National College of Ireland

Classification of Speaker's Age, Gender and Nationality using Transfer Learning

MSc Research Project
MSc in Data Analytics

Rohan Narayan Koli
Student ID: 19224842

School of Computing
National College of Ireland

Supervisor: Prof. Majd Latifi

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Rohan Narayan Koli
Student ID:	19224842
Programme:	MSc in Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Prof. Majd Latifi
Submission Due Date:	20/12/2018
Project Title:	Classification of Speaker's Age, Gender and Nationality using Transfer Learning
Word Count:	5120
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	23rd September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Abstract

Audio classification remains one of the most complex and challenging problems in the 21st century. While much analysis and research has been adopted in audio classification in sub-categories of audio scene classification and bio-acoustics, there has been very few researches adopting Human voice classification. In this research I explored the use of pretrained deep convolutional neural networks learning models for the classification task on log-Mel Spectrograms. Five pretrained models (Xception, Vgg16, Vgg198, ResNet50, Inception V3) along with model stacking are compared with respect to two datasets namely, Mozilla Common Voice and Speech Accent dataset. The research was able to achieve 95% accuracy for gender classification while the age group and nationality classification achieved satisfactory results with accuracy 52% and 48% accuracy respectively which can further be utilized to develop enhanced models.

Area Audio Classification, Pretrained Networks, Stacked Ensemble Model, Transfer Learning.

Contents

1 Introduction **1**
1.1 Background 1
1.2 Motivation and Application 1
1.3 Research Goal 2
1.4 Research Question 3
1.5 Report Structure 3

2 Related Work **3**
2.1 Audio Classification based on Machine Learning or Deep Learning 3
2.2 Audio Classification based on Transfer Learning or Pre-trained networks 6

3 Methodology **8**
3.1 Dataset 9
3.2 Data Preparation 9

4 Design Specification **11**
4.1 Design Framework 11
4.2 Pre-trained Model Architecture 11

5 Implementation **14**
5.1 Development Environment 14
5.2 Data Handling or Data Wrangling 15
5.3 Model Implementation 15

6 Evaluation **16**
6.1 Experiment 1: Gender Classification 16
6.2 Experiment 2: Age-Group Classification 17
6.3 Experiment 3: Nationality or Region Classification 17
6.4 Discussion 18

7 Conclusion and Future Work **18**

Classification of Speaker's Age, Gender and Nationality using Transfer Learning

Rohan Narayan Koli
19224842

1 Introduction

1.1 Background

Audio classification, also known as sound classification, is the process of listening and analysing audio signals. The AI is trained to learn how and what to listen and develop differentiation between different sounds to predict specific tasks. The latest technology utilizing such trends include virtual assistants (Alexa, Siri, Google voice assistants), text to speech applications (Microsoft Azure, Watson), speech recognition systems (IBM, Dragon Anywhere). Further, audio classification can be classified into five major classes based on use-cases, namely, Acoustic data, environmental sound, music, natural language utterance, human organ (medical domain) audio classification[1].

In Acoustic data classification (event classification), environment is monitored or classified based on the source of the audio recorded such as schools, offices, busy traffic streets, homes and so on (Chandrakala and Jayalakshmi; 2020). Secondly, Environmental sound classification involves sound classification within a given set of environments. For example, classifying and differentiating urban sounds such as horns, sirens of police or ambulance, human chatter, construction work in a city (Toffa and Mignotte; 2020). Music classification consists of classifying music genre as proposed by M R and Mohan B S (2020), instrument identification and separation by Bhagwat et al. (2020) and music generation and recommending music based on preferences and trends (Eg. Apple Music, Spotify). In natural language utterance classification, recorded voices of human speech are classified based on language, dialect, and semantics (Hossain et al.; 2020). This type of classification forms a base for virtual assistants proposed by Sripriya et al. (2020) like Apple Siri, Amazon Alexa, Google voice and for chatbots (Muhammad et al.; 2020). Lastly, in human organ sound classification, the audio from human organs such as heart by Evangelista et al. (2020), lungs by Shethwala et al. (2021), throat by Anupam et al. (2021) are classified for medical diagnosis of diseases, abnormalities or any symptoms which signify underlying conditions.

1.2 Motivation and Application

As mentioned above, Natural language classification is a type of sound classification technique which is used to classify human speech. Human speech varies according to their age, gender and the region they live in. The human speech system comprises of

[1] `https://www.telusinternational.com/articles/what-is-audio-classification`

three main components, namely, the vocal chords (voice box), lungs and the nasal and oral cavities [2]. The changes in human voices are very minute and subtle and they can only be recognized by trained professionals such as Speech-Language pathologists. Another way of identifying such changes can be through the use of deep learning techniques when a machine is trained on a relative dataset. Such a model can have multi-fold applications as listed below:

1.2.1 E-commerce and Retail Industry

Different people with different demographics can have different needs and requirements. Some people can be more inclined towards buying technological products such as laptops, mobile phones while some inclined towards home décor and personal cosmetic products. Similarly, people from different regions around the globe may have different preferences while shopping in a supermarket, for example, people from India will buy items such as rice, foodgrains (wheat, pulses) and spices whereas people from America are more likely to buy Mac and Cheese, Waffles, Pizza, cookies. Also, people of different age groups demand different products. A child may be require toys, sanitary items (diapers, baby wipes), stationery items while a grown up may demand better clothing and accessories. Hence, segmenting the buyers based on gender, age-group and region can be very beneficial to target audience for their respective needs which can be achieved through human voice classification.

1.2.2 Criminal Forensics

Human voice recordings such as telephonic conversations or voices recorded on CCTV can act as a evidence in the court of law or any criminal investigations. These sources can as a media to profile the culprits, or any suspicious people while investigating. Thus, the suggested model can be used to profile criminals and tried in the courtrooms.

1.2.3 Security and Access

The speech audio data can also be used at border checks or in collaboration with biometric scanners to provide access to specific people at public restricted places. The system can be used to keep a check on the people transiting to or arriving from different countries. Such a system can also be employed at highly secure facilities to grant access.

1.3 Research Goal

The main goal of the research is to explore and implement the use of pre-trained neural networks for classification of speaker's age, gender and nationality using speed audio. The fundamental idea behind the research project is that, the audio data can be represented as a 2-D image. The audio data when subjected to a fast Fourier transform (FFT), and plotted on Mel scale, I derived a Mel Spectrogram for the respective audio signal. This technique is extensively used for audio classification, which essentially becomes an image classification problem. To address the image classification problem, convolution neural networks (CNN) are widely used and have been proven as state-of-the-art technique in the field of computer vision.

[2] `https://www.britannica.com/science/phonetics`

Further, there are many open-source pre-trained CNN models which have been tested on huge image classification datasets such as COCO[3], ImageNet[4], MNIST[5], CIFAR[6] that achieved commendable classification results. Models such as VGG-16/19, GoogleNet, AlexNet, ResNet50 have already been trained using superior hardware (GPU's and RAM with high memory capacities) and which are very expensive. Hence, utilizing such pre-trained models with pre-defined saved weights can give us an edge over the basic CNN architecture. Hence, such a model can achieve the task of natural speech classification efficiently with shortest amount of time.

1.4 Research Question

To what extent can a pre-trained deep learning approach utilizing data preprocessing provide better performance for identification of an individual's age, gender and native language (nationality) through audio speech data?

1.5 Report Structure

In continuation to the discussion, I shall discuss the recent work or accomplishments related to the field of human speech classification in section 2. In section 3 the implemented methodology will be discussed in detail followed by section 4 wherein the requirements and architecture of the model are discussed. Section 5 describes the actual model implementation along with outputs followed by section 6 where the results using evaluation metrics are discussed. Lastly, in section 7, I will discuss limitations and future work to improvise the model and approach at a future date.

2 Related Work

In this section, an overview of the previously conducted research based on audio classification are discussed. I also provided a comparison and critical review of the related research work using two sections, based on machine learning or deep learning approaches and secondly using pre-trained deep learning models.

2.1 Audio Classification based on Machine Learning or Deep Learning

Audio classification being one of the most complex challenges in the 21st century, there have been multiple research and efforts to demystify the challenges based on different categories of audio classification.

