Nat. Volatiles & Essent. Oils, 2021; 8(5): 12261-12273

NVEO
Natural Volatiles & Essential Oils

A Comparative Study of Deep Learning and Machine Learning Approaches in Speech Emotion and Gender Recognition System

V G Nandan[1], Sukruth Shivakumar[2] , J Sangeetha[3] , Mukund Pandurang Nayak[4], Nishanth S K[5]

[1,2] UG Student, Department of CSE, M S Ramaiah Institute of Technology, Bangalore, India

[3] Associate Professor, Department of CSE, M S Ramaiah Institute of Technology, Bangalore, India

[4,5] UG Student, Department of CSE, M S Ramaiah Institute of Technology, Bangalore, India

Abstract

Emotions have a major impact on mental health and wellbeing of a person. This study focuses on the effective role that machine learning and deep learning approaches have in the early detection process of depression, thus preventing people from taking drastic measures. This can be done with the help of a speech emotion recognition system, through which one can identify and understand the emotional state of a person just by listening to them when they talk. In this work, audio datasets of Kannada and English languages are collected and classified into four categories: happiness, anger, neutral and sadness. For speech emotion recognition systems, the performance of Machine Learning algorithms is compared to deep learning algorithms when applied on both English and Kannada language datasets. Deep Learning algorithms show better accuracy. For speech gender recognition systems, a gaussian mixture model is used which gives satisfactory values for accuracy, precision and recall.

Keywords: Speech Emotion Recognition System, Speech Gender Recognition System, Gaussian Mixture Model, Random Forest, Convolutional Neural Network

1. Introduction and Literature Survey

Emotions play a tremendous role in day-to-day human interactions. They are very essential in making rational as well as intelligent decisions. By playing a vital role in shaping human social interaction, it indicates the mental state of a person. We can identify the emotions of others through speech. So, this is the motivation to consider speech signals as an effective way of understanding human emotions. Even if there has been a considerable improvement in speech recognition, computers are not yet capable of understanding the emotions of humans. Speech emotion recognition has many challenges like selecting the features that are the best and have enough power to differentiate between distinct human emotions, presence of several languages, accent, statements, speakers, variation in pitch, energy and a lot more. Speech emotion recognition systems have lots of benefits in today's applications. We come across a lot of areas where human-computer interaction is needed like customer service, medical analysis, speech synthesis, education etc. In today's world, where most of the meetings and calls are being held online or through phone calls, it is very difficult to understand the actual emotional condition of a person. We are capable of measuring and analyzing all aspects of physical health but we are not yet tracking and quantifying emotions to get a total understanding of our overall well being. It is very important to find out if people are depressed at the initial stages itself. The increasing rate of disability due to physical and mental health problems globally due to depression is very disturbing. Upon that, it is considered to be one of the primary reasons for suicide as well. The absence of help for treating this in the early stages is an alarming problem.

The study done in 2013 [1] the authors propose a system that recognizes the emotional state of a person by using audio signals. The emotions recognized are happiness, anger, sadness, fear, disgust, boredom, and neutrality. They perform Gender Recognition (GR) by using the pitch frequency estimation method before detecting emotion by SVM. The study done in [2-3] discusses how various approaches to the task of recognizing emotions are compared and an effective solution is proposed based on the combination of all

these approaches. The features explored in this study include modulation Mel-Frequency Cepstral Coefficients (MFCC), spectral features, pitch, linear prediction cepstral coefficient and energy. The limitations and performance of speech emotion recognition systems are also addressed. Few studies discuss extensively about the application of Machine Learning algorithms for the purpose of speech emotion recognition. In [4-6], the focus is on identifying human emotion in speech signals and distinguishing them into seven major classes which are disgust, anger, happy, sad, neutral, boredom and anxiety. This paper uses a method which mainly relies on the energy of speech signals and MFCC. Using this method, feature vectors will be extracted and those vectors will be sent to classification algorithms like Random Forest, Gradient Boosting, Support Vector Machine and K-Nearest Neighbors. The feature extraction and classifier training are the two main steps which are necessary for emotion recognition. The effect of increasing the number of features given to the classifier is also explained.

In [7], the study presents a real-time speech emotion recognition system that takes recorded speech as input, extracts 34 audio features and gives emotion as output. Emotions tested here are happiness, sadness, anger, and neutral. They concluded that while testing with their databases (RAVEDESS and SAVEE) SVM performed best and for live speech, gradient boosting performed best. In [8], the author has proposed a system that uses the support vector machine algorithm and recognizes the emotional state of the person from audio signals.

The studies in [9-12] discuss the application of deep learning for speech emotion recognition. Apart from the traditional methods, deep learning algorithms can also be used as alternatives. An overview of deep learning algorithms which can be used to identify the emotion of a person is discussed. The usage of deep neural networks to obtain high-level features from the raw data is considered. These features are efficient for speech emotion recognition. For each speech segment, an emotion state probability distribution is first produced by making use of DNNs. The authors have documented the development of speech emotion recognition systems using CNN as well. In [13], the author proposes an emotion recognition algorithm that doesn't depend on perturbation, noise measures and other acoustic measures. Deep Learning algorithms which use raw speech signals have been used to determine supreme information. Also, an algorithm which blends gender information block with Residual Convolutional Neural Network has been put forward. In [14], authors have recorded the German language for speech emotion recognition. Six emotions were uttered by 10 authors (big four plus boredom and disgust). To recognize the emotion and its naturalness, a perception test has been done and those who passed the criteria have been selected. This database is kept open on the internet for anyone to use.

