93.14%	94.76%	204
Welsh	94.19%	100%	97.01%	162
Accuracy			95.87%	581
Weighted average	95.78%	96.16%	95.93 %	581

Confusion Matrix for our proposed CNN model on the the LibriTTS-British-Accents being finetuned.

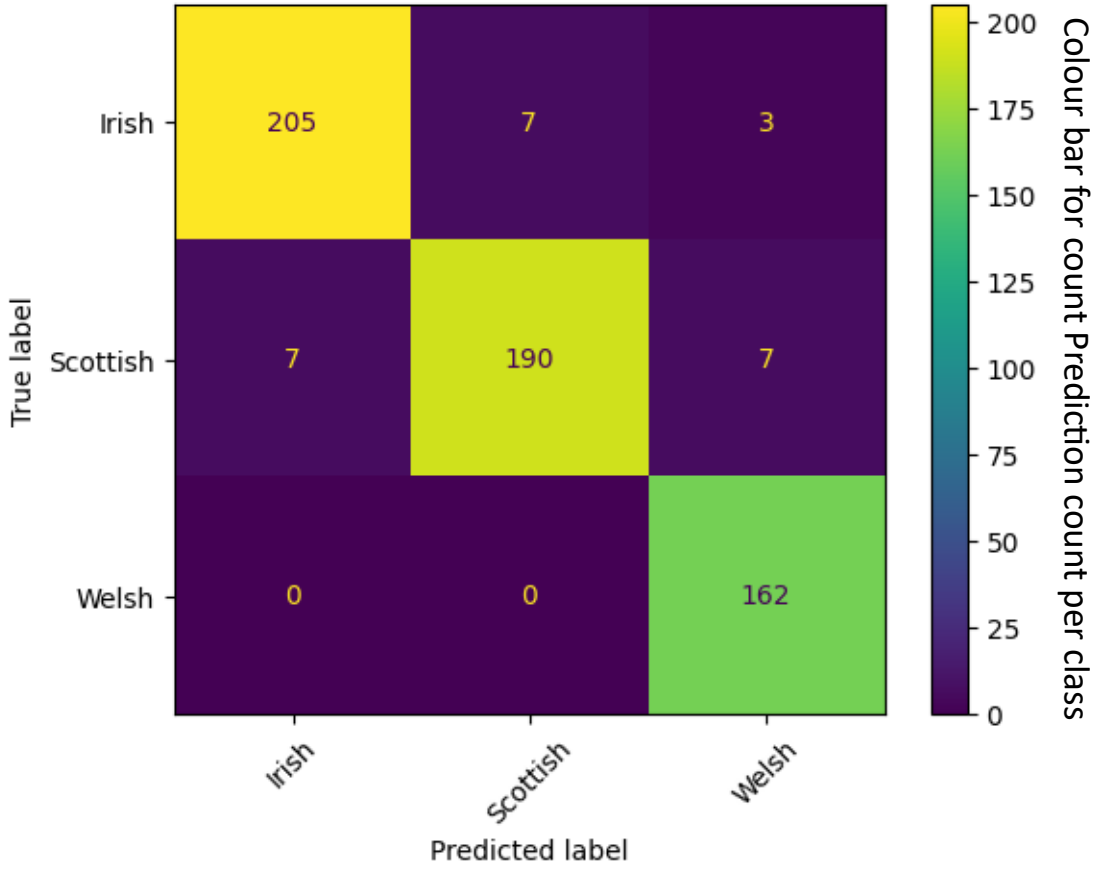

Figure 7. Confusion matrix showing the predictions of our proposed CNN model trained on MFCC features on the LibriTTS-British-Accents[5] dataset after being finetune on it. The True label represents the correct label of the accent, the predicted label represents what the class was predicted as, and the colour bar just aids the visualization to see when there are high numbers where the brightness of the colour is directly proportional to the numbers.