2.1.1 Classifying Human Age and Gender

In a research by Pandey (2020), the researcher trained two different models based on CNN and LSTM-RNN architecture to classify human age group. The models were trained on

[3]COCO dataset : `https://cocodataset.org/`
[4]ImageNet Dataset : `https://image-net.org/`
[5]MNIST dataset : `http://yann.lecun.com/exdb/mnist/`
[6]CIFAR dataset : `https://www.cs.toronto.edu/~kriz/cifar.html`

Speech Accent Archive dataset which were converted to Mel Frequency Cepstal Coefficient (MFCC) and trained on CNN architecture and LSTM architecture. The LSTM architecture (66.07% accuracy) provided better results compared to the CNN architecture (62.45% accuracy). Even though the LSTM model achieved better results, the CNN model was not optimized or fine-tuned using different parameters and activation functions.

In a similar research by Kuchebo et al. (2021), Mozilla Common Voice dataset was used to classify age and gender using a CNN network. Two different CNN architectures were used for separate tasks of classifying age and classifying gender. The task of gender classification achieved an accuracy of 90.08% where as the task of age group prediction achieved poor results accuracy of nearly 43%. The researcher could have tried different architectures along with very deep convolution architectures.

In another research by Qawaqneh et al. (2017), instead of MFCC, a transformed MFCC was utilized using bottle-neck feature (BNF) extractor with the deep neural network model to generate robust features. The results were evaluated based on I-Vectors as well as DNN which were 13% more improved compared to the traditional MFCC extraction technique. Although the accuracy improved by 13%, they were still below the model created by Pandey (2020) i.e in the range of 55-60% and required more tuning.

In the proposal by Kaya et al. (2017), Partial least Square (PLS) method and min-max normalization was used and achieved highest unweighted average recall (UAR) is achieved (57%). The research states that humans perceived the classification more accurately than the automatic systems. The automatic systems could have yielded better classification results by utilizing CNN networks.

In another proposal by Bahari et al. (2014), the research utilizes modelling based on i-vectors. The approach yields a pearson's correlation coefficient of 0.772 and a mean absolute error of 6.08 compared to the baseline model of 2% and 5% respectively. The technique takes into account MAE and pearson's correlation coefficient and ignores the accuracy, sensitivity and specificity as evaluation metrics which would have resulted in a clearer feasibility of the model.

2.1.2 Classifying Human Emotions and Speech fluency

In a research by Atsavasirilert et al. (2019), human emotions from 3 second audio recordings were classified using deep CNN networks. The model resulted in 85.54 weighted average recall (WAR), 87.16 unweighted average recall (UAR) and an accuracy of 77.99%. The researcher also noted that increasing the number of log Mel spectrograms resulted in improved performance. The model though used fewer parameters than AlexNet, should have been compared with other pre-trained models as well along with grid search algorithms to find the best hyper-parameter tuning.

In another research by Preciado-Grijalva and Brena (2018), fluency of individuals were tested in the English language utilizing the Avalingua audio dataset. Five different ML models were tested, out of which support vector machine (SVM) achieved an accuracy of 94.39% where as others achieved significant accuracy of around 89%. At a 20-Mel filter bank, the researcher compared the SVM, random forest, CNN, RNN and MPL algorithms.

2.1.3 Classifying Human Organ Sounds

Balamurugan et al. (2019) in their proposal made use of stacked Residual Networks

(ResHNet) to classify heart sounds of people with normal and abnormal heart rhythms. The audio signals were converted into short FFT before plotting on Mel Scale (Mel Spectrograms). The ResHNet architecture was unique containing pooling and convolution layer at lower levels followed by stacked residual modules and 3 CNN embedded in them. The model notably yielded a sensitivity of 97.4% and specificity of 97.1%. The dataset used was imbalanced with only limited 3240 audio recordings.

2.1.4 Classifying Environmental Sounds

The research by Chi et al. (2019) illustrates the application of deep learning in classifying environmental sounds, the researcher suggested the use of both log-Mel spectrogram (LMS) and Log-Gammatone spectrogram (LGS) using just the first 4 blocks of VGG-13 architecture. The model resulted in a classification accuracy of 83.8% and 80.3% accuracy with the ESC-50 and Urban8k dataset respectively. The model was trained using standard hyper-parameters and validated using k-fold cross validation technique.

In another research by Ramirez et al. (2018), Inverse Mel frequency cepstral coefficients were used to classify bird species according to bio-acoustics. This approach resulted in better accuracy compared to the traditional MFCC approach. The data was further trained on hidden Markov model (HMM) only and other models like SVM, random forest were ignored which have the potential for better classification accuracy.

Table 1: Summary of related work based on Machine or Deep Learning

Title	Approach	Advantages	Limitations
Classification of Human Age Group by Implementing Deep Learning Models on Audio Data	MFCC features with CNN and LSTM	Accuracy of 66.07%	Avoided hyper-parameter tuning
Convolution Neural Network Efficiency Research in Gender and Age Classification From Speech	Two different CNN for separate tasks of age and gender classification	Class. Accuracy: Gender: 90.08% Age: 43%	Did not explore different DNN architectures
Deep neural network framework & transformed MFCCs for speaker's age and gender classification	BNF + DNN with MFCC feature extraction	13% better classification	Comparatively below average results
Emotion, age, and gender classification in children's speech by humans and machines	PLS method with min-max normalization	UAR : 57%	Human perception more accurate
Speaker age estimation using i-vectors	Normalization + LSSVR technique	Pearson's: 0.772 MAE : 6.08	Evaluation parameters Sensitivity, Specificity ignored.
A Light-Weight Deep Convolution Neural Network for Speech Emotion Recognition using Mel-Spectrograms	DNN with hyper-parameter tuning	WAR : 85.54 UAR : 87.16	Comparison with other pretrained model missing
Speaker Fluency Level Classification Using Machine Learning Techniques	SVM for classification	Accuracy : 94.39%	DNN model were not hyper-parameter tuned for comparison
ResHNet: Spectrograms Based Efficient Heart Sounds Classification Using Stacked Residual Networks	ResHNet : CNN + Stacked residual modules	Sensitivity : 97.4%; Specificity : 97.1%	Imbalanced dataset with few instances
Deep Convolutional Neural Network Combined with Concatenated Spectrogram for Environmental Sound Classification	LMS + LGS + VGG-13 architecture	Accuracy: ESC- 50 : 83.8%; Urban8k : 80.3%	Model lacked hyper-parameter tuning
A comparative between MFCC and IMFCC features for an Automatic Bird Species Recognition System	IMFCC + HMM	TRR : 61.90	SVM, RF could have had better accuracy

2.2 Audio Classification based on Transfer Learning or Pre-trained networks

Along with ML algorithms, researchers also applied deep learning pre-trained models to classify audio data.

2.2.1 Human-based Audio Classification

In the research by Le et al. (2019), baby cries were classified using Mel Spectrograms as input features to transfer learned networks. Three models were tested separately, namely, ResNet50, SVM and an ensemble of ResNet50 with SVM. The ensemble model achieved a better accuracy of 91.1% whereas ResNet50 alone acheived an accuracy of 90.80%. For evaluation, sensitivity and specificity are better metrics rather than accuracy.

In another porposal by Koike et al. (2020) and Wodzinski et al. (2019), the researchers made use of pretrained networks to classify heart sounds and detect Parkinson's disease respectively. In research by Koike et al. (2020) the Mel Spectrograms were trained on seven distinct pre-trained models. The PANNs architecture (pretrained audio neural networks by Kong et al. (2020)) achieved better results with UAR of 89.7% along with sensitivity and specificity of 96.9% and 88.6% respectively.

On the other hand, in the research proposed by Wodzinski et al. (2019), similarly Mel Spectrograms were used as input features to the ResNet model . Secondly, 10-fold cross validation technique was implemented to generalize the results. The model achieved and accuracy of 90.80% but lacked other evaluation parameters for comparison with similar architectures.

2.2.2 Bio-acoustic Audio Classification

In the research proposed by Chandu et al. (2020), the researcher implemented pretrained models to classify bird acoustic data. Since the avian sounds mostly fall in the higher frequency domains which were pre-emphasized and convereted into sprectrograms which acted as input to the DCNN models. For the research, AlexNet was considered which achieved an accuracy of 97%. A comparison between different pretrained models is necessary to see if an opportunity for better classification results exists. Also, the DCNN model should have been fine-tuned for better results.