From the literature survey, it is clear that almost all studies have addressed Speech Emotion Recognition where the dataset considered is of the English language. The datasets of Kannada language, which is a regional language, has not yet been considered. So, we decided to include Kannada along with English in our research work. We will be implementing a Speech Emotion Recognition System using several Machine Learning and Deep Learning approaches. By implementing these approaches, it would be easier for everyone to understand the emotions of a person just by listening to their voice when they are talking. From there, the system can draw conclusions regarding the user's mental health. With the help of effective diagnosis of depression and subsequent treatment for the same, it would be very effective in alleviating the symptoms and decreasing suicide rates. We have also developed a Speech Gender Recognition System to identify the gender of the user when the user's voice is given as the input. These recognition systems work on both English and Kannada languages. This study analyzes the accuracy of the machine learning and deep learning

approaches which will be used to identify the emotion and gender through speech when English and Kannada language statements are given as the input.

This paper is organized as follows: Section 2 outlines the methodology adopted for the extraction of emotions, data preprocessing and augmentation as well as about the classification techniques. Section 3 presents the results that have been obtained from this study. Finally, Section 4 describes the conclusion of our study and findings as well as the learning limitations and potential future work.

2. The Methodology

This section elaborates the methodologies of Speech Gender Recognition System and Speech Emotion Recognition System. In the proposed approach, we formed some sentences which exhibit emotions like sad, anger, happy and neutral. We asked the speaker to speak those sentences with the respective emotions. The spoken sentences are recorded in the ".wav" format. The collected audio samples are used for the Speech Emotion Recognition System (SERS) and Speech Gender Recognition System (SGRS). In SGRS, Mel-Frequency Cepstral Coefficient (MFCC) features are extracted from the collected samples. Necessary preprocessing is performed on extracted features and the gender is recognized. In SERS, data augmentation is applied on the collected audio samples. Then MFCC feature extraction followed by preprocessing is performed. Finally, the emotion is recognized.

Dataset Preparation

The first phase of our work was dataset preparation. It was challenging to get the datasets since we had to manage to get the audio samples of both female and male in both English and Kannada languages. To work on the Speech Gender Recognition System, we collected 192 male and 214 female voice samples in English language and 136 male and 146 female voice samples in Kannada language. To work on the Speech Emotion Recognition System, 640 English samples and 450 Kannada samples of different emotions (i.e., anger, sad, happy and neutral) were collected. For each emotion, we had 160 English audio samples and 112 Kannada audio samples. After the data collection, data augmentation & data cleaning is applied.

Data Pre-processing and Data Augmentation

Oftentimes while working with complex tasks, it is very difficult to get a large volume of data to train the model. So, we apply different transformations to the existing data to synthesize new data. This process is known as Data Augmentation. After the data collection phase, we apply data augmentation to get synthesized data. The augmentation techniques which we have used in our work are addition of noise, stretch, shift, pitch, higher speed and lower speed.

After data augmentation we perform data cleaning in the next step. We perform data cleaning because raw data is often not pre-possessed and noisy containing redundant fields, missing values, outliers and inconsistent data. In this research work, we have handled missing values and outlier detection. We have replaced missing values with field mean values so that we do not lose any information. Identification of outliers was done by the Interquartile Range (IQR) method. The quartile of dataset divides the dataset into four major sections and each one of them will contain 25% of data. Then interquartile range is calculated as IQR = Q3 – Q1. The lower bound and upper bound will be marked and all the data points which lie outside that particular range are treated as outliers and removed for further calculation. Finally, we applied Z-score standardization to normalize the value.

$$LB = Q1 - 1.5(IQR) \quad \textbf{(1)}$$

$$UB = Q3 + 1.5(IQR) \quad \textbf{(2)}$$

From Equation (1) LB represents Lower Bound and IQR represents Interquartile Range. From Equation (2) UB represents Upper Bound and IQR represents Interquartile Range.

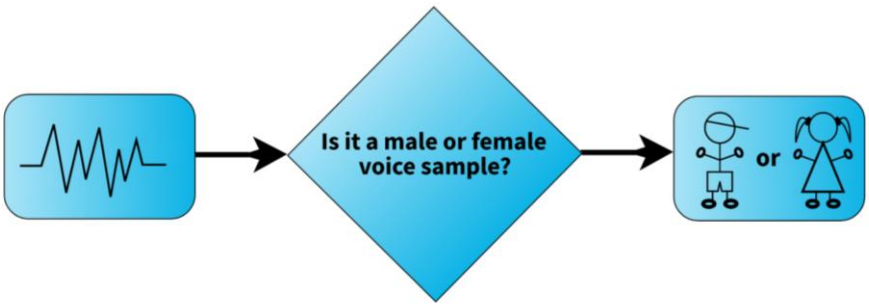
Feature Extraction for Gender Recognition and Emotion Recognition

There are 39 features [15] present in Mel-Frequency Cepstral Coefficients which we extract from the audio sample. The main reason behind extracting MFCC is because it has 12 parameters which are related to the amplitude of the frequency [15]. So, they provide enough frequency channels to analyze the audio. There are several python libraries that are present to extract them easily. The one which we have used in our work is "Librosa" and we set the sampling rate manually. After feature extraction, extracted features with suitable labels are concatenated.