Accent Classes	**Precision**	**Recall**	**F1-Score**	**Support**
Irish	93.68%	98.89%	96.22%	90
Midlands	78.41%	99.28%	87.62%	139
Northern	94.62%	95.75%	95.20%	569
Scottish	91.21%	96.06%	93.58%	508
Southern	98.02%	93.29%	95.59%	1698
Welsh	96.35%	97.36%	96.10%	569

4. DISCUSSION

By comparing the F1-score across all the classes in

Table 5, we make the collusion that our proposed model handled the class imbalance properly. One may make the argument that the model struggles to handle the imbalance because the Midlands class had an F1-Score of 87.62% while the other classes have their F1-scores in the mid-90s, but this is not the case. In fact, looking at the false positives for the midlands class in Figure 6 in relation to its support in

Table 5, we see that the midlands had the highest false positive rate among all the classes and this is because the midlands accent has the most similarities with all the other accents in the UK according to Hughes, Trudgill and Watt (2013).

The precision of the model is also high (90 and above) for all the classes apart from the midlands class which has a precision of 78.41% and the reason is the same as above because if the precision is given by equation (3), then the number of False Positive would be inversely proportional to the precision.

We investigated how well our proposed model would perform in transfer learning using the LibriTTS-British-Accents[5]. As seen in Table 2, the dataset has a total of only 581 samples, and this was good because according to Cai et al. (2020), models cannot perform well on small datasets without leveraging on transfer learning which implied that we could properly test the transfer learning capabilities of our model. We finetuned our pretrained proposed model to fit and make predictions for the LibriTTS-British-Accents dataset[5] and upon testing, we achieved an accuracy of 95.87% and average weighted, F1-score, recall and precision of 95.93%, 96.16% and 95.78% respectively as seen in Table 6. Also, from Figure 7 we argue that the model is nearly perfect on the dataset because it misclassifies only 23 out of 581 samples. These results tell us that our proposed model is suitable for transfer learning in accent detection.

4.1. COMPARISON WITH WIDER LITERATURE

Research Work	Dataset	Features	Techniques	Accuracy	F1-score
(Kristiawan Nugroho, et al., 2023)	Speaker Accent Recognition (2020)	MFCC	Simple Neural Network	82.68%	83.00%
(Cetin, 2022)	Crowdsourced high-quality UK and Ireland English Dialect speech data set[1]	Spectrogram Images	AlexNet	93.39%	93.19%
(Cetin, 2022)	Crowdsourced high-quality UK and Ireland English Dialect speech data set[1]	Spectrogram Images	CNN	92.92	92.67%
(Parikh et al., 2020)	speech accent archive[2]	Waveform and MFCCs	DNN-RNN-CNN	68.67%	67.00%
(Alam, Bhuiyan and Monir, 2023)	Bangla accents dataset gathered by Islam et al. (2020).	Zero Crossing Rate (ZRC), MFCC, Root Mean Square (RMS) and Mel-spectrograms	DNN	94.00%	94.00%

Accent Classes	**Precision**	**Recall**	**F1-Score**	**Support**
Irish	93.68%	98.89%	96.22%	90
Midlands	78.41%	99.28%	87.62%	139
Northern	94.62%	95.75%	95.20%	569
Scottish	91.21%	96.06%	93.58%	508
Southern	98.02%	93.29%	95.59%	1698
Welsh	96.35%	97.36%	96.10%	569

Our Research	Crowdsourced high-quality UK and Ireland English Dialect speech data set[1]	MFCC	CNN (proposed)	95.10%	94.18%
Our Research	LibriTTS-British-Accents[5]	MFCC	CNN	95.87%	96.16%

Table 7. Table showing the performance comparison between the results of our best performing model (CNN) and models built in other research works. The accuracy metric was used so that we could compare the overall correctness of the model while F1-Scores was used so that we could how the models performed in the face of imbalance.