In another research by Nanni et al. (2020), for the input features four types were considered, namely, scattergrams, spectrograms, harmonic percussion images and combined non-overlapping images of previous three images. The datasets used were based on different animal species and for deep pretrained models, an ensemble of six architectures were considered. Though the ensemble model achieved superior results, it required a lot of computing for initial model training and pre-processing.

In a similar work by Zhong et al. (2020) to classify multi-species, three models VGG-16, ResNet50, and ResNet50 with pseudo labelling was used. In the first model, Adam optimizer was used while, in the second model, initial layers of the ResNet50 were freezed followed by fully connected layers. The model with pseudo labelling achieved a superior classification accuracy of 95%. Compared to Nanni et al. (2020), the researched could have considered other pretrained models as well and different evaluation parameters.

2.2.3 Music Instrument Audio Classification

The research by Hall et al. (2019) and Shukla et al. (2020) suggests implementation of pretrained network for classifying musical signals based on different instruments. Hall et al. (2019) made use of spectrograms by applying SFT(Short-Time Fourier Transform) to the audio samples for creating input features. Secondly, AlexNet was used as a model to classify these spectrograms. Instead of 11 classes, only 7 classes were trained for 30 epochs which resulted in an accuracy of 73.7%.

Similarly, in the research by Shukla et al. (2020), only string instruments were classified by adding preprocessing techniques as well (MFCC, Constant-Q transform, Mel Spectrograms). Pretrained network VGG preformed well with Mel Spectrograms with the accuracy of 77.5%. Notably, on comparison, VGG performs better than AlexNet. Also, although the performance of both the architecture was commendable, they lacked proper optimization and cross-validation techniques.

2.2.4 Environmental Scene Audio Classification

In a research by Kong et al. (2020), various pretrained models such as (Google CNN, DeepRes, Single/Multi-level attention TAL Net) were compared and a custom architecture was suggested based on Wavegram-LogMel-CNN features which achieved a mAP(mean average precision) of 0.439. The model still remains to be tested on different datasets to ensure feasibility.

The datasets DCASE [7], ESC-50[8] and UrbanSound8k[9] are extensively used for acoustic scene classification. In the work by Palanisamy et al. (2020), the above three datasets were considered and convereted to Mel Spectrograms as input features to pretrained models DenseNet, ResNet and Inception out of which the models DenseNet and ResNet were able to achieve highest validation accuracy. In a similar research by Copiaco et al. (2019), six pre-trained models were tested based on scalograms obtained from DCASE and SINS datasets. The architecture of AlexNet performed better on DCASE dataset (F1-score 93.37%) whereas Xception model performed well on the SINS dataset (F1-score 94.31%).Taking a different approach, Zhou et al. (2018) used two different channels (left and right channels) to extract data and plot spectrograms as input features. Out of the three pretrained models (AlexNet, VGGNet and ResNet), VGG-16 with MLP achieved a better accuracy of 77.8%. Lastly, McMahan and Rao (2017) proposed DenseNet to be the preffered model architecture for classification based on Urban8k dataset with an accuracy of 71.86%. The three research lacked optimization by fine tuning hyper parameters and pre-processing techniques like silence removal, pre-emphasis were missing.

Hence, it can assert that, pretrained networks perform better than vanilla deep learning models and basic statistical models. The Table 2 illustrates the summary of the discussed research utilizing pre-trained models.

[7]DCASE Dataset : `http://dcase.community/`

[8]ESC-50 Dataset : `https://www.cs.cmu.edu/~alnu/tlwled/esc50.htm`

[9]UrbanSound8k Dataset : `https://urbansounddataset.weebly.com/urbansound8k.html`

Table 2: Summary of related work based on Transfer Learning

Title	Approach	Advantages	Limitations
Using Transfer Learning, SVM, and Ensemble Classification to Classify Baby Cries Based on Their Spectrogram Images	ResNet50, SVM, Emsemble (ResNet50+SVM)	Accuracy : 90.80%	Missing Sensitivity, Specificity
Audio for Audio is Better? An Investigation on Transfer Learning Models for Heart Sound Classification	VGG-16/19, MobileNet v2, ResNet-18/50/101 and PANNs CNN16	PANNs model: UAR : 89.7% Sensitivity : 96.9% Specificity : 88.6%	Hyper-parameter tuning missing
Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification	Pretrained ResNet + Cross-validation	Accuracy : 90.80%	Other Evaluation metric missing
Automated Bird Species Identification using Audio Signal Processing and Neural Networks	AlexNet	Accuracy : 97%	Different models should be considered
Ensemble of convolutional neural networks to improve animal audio classification	Ensemble of six pretrained models	Superior Classification results	High computation required
Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling	VGG-16, ResNet50, ResNet5	Pseudo labelling	Different models should be considered
An Efficient Visual-Based Method for Classifying Instrumental Audio using Deep Learning	SFT + AlexNet	Accuracy : 73.7%	Only 7 classes considered
Instrument Classification using image based Transfer Learning	VGG	MFCC, Constant Q transform, Mel Spectrograms	Hyper-parameter tuning missing
PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition	Wavegram-LogMel-CNN	mAP : 0.439	Tested only on one dataset
Rethinking CNN Models for Audio Classification	DenseNet, ResNet, Inception	DenseNet, ResNet performed better	Hyper-parameter tuning, audio preprocessing missing
Scalogram Neural Network Activations with Machine Learning for Domestic Multi-channel Audio Classification	Six pretrained models	F1-score: AlexNet(DCASE) : 93.37% Xception(SINS) : 94.31%	Hyper-parameter tuning, audio preprocessing missing
An Investigation of Transfer Learning Mechanism for Acoustic Scene Classification	Dual Channels on AlexNet,VGGNet, ResNet	Accuracy : 77.8%	Hyper-parameter tuning, audio preprocessing missing
Listening to the World Improves Speech Command Recognition	DenseNet	Accuracy : 71.86%	Hyper-parameter tuning, audio preprocessing missing

3 Methodology

The methodology followed in this research was *CRoss Industry Standard Process for Data Mining* (CRISP-DM) as illustrated in the Fig.1. The methodology is proven to be robust providing a structured approach followed by many data scientists over the globe to find business solutions[10]. Secondly, since our main idea was to develop an algorithm to assist the sales and marketing team and to contribute to a speech development model, CRISP-DM was the most appropriate methodology to be followed. Lastly, the methodology is very flexible by allowing the tasks to be performed in different order of importance along with the ability to backtrack tasks and actions.

[10] `https://www.sv-europe.com/crisp-dm-methodology/`

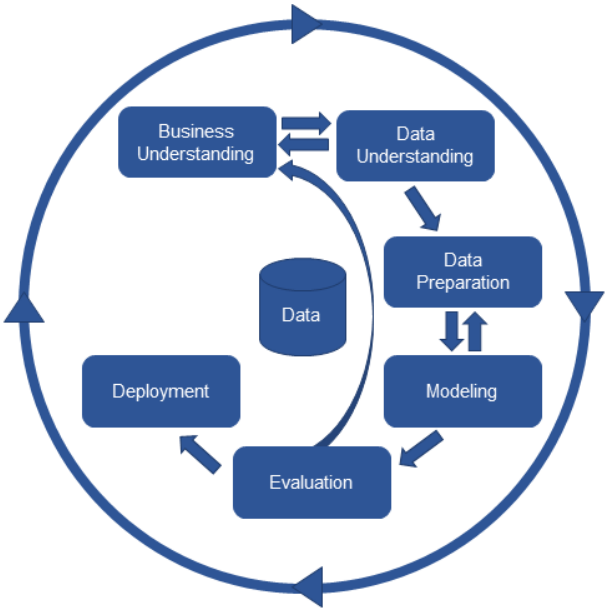

Figure 1: Methodology : CRISP-DM

3.1 Dataset

For the research, I have used Mozilla Common Voice[11] and Speech Accent dataset[12]. The datasets are freely available to anyone and distributed under the *Creative Commons License* for research purpose. Common Voice dataset contains 75,879 English language 5 seconds audio samples of people from different countries and nationalities whereas the Speech Accent Dataset contains 2,140 samples from 177 countries over 214 native languages. The Common Voice dataset is being continuously validated by many individuals worldwide to ensure authenticity of the uttered words. Both datasets contains audio recordings (.mp3 file format) along with meta-data file (.csv file format) containing the demographic distribution.