Speech Gender Recognition System

Speech Gender Recognition System (SGRS) is the one which takes audio samples as input and predicts the gender of the sample as output. As shown in Figure 1, it takes the input in audio format and tells the gender of the audio as output. We have used Gaussian Mixture Model (GMM) for gender recognition.

Figure 1. Gender Recognition System



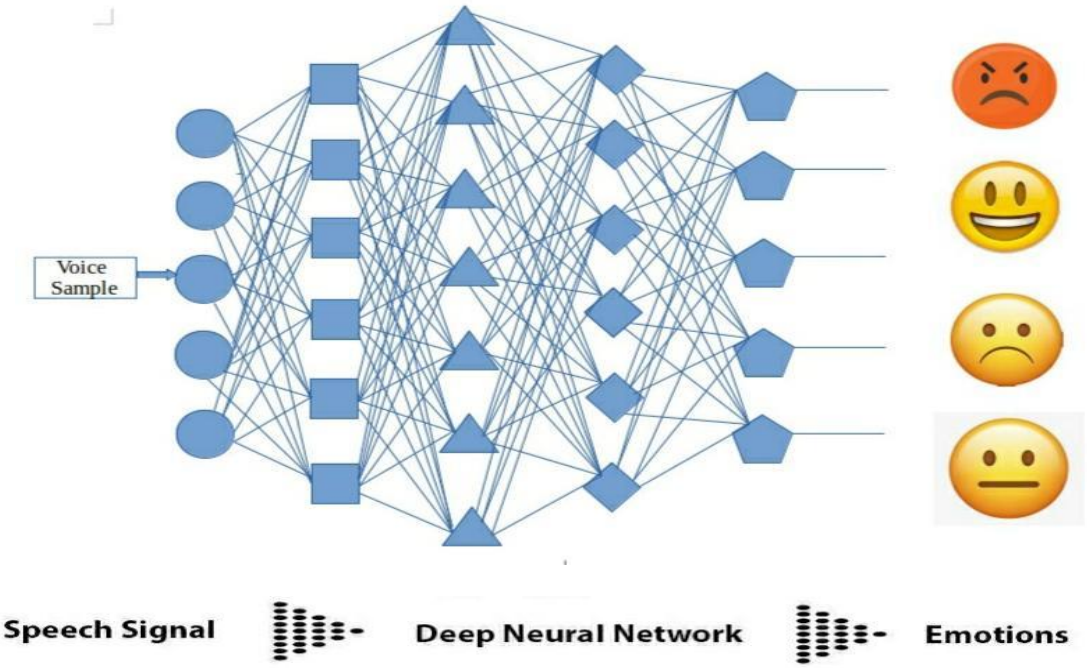
GMM is a powerful soft clustering algorithm. It uses a probabilistic model to distribute points in different clusters. This algorithm uses Expectation-Maximization (EM) technique to determine the values of parameters μ and σ^2. Where μ is mean and σ^2 is a variance matrix [16]. Probability density function of the Gaussian distribution is given by Equation (3).

$$f(x \mid \mu,\Sigma) = \frac{1}{\sqrt{2\Pi|\Sigma}} exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right] \quad \textbf{(3)}$$