From Table 7, We observe that firstly, our proposed CNN model outperformed the models in the wider literature on the English Dialect speech dataset[1] dataset in both accuracy and F1-score metrics in the wider literature like Cetin (2022). Secondly, CNN models perform better when trained on MFCC features rather than spectrogram images as suggested by Cetin (2022). Thirdly, observe that our proposed model does not only generalize well but is also effective for transfer learning because it was finetuned on the LibriTTS-British-Accents[5] dataset and showed an improved performance. Fourthly we argue that our results are state of the art because our model achieved the best in comparison to the wider literature on the "Crowdsourced high-quality UK and Ireland English Dialect speech[1]" and "LibriTTS-British-Accents[5]" datasets with accuracies of 95.10% and 95.87%. This implies that the model generalized well.

4.2. APPLICATIONS

Although our model is not practical in crime forensics to identify the ethnicity of crime suspects, because of its issue of not performing well on noisy data, it can still be applied to tailoring language learning programs to help individuals learn pronunciations of and speech patterns of the different accents in the UK.

Also, our proposed model can give insights to the UK accents that a non-UK accent is closest to so far as the audio sample is clean. This can help with research to identify the origins of various non-UK English accents.

Finally, it is evident that our proposed model finetuned on the LibriTTS-British-Accents[5] performed very well with an accuracy of 95.87%. This tells us that our proposed model can easily be used for transfer learning especially in cases where the data is limited.

5. CONCLUSION

In this research, we proposed a CNN model trained on MFCC features extracted from the Crowdsourced high-quality UK and Ireland English Dialect speech data set[1] to classify English accents in the UK including the Irish, southern, northern, midlands, Scottish, and Welsh accents and it attained an accuracy, F1-Score, recall and precision of 95.10%, 94.18%, 96.78% and 92.05% respectively, which surpasses every other model used for the task in our research as seen in Table 3. We also made the argument that our proposed model is one of if not the best in the accent classification task in the deep learning domain by making comparisons with wider literature as seen in Table 7. By finetuning our proposed model on the LibriTTS-British-Accents[5] dataset and attaining an accuracy, F1-Score, Recall and precision of 95.87%, 95.93%, 96.16% and 95.78% respectively, we proved that our proposed model qualifies to be highly recommended for transfer learning, especially

in situations where data is limited. Finally, we filled the gap of exploring established audio transfer learning on accent classification tasks by using the VGGish and Yamnet models and proved that the VGGish model is a very good choice for accent classification. This is because other than our proposed model, the VGGish models performed best with the VGGish-LSTM leading the forefront with an accuracy, F1-Score, recall, and precision of 86.65%, 86.38%, 87.70% and 85.70% respectively.

5.1. FUTURE WORKS

From our research, a few works could be considered for the future. Firstly, results in Table 3, we determine that the second best of models are the VGGish models which were pretrained on spectrogram features. Because we have established that MFCC features perform better that spectrogram in accent classification it would be interesting to determine if a version of the VGGish model trained on MFCC features rather than spectrograms would outperform CNN models in the task of accent classification.

Secondly, despite our efforts using augmentation efforts, our proposed model has a low accuracy on unclean audio samples. It would be good if research could be done to not only make accent classification models that would be robust to noise, but also maintain a high level of performance while doing so.

Thirdly, there are only a few research works like the one done by Alam, Bhuiyan and Monir (2023) to classify accents in languages in languages other than English. It would be interesting to see research to determine if the best models and features for accent classification change across different languages.

REFERENCES

Alam, K., Bhuiyan, M.H. and Monir, M.F. (2023) Bangla Speaker Accent Variation Classification from Audio Using Deep Neural Networks: A Distinct Approach. *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*. [online] Available online: https://www.researchgate.net/publication/375842280_Bangla_Speaker_Accent_Variation_Classification_from_Audio_Using_Deep_Neural_Networks_A_Distinct_Approach [Accessed 2 Dec. 2023].

Ashiq, W. (2022) *How many different accents are there across the UK?* Great British Mag. Available online: https://greatbritishmag.co.uk/uk-culture/how-many-british-accents-are-there/ [Accessed 12 Dec. 2023].

Brownlee, J. (2016) *Crash Course On Multi-Layer Perceptron Neural Networks*. Machine Learning Mastery. Available online: https://machinelearningmastery.com/neural-networks-crash-course/ [Accessed 12 Dec. 2023].

Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L. and Pei, J. (2020) Transfer Learning for Drug Discovery. *Journal of Medicinal Chemistry*, [online] 63(16), 8683–8694. Available online: https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.9b02147 [Accessed 27 Mar. 2023].

Cetin, O. (2022) Accent Recognition Using a Spectrogram Image Feature-Based Convolutional Neural Network. *Arabian Journal for Science and Engineering*, [online] 48(2). Available online: https://link.springer.com/article/10.1007/s13369-022-07086-9 [Accessed 12 Dec. 2023].

Demirsahin, I., Kjartansson, O., Gutkin, A. and Rivera, C. (2020) *Open-source Multi-speaker Corpora of the English Accents in the British Isles*. ACLWeb. Available online: https://www.aclweb.org/anthology/2020.lrec-1.804 [Accessed 1 Dec. 2023].

Douglas, S., Duncan, B., Sinclair, I., Brice, R., Hood, J.L., Singmin, A., Davis, D., Patronis, E. and Watkinson, J. (2009) *Audio Engineering: Know It All*. *Google Books*, Newnes, 411–413. Available online: https://books.google.co.uk/books?hl=en&lr=&id=RDfKSh3Q8kAC&oi=fnd&pg=PP2&dq=what+is+audio+waveform [Accessed 5 Dec. 2023].

Flamme, G.A., Stephenson, M.R., Deiters, K., Tatro, A., van Gessel, D., Geda, K., Wyllys, K. and McGregor, K. (2012) Typical noise exposure in daily life. *International Journal of Audiology*, [online] 51(sup1), S3–S11. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4685462/ [Accessed 12 Dec. 2023].

Gholamy, A., Kreinovich, V. and Kosheleva, O. (2018) Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)*. [online] Available online: https://scholarworks.utep.edu/cs_techrep/1209/ [Accessed 14 Dec. 2023].

Hassan, M. (2023) *Farewell to King's English? Traditional accents on the decline*. The National News. Available online: https://www.thenationalnews.com/world/uk-news/2023/10/31/farewell-to-kings-english-traditional-accents-on-the-decline [Accessed 29 Nov. 2023].

Hughes, A., Trudgill, P. and Watt, D. (2013) *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles, Fifth Edition*. *Google Books*. Routledge. Available online: https://books.google.co.uk/books?hl=en&lr=&id=a80iE9jWx1AC&oi=fnd&pg=PP2 [Accessed 8 Dec. 2023].

IBM (2023) *What are Convolutional Neural Networks? | IBM*. www.ibm.com. Available online: https://www.ibm.com/topics/convolutional-neural-networks [Accessed 12 Dec. 2023].

Islam, S., Rahaman, H., Farea Rehnuma Rupon and Sheikh Abujar (2020) Bengali Accent Classification from Speech Using Different Machine Learning and Deep Learning Techniques. *Advances in intelligent systems and computing*, [online] 503–513. Available online:

https://www.researchgate.net/publication/347219331_Bengali_Accent_Classification_from_Speech_Using_Different_Machine_Learning_and_Deep_Learning_Techniques [Accessed 2 Dec. 2023].

Kiran, U. (2021) *MFCC Technique for Speech Recognition*. Analytics Vidhya. Available online: https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition [Accessed 5 Dec. 2023].

Kristiawan Nugroho, Edy Winarno, Eri Zuliarso and Sunardi Sunardi (2023) Multi-Accent Speaker Detection Using Normalize Feature MFCC Neural Network Method. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, [online] 7(4), 832–836. Available online: https://www.researchgate.net/publication/373120975_Multi-Accent_Speaker_Detection_Using_Normalize_Feature_MFCC_Neural_Network_Method [Accessed 1 Dec. 2023].

Kulkarni, A., Chong, D. and Batarseh, F.A. (2020) *5 - Foundations of data imbalance and solutions for a data democracy*. ScienceDirect. Available online: https://www.sciencedirect.com/science/article/abs/pii/B9780128183663000058 [Accessed 15 Dec. 2023].