3.2 Data Preparation

The following steps were followed in the process of data preparation.

3.2.1 Dataset Cleaning

The datasets are first accessed and downloaded from the respective links. Both the datasets are checked for missing values in the metadata and the presence of files in the directory. I dropped the records if they contain missing values (NA and NaN).

3.2.2 Audio Conversion and Feature Extraction

Initially the audio were sample at 44,100 Hz and a bit rate of 128 kb/s. For easing computation, I down-sample the audio at 16,000 Hz. We used the pre-emphasis technique to enhance the audio signals. Further, Short-Time Fourier Transform (STFT) is obtained and the audio signals are plotted on Mel scale to get Mel-Frequency Cepstrum co-efficients

[11] `https://www.kaggle.com/mozillaorg/common-voice?select=cv-valid-dev.csv`
[12] `https://www.kaggle.com/rtatman/speech-accent-archive?select=recordings`

(MFCC) which essentially is the time power spectrum of the audio sample with 128 components. Lastly, I plotted the signal on the log scale as shown in Fig.2 and stored in the respective directories.

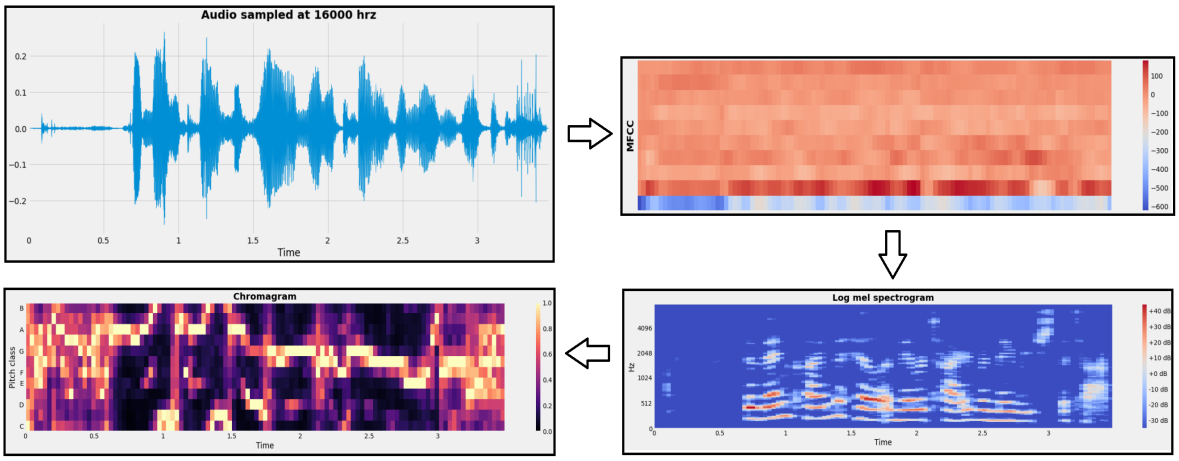

Figure 2: Audio Pre-Processing

3.2.3 Exploratory Data Analysis

In this part, we visualize the meta-data based on each subsets to get an idea on the audio sample distribution across each classes (Age, Gender and Nationality) and compare each datasets.

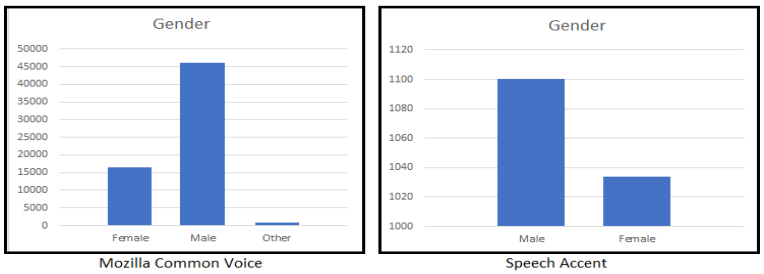

Figure 3: Gender Data Distribution

By grouping the dataset according to gender, it is evident that in the Common Voice Dataset, the Female population distribution is 26% compared to 76% of males. On the other hand, the gender distribution is almost equal in the speech accent dataset as shown in Fig.3.

Next, it was found that in the dataset, population with age group of twenties was the highest followed by thirties, forties wand so on as depicted in the Fig.4.

Lastly, on analyzing the region data in both the datasets, in both the datasets, most of the population belonged to the US, followerd by UK, India and Canada as shown in Fig.5

3.2.4 Model Preparation

Sample Database creation : For Classifying the age, gender and accent in the Common Voice dataset, first 500, 200 and 200 audio samples from respective the dataset are

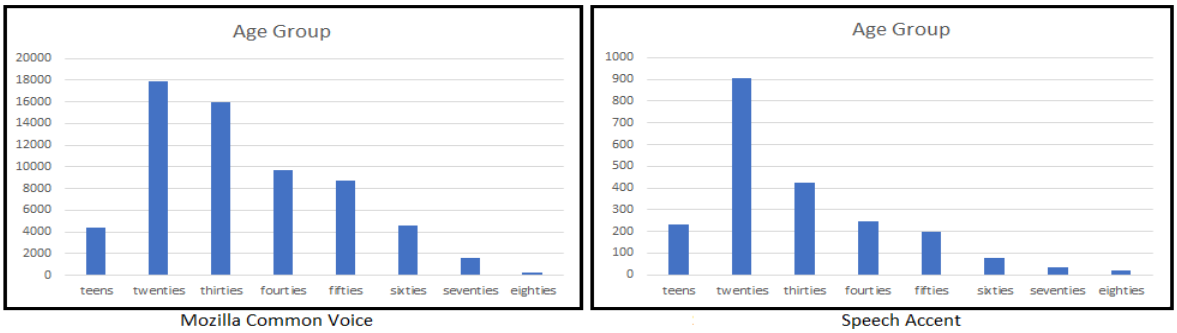

Figure 4: Age-Group Data Distribution

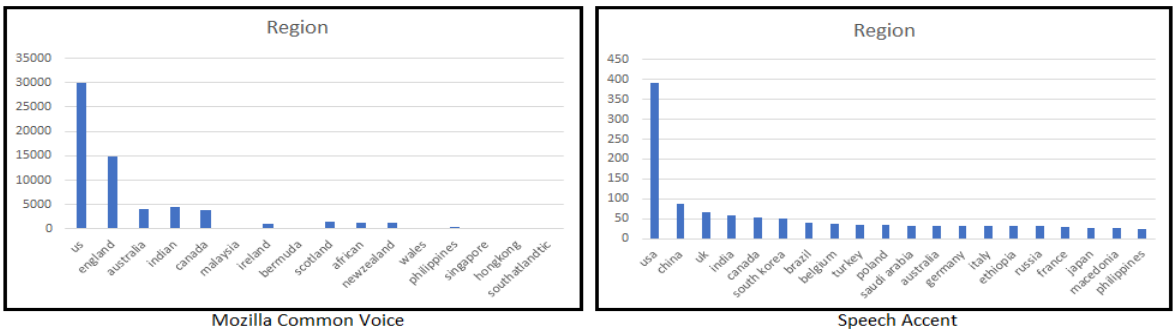

Figure 5: Region Data Distribution

sampled and extracted as an RGB image (128 x 128 resolution, 3) and stored in the respective directories. On the other hand for Speech Accent Dataset, I sampled all the audio and stored.

Extracting training features : The input features from each stored images are normalized and converted into Numpy arrays which act as an input to the CNN models. Lastly, the datasets are split into training and testing data in 80:20 ratio before using it as an input feature.

4 Design Specification

4.1 Design Framework

The outlined work flow diagram in Fig.6 depicts each steps followed to achive the respective classification. As shown in the first step the audio samples for each categories are processed and Log-Mel-Spectrograms are extracted. Next, these images are passed as vectors to the five pre-trained models (VGG-16/19, ResNet50, Incep-tionV3 and Xception). Simultaneously, the predicted outputs from the stacked classifiers are passed as input to the logistic regression model (model stacking). Finally, the outputs from each individual pre-trained model and the stacked model are evaluated.