Speech Emotion Recognition System

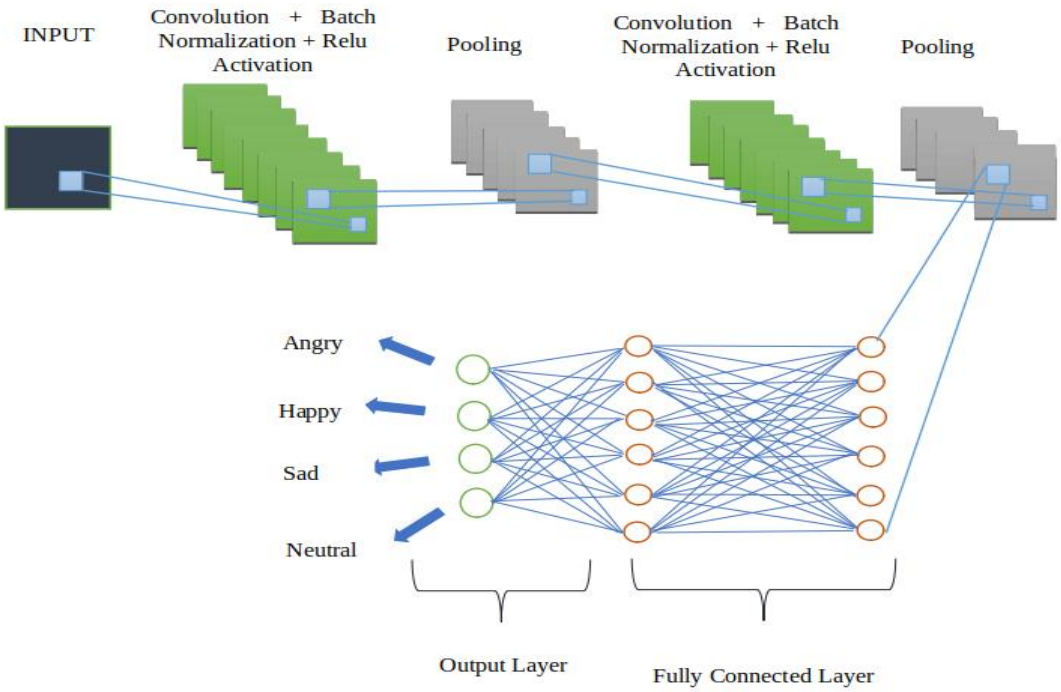
Speech Emotion Recognition System (SERS) is the one which takes a voice sample as input and predicts the emotion present in the voice sample. For example, as shown in Figure 2, a speech signal is sent to a deep neural network and it predicts the emotion present in the speech signal. In our work we are predicting four major emotions namely anger, happy, sad and neutral.

Figure 2. Speech Emotion Recognition System



In our work we are considering audio data in ".wav" format and we are sending them to five machine learning classifiers and two deep learning algorithms for emotion recognition. The machine learning classifiers which we have used in our work are K-Nearest Neighbor [17], Naive Bayes, Decision Tree [18], Random Forest [19] and Support Vector Machine [20]. The deep learning algorithms which we used in our work are Convolutional Neural Network and Multi-Layer Perceptron Classifier.

Figure 3. Convolutional Neural Network



A

Convolutional Neural Network (CNN) is a deep learning algorithm which is mostly used for visualizing images. When we feed an image the input, as shown in Figure 3, to the network it assigns importance to every aspect of that image in the form of weights and biases which are learnable parameters so that it can be able to

differentiate one image from others. It requires little to no pre-processing and can work well with unstructured data.

3. Results and Discussion

Students of today's generation are emotionally weak. They are impatient and reckless, leading them to take drastic measures even for small problems. To avoid that, teachers need to understand the mental behavior of students. Sometimes, teachers may face a situation where they have to face students virtually (a situation like the Covid-19 pandemic where all academic activities were held online). In such a situation we need to build a better understanding between teachers and students so that they can help the students in improving their mental health. To solve this problem, we have designed two recognition systems: Speech Gender Recognition System (SGRS) and Speech Emotion Recognition System (SERS). The main aim of our research is to help teachers and students so that the teachers can prevent students from taking drastic steps.

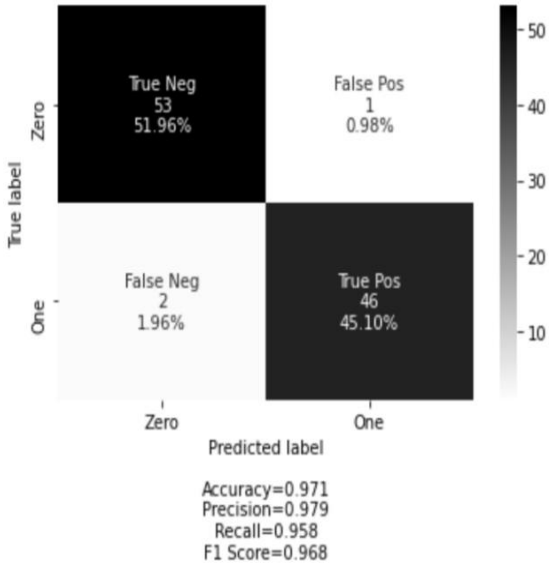
Speech Gender Recognition System

To identify the gender of a student, we have designed a Speech Gender Recognition System (SGRS). SGRS takes the voice data in ".wav" format and predicts the gender of the speaker. We have chosen Gaussian Mixture Model (GMM) as our machine learning classifier. Our work started with collecting audio samples in the English language. Then we performed Mel-Frequency Cepstral Coefficient (MFCC) feature extraction on collected audio files. Extracted MFCC features will be sent to GMM. GMM makes two clusters namely male and female. Then for each audio file, it calculates male score and female score. Once the MFCC features are extracted for each test audio file, male score and female scores will be calculated. These scores will be compared with female.gmm and male.gmm which already contains standard male score and female score on training data. If the female score of a file is greater than male score then it will be classified as female audio or vice versa.

English Language Dataset

The English test data set consists of 54 male samples and 48 female samples. When we tested these samples on the GMM model, we got an accuracy of 97.1% with 97.9% precision and 95.8% recall (Figure 4). Figure 4 shows that 53 samples of females have been classified as female and one sample of female has been wrongly classified as male and 46 samples of male have been correctly classified as male and 2 samples are wrongly classified as female on the English dataset.