Kunchur, M.N. (2007) Probing the temporal resolution and bandwidth of human hearing. *Proceedings of Meetings on Acoustics*. [online] Available online: https://pubs.aip.org/asa/poma/article/2/1/050006/842033/Probing-the-temporal-resolution-and-bandwidth-of [Accessed 12 Dec. 2023].

Nikhate, S.N., Chaudhari, A. and Kulkarni, J. (2019) Determination of Extent of Similarity between Mimic and Genuine Voice Signals Using MFCC Features. [online] Available online: https://ieeexplore.ieee.org/document/8822061 [Accessed 8 Dec. 2023].

Olah, C. (2015) *Understanding LSTM Networks*. Github.io. Available online: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed 12 Dec. 2023].

Parikh, P., Velhal, K., Potdar, S., Sikligar, A. and Karani, R. (2020) English Language Accent Classification and Conversion using Machine Learning. *SSRN Electronic Journal*. [online] Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3600748 [Accessed 2 Dec. 2023].

Pattanayak, S. (2023) Introduction to Deep-Learning Concepts and TensorFlow. *Pro Deep Learning with TensorFlow 2.0*, [online] 109–197. Available online: https://link.springer.com/chapter/10.1007/978-1-4842-8931-0_2 [Accessed 15 Dec. 2023].

Por, E., Van Kooten, M. and Sarkovic, V. (2019) *Nyquist-Shannon sampling theorem 1 Theory 1.1 The Nyquist-Shannon sampling theorem*. *https://home.strw.leidenuniv.nl/~por/AOT2019/docs/AOT_2019_Ex13_NyquistTheorem.pdf*. Available online: https://home.strw.leidenuniv.nl/~por/AOT2019/docs/AOT_2019_Ex13_NyquistTheorem.pdf [Accessed 12 Dec. 2023].

Rabiner, L.R. and Juang, B.-H. (1993) *Fundamentals of Speech Recognition*. *Google Books*. Pearson Education. Available online: https://books.google.co.uk/books/about/Fundamentals_of_Speech_Recognition.html?id=dVUGAAAACAAJ&redir_esc=y [Accessed 8 Oct. 2023].

Saxena, S. (2021) *What is LSTM? Introduction to Long Short-Term Memory*. Analytics Vidhya. Available online: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm [Accessed 12 Dec. 2023].

Singh, T. (2019) *MFCC's Made Easy*. Medium. Available online: https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040 [Accessed 12 Dec. 2023].

Speaker Accent Recognition. (2020) UCI Machine Learning Repository. Available online: https://doi.org/10.24432/C52329 [Accessed 1 Dec. 2023].

Tantisatirapong, S., Prasoproek, C. and Phothisonothai, M. (2018) *Comparison of Feature Extraction for Accent Dependent Thai Speech Recognition System*. IEEE Xplore. Available online: https://ieeexplore.ieee.org/document/8465705 [Accessed 12 Dec. 2023].

Walfisz, J. (2023) *Would You Adam and Eve it? Cockney Accents Are Disappearing in Britain*. Euronews. Available online: https://www.euronews.com/culture/2023/10/31/cockney-and-the-kings-english-accents-no-longer-predominate-in-london [Accessed 29 Nov. 2023].

Wang, W., Li, Y., Zou, T., Wang, X., You, J. and Luo, Y. (2020) *A Novel Image Classification Approach via Dense-MobileNet Models*. Mobile Information Systems. Available online: https://www.hindawi.com/journals/misy/2020/7602384/ [Accessed 12 Dec. 2023].

Yamashita, R., Nishio, M., Do, R.K.G. and Togashi, K. (2018) Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, [online] 9(4), 611–629. Available online: https://pubmed.ncbi.nlm.nih.gov/29934920/ [Accessed 12 Dec. 2023].

Zhang, B., Leitner, J. and Thornton, S. (2019) *Audio Recognition using Mel Spectrograms and Convolution Neural Networks*. Available online: http://noiselab.ucsd.edu/ECE228_2019/Reports/Report38.pdf [Accessed 5 Dec. 2023].