4.2 Pre-trained Model Architecture

For the audio signal classification, I have used three deep neural network (DNN) architectures based on transfer learning methodology namely, VGG-16/19, ResNet50, Incep-

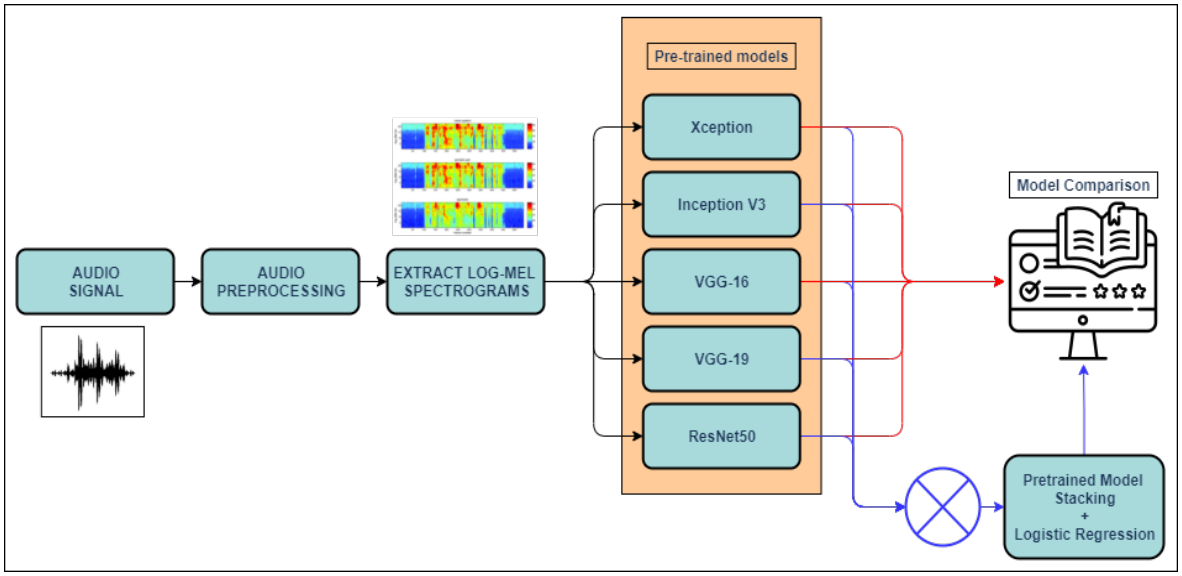

Figure 6: Design Framework

tionV3 and Xception whose architecture is explained in the following sections.

The pre-trained networks have the advantage of already being trained on millions of images with better techniques and hardware which resulted in superior results. Hence, weights of these models have been retained (freezed Imagenet weights) and after flattening the pretrained base model layer only dense layers with 'ReLu' activation function have been added. Finally, the last classification layer with 'softmax' activation is added with respect to the type and nature of the classifications (age, gender or nationality).

4.2.1 VGG-16 and VGG-19

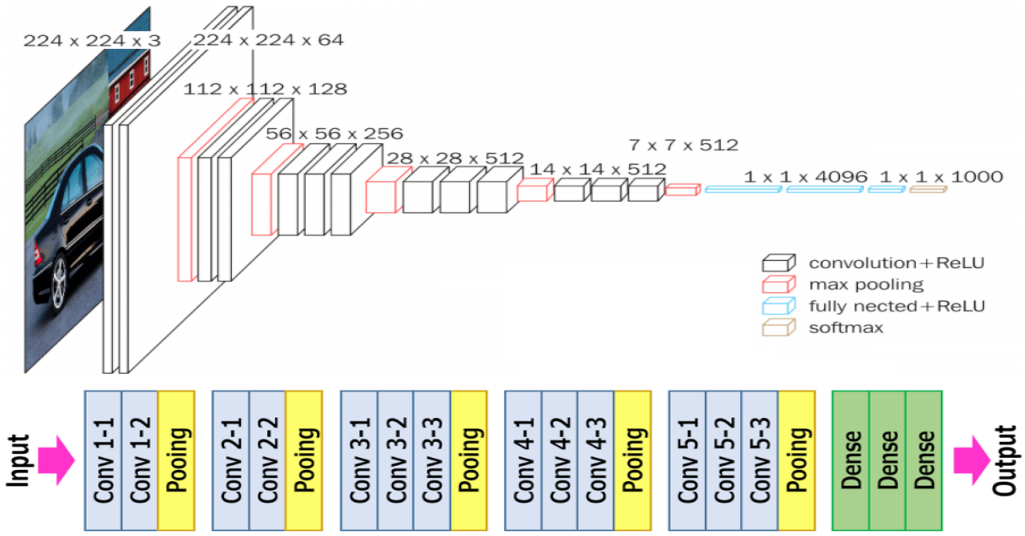

Figure 7: VGG-16 Architecture

VGG-16 is a deep neural convolution network architecture proposed by Simonyan and Zisserman (2015) in their research which secured 2nd place in the ILSVR classification task with an error rate of 6.8% in the year 2014. The model accepts an input of 224 x 224 RGB images. It is created using small 3 x 3 convolution layers throughout the

architecture and a stride of 1 pixel. The convolutions blocks are are padded by five max pooling layers (2 x 2 with stride 2)to reduce spatial resolution. Lastly, the pooling layers are connected to three fully connected layers out of which the first two layers have 4096 channels followed by the last layer having 1000 channels. The VGG-19 architecture is similar to VGG-16 and uses 1 extra convolution layers at blocks (total 3 CNN layers).

4.2.2 ResNet50

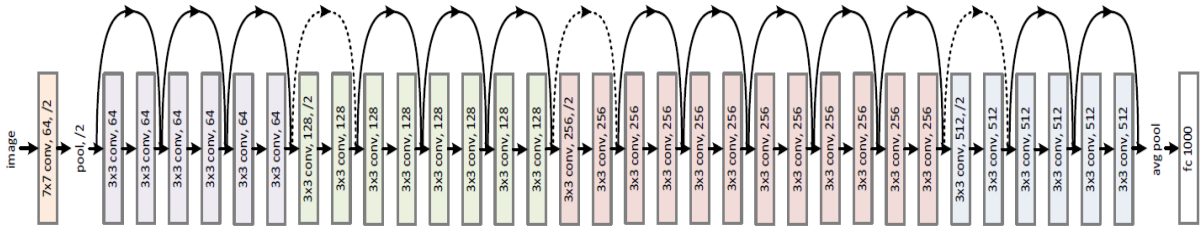

Figure 8: ResNet50 Architecture

The ResNet50 architecture presented by He et al. (2015) was the winner of 2015 ILSVRC and COCO competition with top-5 error rate of 3.57% which introduced a unique architecture of *Residual Blocks*. The addition of residual mappings addressed the limitations of vanishing gradient in deeper convolution networks. The architecture is made up of 50 deep layers as shown in Fig.8. The initial layer of network accepts inputs with multiple of 32 x 32 dimension images with RGB channel and has 7 x 7 and 3 x 3 convolution kernels followed deep convolution blocks with residual mapping. Lastly, the output layer contains global pooling followed by 1000 channel fully connected dense layer with 'softmax' activation.

4.2.3 Inception V3

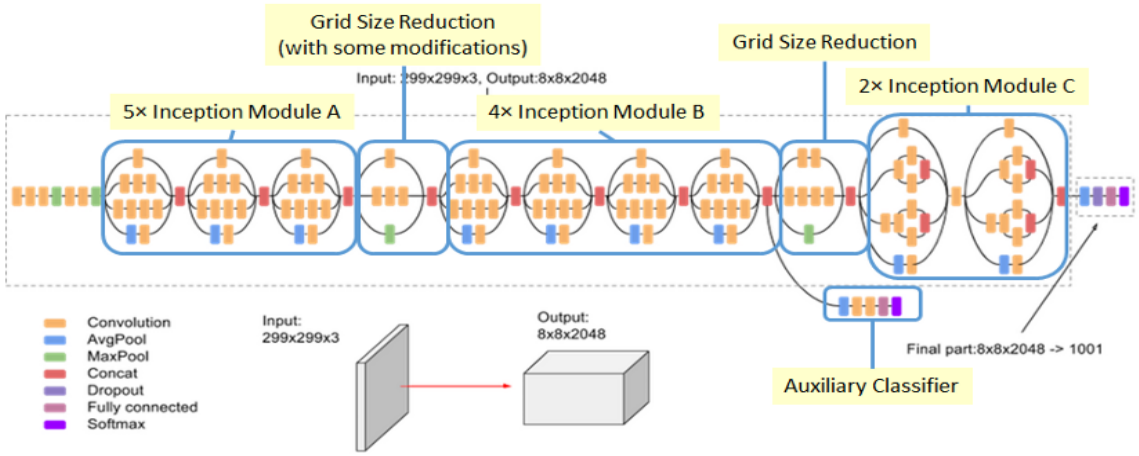

Figure 9: Inception V3 Architecture

Inception V3 model suggested by Szegedy et al. (2015) is an improvement over the previous versions, which mainly focuses on utilizing less computational power by factorizing convolutions. The model achieved superior results with top-5 error of 5.6% compared to the contemporary models. The architecture is made up of 9 inception modules stacked

linearly making upto 22 layers utilizing global average pooling at the end as shown in the Fig.9. Auxiliary classifiers are implemented to prevent the network from dying out using batch normalization.