Figure 4. Confusion matrix of gender recognition on English dataset



Kannada Language Dataset

Next, we have considered the same algorithm and followed the same procedure for the Kannada dataset. The Kannada test set consists of 50 male data and 50 female data. We have obtained an accuracy of 84% from the Kannada dataset with 85.4% precision and 82% recall (Figure 5). Out of those 100 samples 45 female samples 40 male samples have been classified correctly and 5 female samples 10 male samples have been wrongly classified.

Figure 5. Confusion matrix of gender recognition on Kannada dataset

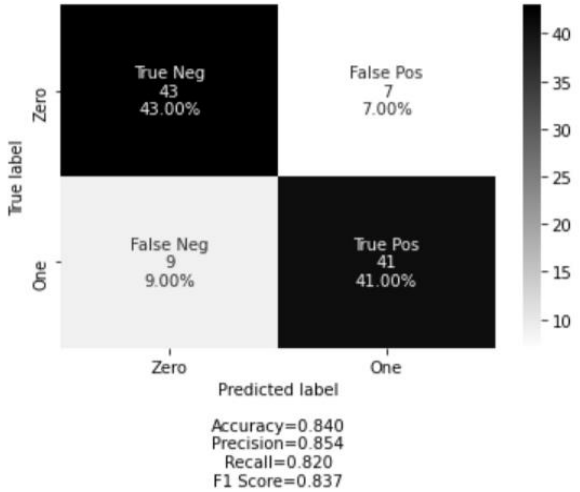


Figure 5 shows 43 samples of females have been classified as female and seven samples of females have been wrongly classified as male and 41 samples of male have been correctly classified as male and 9 samples are wrongly classified as female on the Kannada dataset.

GMM model achieved an accuracy of 97% on the English dataset and 84% accuracy on the Kannada dataset. This shows the GMM model on the English dataset performed better compared to the Kannada dataset. So, our choice of GMM algorithm for gender recognition over other classifiers gives us satisfactory results.

Speech Emotion Recognition System

We have designed a Speech Emotion Recognition System (SERS) which will predict the emotion of the speaker. We are testing this on multiple machine learning classifiers as well as deep learning techniques to see which algorithms perform better. For this purpose, a large amount of data is needed with proper data preprocessing. So, we collected around 640 audio samples in English language and 450 audio samples in Kannada language. Then we applied data augmentation techniques on each data sample. We chose six augmentation techniques for our audio data such as adding noise, stretch, shift, pitch, higher speed and lower speed. After the addition of augmented audio files, we ended up with around 4200 samples for the English dataset and 3150 samples for Kannada dataset which was sufficient for a deep learning model. Then we extracted MFCC features from each of these files. Data preprocessing is performed on extracted MFCC feature values. Finally, we applied Z-score transformation to get the data values in a common range.

Machine learning approach on English and Kannada Languages

After data preprocessing, we started applying machine learning algorithms on the English dataset and the Kannada dataset. The machine learning algorithms which we have considered for our research work are Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The performance of various machine learning algorithms on English and Kannada dataset are shown below.

Figure 6. Performance of Machine Learning Classifiers on English Dataset

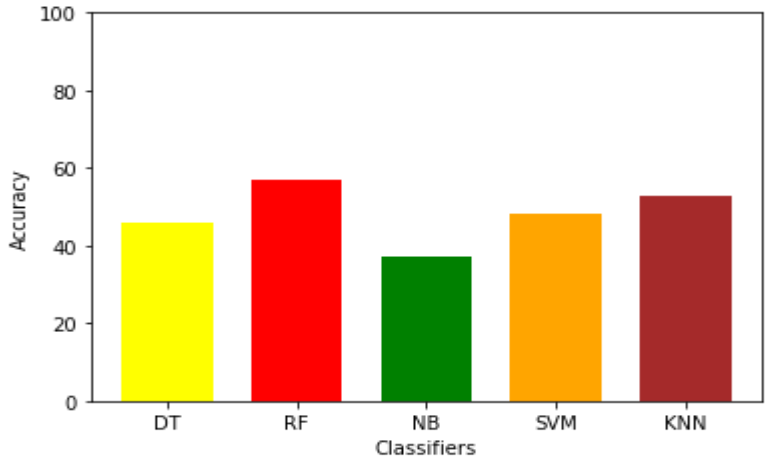
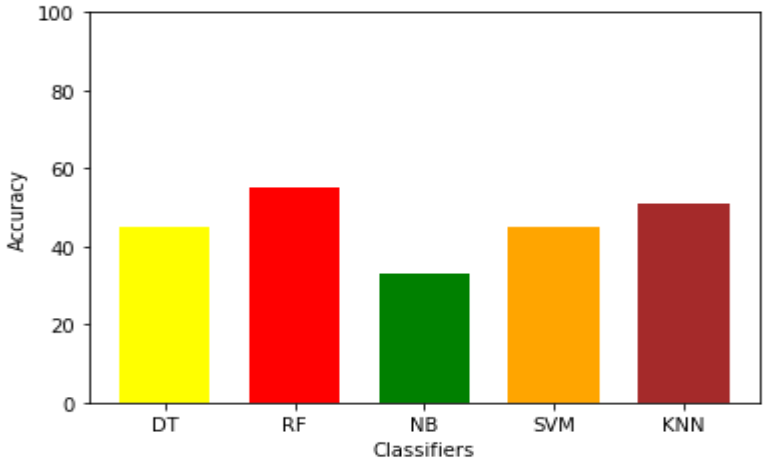