4.2.4 Xception

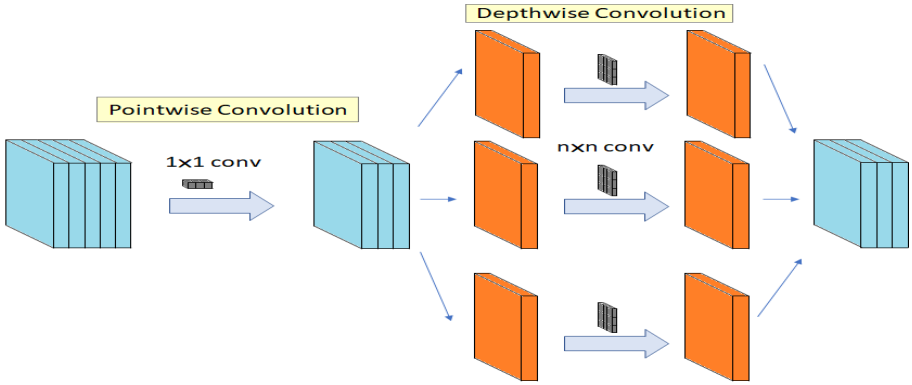

Figure 10: Xception Architecture

The Xception model was proposed by Chollet (2017) as an enhancement to the Inception V3 model by introducing point-wise and depth-wise convolution separation. The filters are applied on each depth followed by 1 x 1 convolution as shown in Fig.10. Unlike the Inception model, the Xception model does not include ReLU non-linearity after the first operation.

5 Implementation

5.1 Development Environment

The project was developed using Python 3.7.10 programming language on a Kaggle Notebook Environment (Session: 9 hours, Disk Space: 73 gb, RAM: 16gb and GPU: 13 gb) and Google Colab (GPU: Tesla P100-16GB, CPU: Intel Xeon 2.30GHz, Disk: 100 gb) in the cloud. The data is directly accessed in the repository uploaded on Kaggle. For reading Audio files and rendering the Mel Spectrograms, the Librosa package[13] in was used. For Data handling and wrangling and generating model features Numpy[14] and Pandas[15] library was used. The deep learning models were accessed run using the Keras [16]I and TensorFlow[17] API. Finally, for exploratory data analysis as well as to evaluate results, scikit-learn[18], Matplotlib[19] and Seaborn[20] were used.

[13] https://librosa.org/doc/latest/index.html
[14] https://numpy.org/
[15] https://pandas.pydata.org/
[16] https://keras.io/api/applications/
[17] https://www.tensorflow.org/
[18] https://scikit-learn.org/stable/
[19] https://matplotlib.org/
[20] https://seaborn.pydata.org/

5.2 Data Handling or Data Wrangling

Using the meta-data, I analyze the datasets and data-distribution according to each category. I performed exploratory data analysis and plot the count distributions and analyze the audio signals as depicted in Fig. 2, 4, 3 and 5. Using the filenames in the meta-data and the type of classification I first create 3 folders according to the age, gender and nationality classification. For the Speech accent data, we sample all the data in the root folder, where as for the Mozilla common voice dataset, we sample 500, 500 and 200 samples for age, gender and nationality classification respectively.

5.3 Model Implementation

Model implementation is started by defining custom functions for data preprocessing, model architectures, model stacking and evaluation. Next, constants are defined for RGB image as 3, image size to 128 x 128, epochs to 10, initial learning rate to 0.001 and Train:Test split to 80:20. Next, dense architecture which will be connected as a top layer for all the pretrained models is defined as shown in Fig.11 where an architecture using VGG-16 is shown.

```
def dense_model(base_model,num_classes):
    model = Sequential()
    model.add(base_model)
    model.add(Flatten())
    model.add(BatchNormalization())
    model.add(Dense(512, activation='relu'))
    model.add(Dropout(0.5))
    model.add(BatchNormalization())
    model.add(Dense(256, activation='relu'))
    model.add(Dropout(0.5))
    model.add(BatchNormalization())
    model.add(Dense(128, activation='relu'))
    model.add(BatchNormalization())
    model.add(Dense(num_classes, activation='softmax'))

    return model
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg16 (Functional)           (None, 7, 7, 512)         14714688
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
batch_normalization (BatchNo (None, 25088)             100352
_________________________________________________________________
dense (Dense)                (None, 1024)              25691136
_________________________________________________________________
dropout (Dropout)            (None, 1024)              0
_________________________________________________________________
batch_normalization_1 (Batch (None, 1024)              4096
_________________________________________________________________
dense_1 (Dense)              (None, 128)               131200
_________________________________________________________________
batch_normalization_2 (Batch (None, 128)               512
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 387
=================================================================
Total params: 40,642,371
Trainable params: 25,875,203
Non-trainable params: 14,767,168
```

Figure 11: VGG-16 Model Architecture in Python

In the next step, set include_top=False, meaning the top layer of the model will not be included, instead the dense layers are added with respective output layer as the number of categories to predict. Weights of the pretrained models are initiated to *imagenet*, since I will be using predefined weights from the ImageNet competition which secured state-of-the-art performance.

In the next step, the models are compiled using Adam optimizer and the the learning rate is initialized. The decay is set to (learning rate / Epoch) for a gradual gradient descent while training the models. Further, the models are trained and the respective evaluation metric are displayed using inbuilt Keras library functions. The training and loss metrics are carefully noted and checked for model over-fitting as shown in Fig.12. Lastly, the stacked model classifier function is called which utilizes Logistic regression to train the pretrained model outputs. Finally, the models are compared using and selected based on the best evaluation metric.

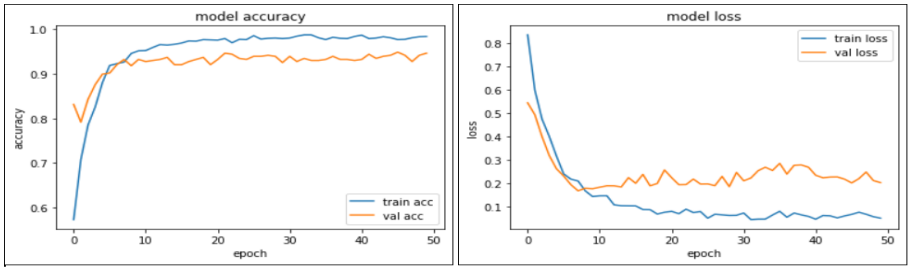

Figure 12: VGG-16 Model training for Gender Classification

6 Evaluation

Based on the objectives of the research, the results are evaluated in 2 dimensions, i.e based on the classification type and the selection of dataset. The validation accuracy graphs for all the pretrained models are depicted in Figs. 13, 14 and 15.

6.1 Experiment 1: Gender Classification

In the first experiment, I tested the models based on gender classification. Considering the Common Voice dataset, ResNet50 achieved the best accuracy (62%) whereas the Xception model achieved the least loss (0.68%) on the test dataset as shown in Table 3. On the other hand, using the Speech Accent dataset, the stacked model performed the best with 95% accuracy followed by ResNet50 (94% accuracy). Also, the V66-19 model achieved the least loss (0.17) off all the models.

Table 3: Evaluation of Gender-Based Classification

Mozilla Common Voice

Model	Loss	Accuracy	Recall
Xception	0.68	0.59	0.59
InceptionV3	0.97	0.55	0.51
VGG16	2	0.6	0.55
VGG19	0.91	0.58	0.56
ResNet50	1.08	0.62	0.62
Stacked Model	-	0.48	0.48

Speech Accent Dataset

Model	Loss	Accuracy	Recall
Xception	0.25	0.91	0.92
InceptionV3	0.22	0.91	0.91
VGG16	0.18	0.93	0.93
VGG19	0.174	0.93	0.94
ResNet50	0.18	0.94	0.94
Stacked Model	-	0.95	0.95

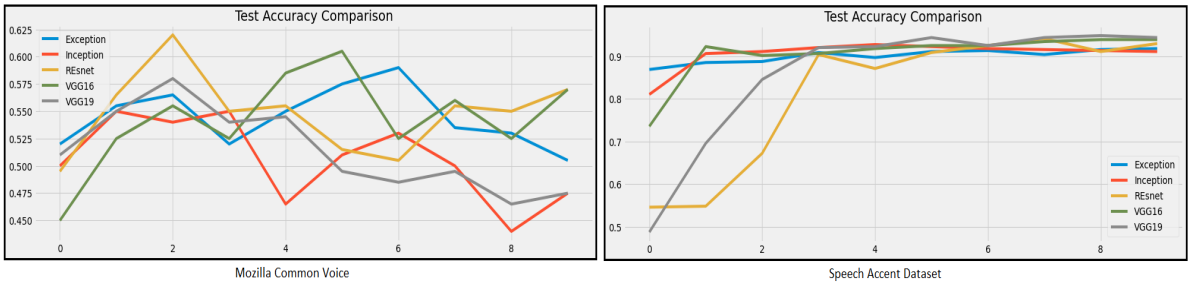

Figure 13: Model Accuracy Evaluation: Based on Gender

6.2 Experiment 2: Age-Group Classification

In the second part of the research, the pretrained models were trained to predict the age-group of individuals. On the common voice dataset, VGG 16 was able to achieve only 22% accuracy, where as pm the Speech accent dataset, VGG-16 achieved 52% accuracy. Similarly, the Inception V3 model was able to achieve the lowest loss with both the datasets. Notably, the stacked model performed similarly to the vgg-16 model as shown in Table 4.

Table 4: Evaluation of Age-Group-Based Classification

Mozilla Common Voice

Model	Loss	Accuracy	Recall
Xception	2.26	0.18	0.39
InceptionV3	2.00	0.20	0.01
VGG16	2.33	0.22	0.08
VGG19	2.02	0.22	0.04
ResNet50	4.29	0.21	0.02
Stacked Model	-	0.14	0.14

Speech Accent Dataset

Model	Loss	Accuracy	Recall
Xception	1.00	0.45	0.10
InceptionV3	0.99	0.49	0.19
VGG16	0.97	0.52	0.23
VGG19	0.99	0.50	0.08
ResNet50	1.00	0.46	0.02
Stacked Model	-	0.52	0.52

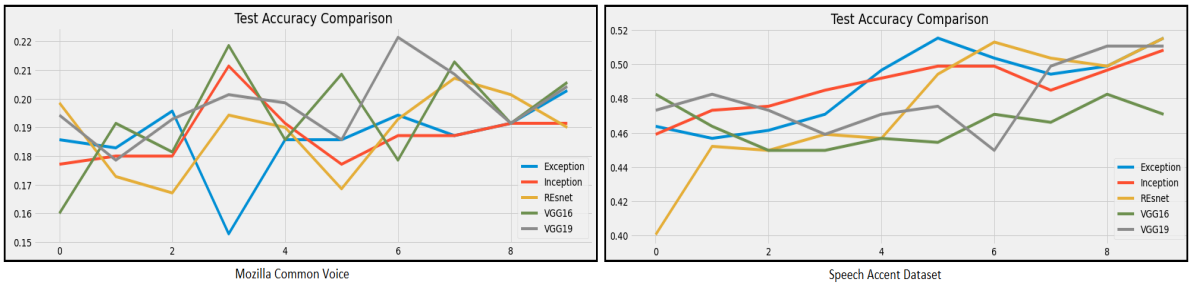

Figure 14: Model Accuracy Evaluation: Based on Age-Group

6.3 Experiment 3: Nationality or Region Classification

In the last experiment, the DCNN pretrained models are re-trained to recognize Nationality. As evident from Table 5, using the Common Voice dataset, the stacked model performed well with 48% accuracy where as the Inception V3 was able to achieve an accuracy of 46% using the Speech Accent Dataset.

Table 5: Evaluation of Nationality-Based Classification

Mozilla Common Voice

Model	Loss	Accuracy	Recall
Xception	1.71	0.43	0.15
InceptionV3	1.58	0.46	0.23
VGG16	1.61	0.47	0.18
VGG19	1.67	0.45	0.21
ResNet50	1.80	0.47	0.12
Stacked Model	-	0.48	0.48

Speech Accent Dataset

Model	Loss	Accuracy	Recall
Xception	1.56	0.33	0.18
InceptionV3	1.58	0.46	0.23
VGG16	1.53	0.36	0.20
VGG19	1.56	0.34	0.12
ResNet50	1.56	0.30	0.15
Stacked Model	-	0.34	0.34

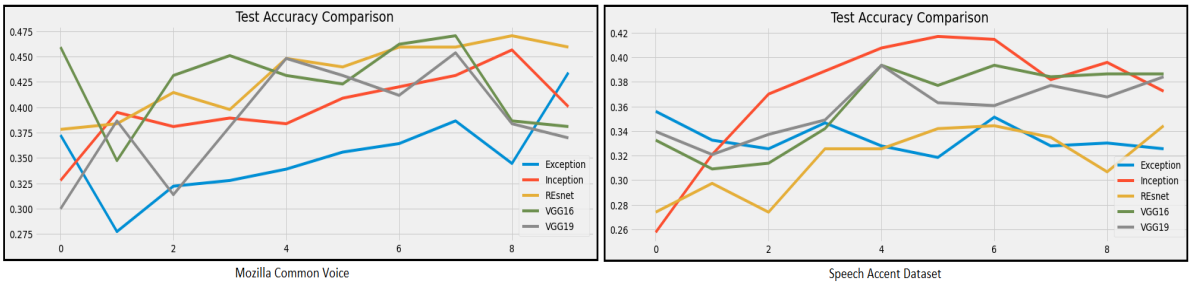

Figure 15: Model Accuracy Evaluation: Based on Nationality or Region

6.4 Discussion

From the results in section 6, it can be seen that, all pretrained models achieved different results based on different tasks. The proposed model was able to achieve superior results compared to the models proposed by Kuchebo et al. (2021) and Pandey (2020). Compared to Pandey (2020), where the author tried classifying the population in two age groups, the proposed research was aimed towards classifying the age-groups across decades. On changing the approach by categorizing the age groups into three categories (based on adolescence), the model was able to achieve better performance(46% accuracy). Lastly, for nationality or region classification, overall the Mozilla Common voice performed better with ResNet50.

The main reason for the differential performance, is due to the skewness and kurtosis of the dataset related to each categories. The data points have unequal distribution causing the models to underfit or overfit. The proposed model has the potential to perform better only if more data points are sampled from the new datasets, which was not feasible due to time constraints. On an average, the pretrained models were able to exceed the comparative results by 5 in gender classification and by 6% in age-group classification. Model stacking worked best with age and gender classification, while it was inefficient to classify nationality. The research was able to get superior results for gender classification, while age group and nationality classification remain averagely rated using this approach. Sampling more datapoints remains the key to decoding the classification problem of human voice based on age, gender and nationality.

7 Conclusion and Future Work

As stated in the Section 6.4, pretrained networks were able to achieve superior accuracy as the weights had been previously assigned using pre-defined architecture restating the hypothesis of usability of pretrained networks in human speech classification through audio. While more research is needed to derive better results, the proposed novel approach is satisfactory and can act as a baseline model for other researchers to contribute to the study. With an accuracy of 95%, 52% and 48% to classify gender, age and nationality, the proposed objective of the research has been successfully achieved. The classification performance was restricted due to the quality of the datasets used.