Figure 7. Performance of Machine Learning Classifiers on Kannada Dataset



We got an accuracy of 48% on SVM, 37% on NB, 57% on RF, 53% on KNN and 46% on DT for kannada language and an accuracy of 45% on SVM, 33% on NB, 51% on KNN classifier, 55% on RF and 45% on DT for kannada

language. Among all classifiers, the highest accuracy was achieved with Random Forest. This is because random forest, among all the classifiers, has the capability of handling large amounts of data and it is also capable of balancing datasets automatically. Since random forest takes the voting average of many decision trees, it will also reduce the high variance and achieve high accuracy.

Deep Learning approach on English and Kannada Languages

After applying our data samples on machine learning models, we tested our dataset on deep learning models. The deep learning algorithms which we have used in our work are one-dimensional Convolutional Neural Network (1D CNN), two-dimensional Convolutional Neural Network (2D CNN) and Multi-Layer Perceptron (MLP) classifier. The data cleaning and preprocessing steps are applied for deep learning as well. Once we were done with the English dataset, we tested that on the Kannada dataset. The architectures of 1D CNN and 2D CNN are shown below.

Figure 8(a). 1D CNN Architecture

```
model.summary()

Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d (Conv1D)              (None, 216, 256)          2304
_________________________________________________________________
activation (Activation)      (None, 216, 256)          0
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 216, 256)          524544
_________________________________________________________________
batch_normalization (BatchNo (None, 216, 256)          1024
_________________________________________________________________
activation_1 (Activation)    (None, 216, 256)          0
_________________________________________________________________
dropout (Dropout)            (None, 216, 256)          0
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 27, 256)           0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 27, 128)           262272
_________________________________________________________________
activation_2 (Activation)    (None, 27, 128)           0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 27, 128)           131200
_________________________________________________________________
activation_3 (Activation)    (None, 27, 128)           0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 27, 128)           131200
_________________________________________________________________
activation_4 (Activation)    (None, 27, 128)           0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 27, 128)           131200
_________________________________________________________________
batch_normalization_1 (Batch (None, 27, 128)           512
_________________________________________________________________
activation_5 (Activation)    (None, 27, 128)           0
_________________________________________________________________
dropout_1 (Dropout)          (None, 27, 128)           0
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 3, 128)            0
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 3, 64)             65600
_________________________________________________________________
activation_6 (Activation)    (None, 3, 64)             0
_________________________________________________________________
conv1d_7 (Conv1D)            (None, 3, 64)             32832
_________________________________________________________________
activation_7 (Activation)    (None, 3, 64)             0
_________________________________________________________________
flatten (Flatten)            (None, 192)               0
_________________________________________________________________
dense (Dense)                (None, 4)                 772
_________________________________________________________________
activation_8 (Activation)    (None, 4)                 0
=================================================================
Total params: 1,283,460
Trainable params: 1,282,692
Non-trainable params: 768
```

Figure 8(b). 2D CNN Architecture

```
model.summary()

dropout (Dropout)            (None, 15, 108, 32)       0
conv2d_1 (Conv2D)            (None, 15, 108, 32)       40992
batch_normalization_1 (Batc  (None, 15, 108, 32)       128
hNormalization)
activation_1 (Activation)    (None, 15, 108, 32)       0
max_pooling2d_1 (MaxPooling  (None, 7, 54, 32)         0
2D)
dropout_1 (Dropout)          (None, 7, 54, 32)         0
conv2d_2 (Conv2D)            (None, 7, 54, 32)         40992
batch_normalization_2 (Batc  (None, 7, 54, 32)         128
hNormalization)
activation_2 (Activation)    (None, 7, 54, 32)         0
max_pooling2d_2 (MaxPooling  (None, 3, 27, 32)         0
2D)
dropout_2 (Dropout)          (None, 3, 27, 32)         0
conv2d_3 (Conv2D)            (None, 3, 27, 32)         40992
batch_normalization_3 (Batc  (None, 3, 27, 32)         128
hNormalization)
activation_3 (Activation)    (None, 3, 27, 32)         0
max_pooling2d_3 (MaxPooling  (None, 1, 13, 32)         0
2D)
dropout_3 (Dropout)          (None, 1, 13, 32)         0
flatten (Flatten)            (None, 416)               0
dense (Dense)                (None, 64)                26688
dropout_4 (Dropout)          (None, 64)                0
batch_normalization_4 (Batc  (None, 64)                256
hNormalization)
activation_4 (Activation)    (None, 64)                0
dropout_5 (Dropout)          (None, 64)                0
dense_1 (Dense)              (None, 4)                 260
=================================================================
Total params: 152,004
Trainable params: 151,620
Non-trainable params: 384
```

The validation loss and accuracies achieved by 2D CNN on English and Kannada datasets are shown below.