While Mozilla Common voice dataset is still being developed and validated, the proposed approach can be tested on new released versions of datasets to achieve better performance. To further widen the scope, the model can be tested on different languages accross regions on different datasets. Further, another approach implementing capsule networks for classification can be explored. Lastly, further work of the study will focus

to identify the best audio recording time needed to classify based on age, gender and nationality.

Acknowledgement

I would like to express my gratitude to my supervisor and mentor Dr. Majid Latifi and all the faculty members of NCI for imparting knowledge and guiding me though out the course of Data Analytics. Lastly, I would like thank my family and friends for their encouragement and support all though my studies.

References

Anupam, A., Mohan, N. J., Sahoo, S. and Chakraborty, S. (2021). Preliminary diagnosis of COVID-19 based on cough sounds using machine learning algorithms, *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, pp. 1391–1397.
URL: *https://ieeexplore.ieee.org/document/9432324/*

Atsavasirilert, K., Theeramunkong, T., Usanavasin, S., Rugchatjaroen, A., Boonkla, S., Karnjana, J., Keerativittayanun, S. and Okumura, M. (2019). A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms, *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, IEEE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/document/9045511/*

Bahari, M. H., McLaren, M., Van hamme, H. and van Leeuwen, D. A. (2014). Speaker age estimation using i-vectors, **34**: 99–108.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0952197614001018*

Balamurugan, A., Teo, S. G., Yang, J., Peng, Z., Xulei, Y. and Zeng, Z. (2019). ResH-Net: Spectrograms based efficient heart sounds classification using stacked residual networks, *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, IEEE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/document/8834578/*

Bhagwat, T., Deolalkar, S., Lokhande, J. and Ragha, L. (2020). Enhanced audio source separation and musical component analysis, *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*, IEEE, pp. 1–6.
URL: *https://ieeexplore.ieee.org/document/9358850/*

Chandrakala, S. and Jayalakshmi, S. L. (2020). Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition, **22**(1): 3–14.
URL: *https://ieeexplore.ieee.org/document/8752027/*

Chandu, B., Munikoti, A., Murthy, K. S., Murthy V., G. and Nagaraj, C. (2020). Automated bird species identification using audio signal processing and neural networks, *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*,

IEEE, pp. 1–5.
URL: *https://ieeexplore.ieee.org/document/9073584/*

Chi, Z., Li, Y. and Chen, C. (2019). Deep convolutional neural network combined with concatenated spectrogram for environmental sound classification, *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, IEEE, pp. 251–254.
URL: *https://ieeexplore.ieee.org/document/8962462/*

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions.
URL: *http://arxiv.org/abs/1610.02357*

Copiaco, A., Ritz, C., Fasciani, S. and Abdulaziz, N. (2019). Scalogram neural network activations with machine learning for domestic multi-channel audio classification, *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, pp. 1–6.
URL: *https://ieeexplore.ieee.org/document/9001814/*

Evangelista, E. B., Guajardo, F. and Ning, T. (2020). Classification of abnormal heart sounds with machine learning, *2020 15th IEEE International Conference on Signal Processing (ICSP)*, IEEE, pp. 285–288.
URL: *https://ieeexplore.ieee.org/document/9320916/*

Hall, J., O'Quinn, W. and Haddad, R. J. (2019). An efficient visual-based method for classifying instrumental audio using deep learning, *2019 SoutheastCon*, IEEE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/document/9020571/*

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep residual learning for image recognition.
URL: *http://arxiv.org/abs/1512.03385*

Hossain, M. F., Hasan, M. M., Ali, H., Sarker, M. R. K. R. and Hassan, M. T. (2020). A machine learning approach to recognize speakers region of the united kingdom from continuous speech based on accent classification, *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, IEEE, pp. 210–213.
URL: *https://ieeexplore.ieee.org/document/9393038/*

Kaya, H., Salah, A. A., Karpov, A., Frolova, O., Grigorev, A. and Lyakso, E. (2017). Emotion, age, and gender classification in children's speech by humans and machines, **46**: 268–283.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0885230816301346*

Koike, T., Qian, K., Kong, Q., Plumbley, M. D., Schuller, B. W. and Yamamoto, Y. (2020). Audio for audio is better? an investigation on transfer learning models for heart sound classification, *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, pp. 74–77.
URL: *https://ieeexplore.ieee.org/document/9175450/*

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W. and Plumbley, M. D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, **28**: 2880–2894.
URL: *https://ieeexplore.ieee.org/document/9229505/*

Kuchebo, A. V., Bazanov, V. V., Kondratev, I. and Kataeva, A. M. (2021). Convolution neural network efficiency research in gender and age classification from speech, *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, IEEE, pp. 2145–2149.
URL: *https://ieeexplore.ieee.org/document/9396365/*

Le, L., Kabir, A. N. M., Ji, C., Basodi, S. and Pan, Y. (2019). Using transfer learning, SVM, and ensemble classification to classify baby cries based on their spectrogram images, *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW)*, IEEE, pp. 106–110.
URL: *https://ieeexplore.ieee.org/document/9059502/*

M R, N. and Mohan B S, S. (2020). Music genre classification using spectrograms, *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, IEEE, pp. 1–5.
URL: *https://ieeexplore.ieee.org/document/9362364/*

McMahan, B. and Rao, D. (2017). Listening to the world improves speech command recognition.
URL: *http://arxiv.org/abs/1710.08377*

Muhammad, A. F., Susanto, D., Alimudin, A., Adila, F., Assidiqi, M. H. and Nabhan, S. (2020). Developing english conversation chatbot using dialogflow, *2020 International Electronics Symposium (IES)*, IEEE, pp. 468–475.
URL: *https://ieeexplore.ieee.org/document/9231659/*

Nanni, L., Costa, Y. M. G., Aguiar, R. L., Mangolin, R. B., Brahnam, S. and Silla, C. N. (2020). Ensemble of convolutional neural networks to improve animal audio classification, **2020**(1): 8.
URL: *https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-020-00175-3*

Palanisamy, K., Singhania, D. and Yao, A. (2020). Rethinking CNN models for audio classification.
URL: *http://arxiv.org/abs/2007.11154*

Pandey, S. (2020). Classification of human age group by implementing deep learning models on audio data, p. 24.

Preciado-Grijalva, A. and Brena, R. F. (2018). Speaker fluency level classification using machine learning techniques.
URL: *http://arxiv.org/abs/1808.10556*

Qawaqneh, Z., Mallouh, A. A. and Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification, **115**: 5–14.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0950705116303926*

Ramirez, A. D. P., de la Rosa Vargas, J. I., Valdez, R. R. and Becerra, A. (2018). A comparative between mel frequency cepstral coefficients (MFCC) and inverse mel frequency cepstral coefficients (IMFCC) features for an automatic bird species recognition system, *2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*,

IEEE, pp. 1–4.
URL: *https://ieeexplore.ieee.org/document/8625230/*

Shethwala, R., Pathar, S., Patel, T. and Barot, P. (2021). Transfer learning aided classification of lung sounds-wheezes and crackles, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, pp. 1260–1266.
URL: *https://ieeexplore.ieee.org/document/9418310/*

Shukla, U., Tiwari, U., Chawla, V. and Tiwari, S. (2020). Instrument classification using image based transfer learning, *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, IEEE, pp. 1–5.
URL: *https://ieeexplore.ieee.org/document/9277366/*

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.
URL: *http://arxiv.org/abs/1409.1556*

Sripriya, N., Poornima, S., Mohanavalli, S., Bhaiya, R. P. and Nikita, V. (2020). Speech-based virtual travel assistant for visually impaired, *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, IEEE, pp. 1–7.
URL: *https://ieeexplore.ieee.org/document/9315217/*

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2015). Rethinking the inception architecture for computer vision.
URL: *http://arxiv.org/abs/1512.00567*

Toffa, O. K. and Mignotte, M. (2020). Environmental sound classification using local binary pattern and audio features collaboration, pp. 1–1.
URL: *https://ieeexplore.ieee.org/document/9248620/*

Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R. and Noth, E. (2019). Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, pp. 717–720.
URL: *https://ieeexplore.ieee.org/document/8856972/*

Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J. P. and Aide, T. M. (2020). Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling, **166**: 107375.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0003682X20304795*

Zhou, H., Bai, X. and Du, J. (2018). An investigation of transfer learning mechanism for acoustic scene classification, *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, pp. 404–408.
URL: *https://ieeexplore.ieee.org/document/8706712/*