Figure 9 (a). 2D CNN validation accuracy on English dataset

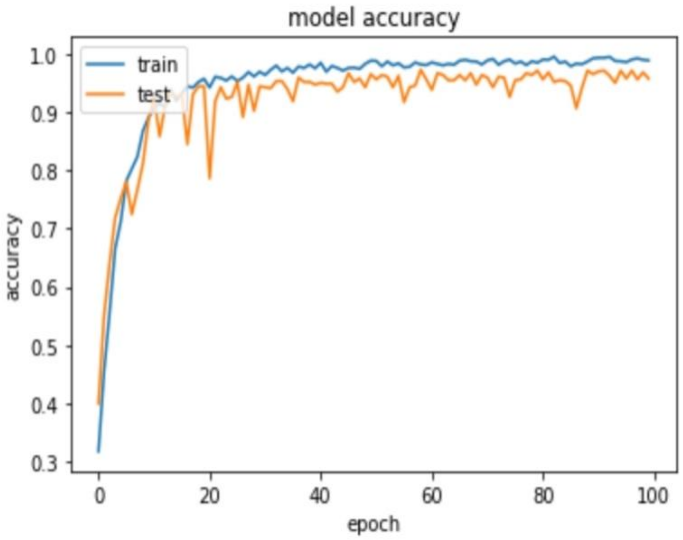


Figure9 (b). 2D CNN validation loss on Kannada dataset

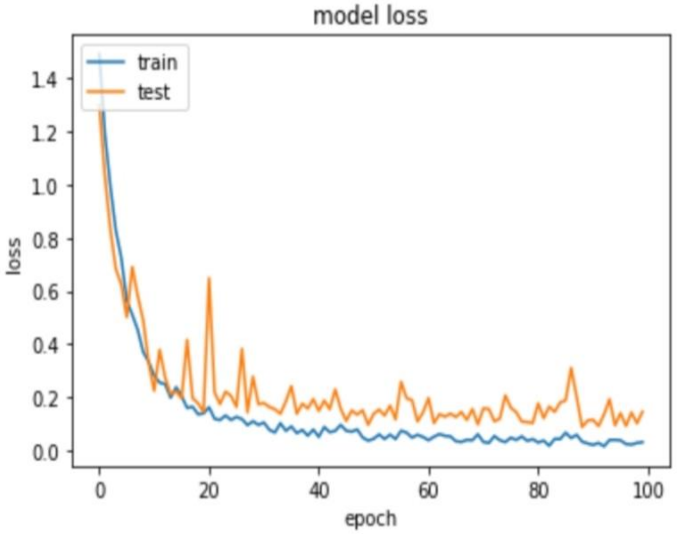


Figure 10 (a). 2D CNN validation accuracy on Kannada dataset

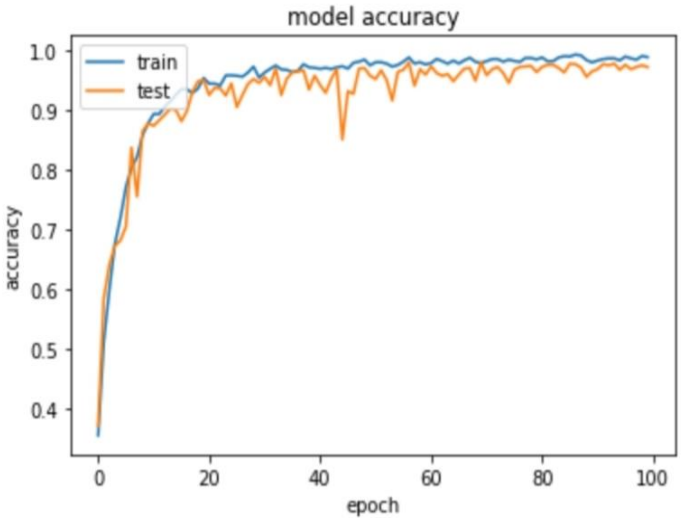


Figure 10 (b). 2D CNN validation loss on Kannada dataset

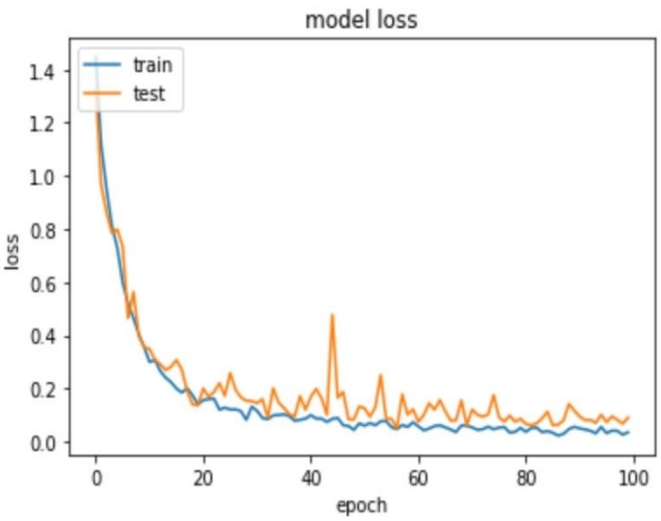


Figure 9 (a) shows the validation accuracy of 98.89% on training data and 97.24% on test data for the English dataset. Figure 10 (a) shows the validation accuracy of 98.02% on training data and 96.77% and on test data. Figure 9 (b) and 10 (b) shows the validation loss of 0.0908 and 0.1458 for English and Kannada dataset respectively. In both the cases validation loss is <1 which clearly indicates there is no overfitting of data. The performance of various deep learning algorithms on English and Kannada dataset are shown below.

Figure 11. Performance of Deep Learning Models on English Dataset

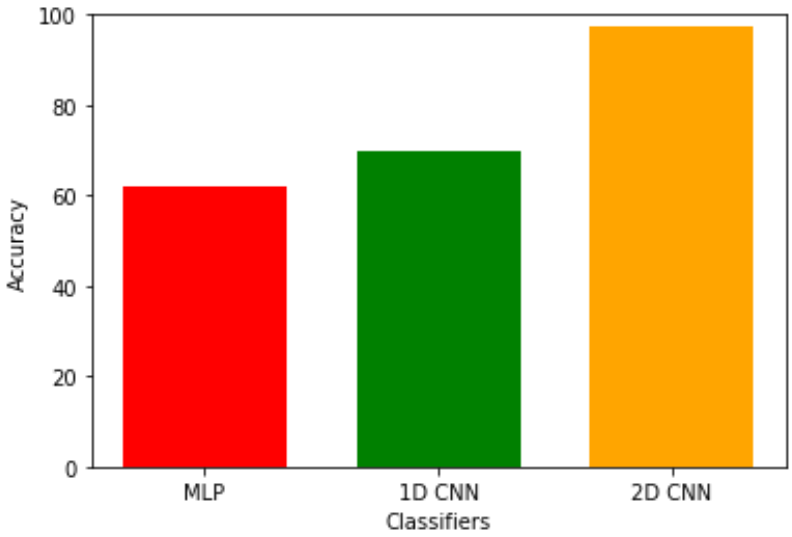
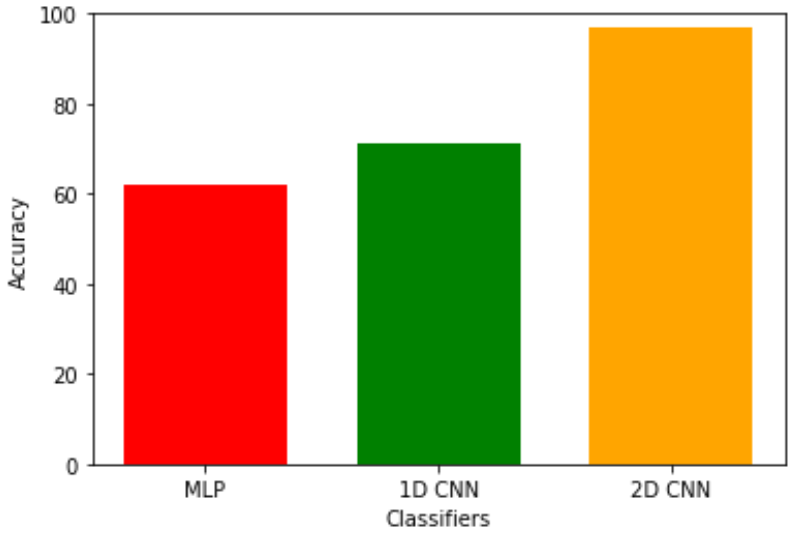


Figure 12. Performance of Deep Learning Models on Kannada Dataset



When we trained our model with a batch size of 16 on 2D CNN we achieved an accuracy of 97.24% for the English dataset. Along with this we achieved an improved accuracy of 70% on 1D CNN and 62% on MLP classifier. For the Kannada dataset, 2D CNN achieved an accuracy of 98.02% on training data and 96.77% on test data. Other models 1D CNN and MLP classifier achieved an accuracy of 70.51% and 61.84% respectively. From all the experiments, we observed that 2D CNN performs much better compared to all other models and also these models performed better on larger datasets compared to the smaller datasets.

4. Conclusion

In this study, we have focused on two types of recognition systems (i.e., Speech Gender Recognition System and Speech Emotion Recognition System). For the Speech Gender Recognition System using the GMM model, we were able to achieve an accuracy of 97% on the English dataset and 84% accuracy on the Kannada dataset. In the Speech Emotion Recognition System, we got lower accuracies with the machine learning approach for both English and Kannada languages. The two main reasons for lower accuracies are (i) a glitch during collection of dataset and (ii) while handling missing values, we were replacing missing values with column

mean which might be similar to other emotions. With the deep learning approach, we got better accuracy compared to machine learning algorithms. We noticed several things in this research work. Accuracy was improved by a significant amount when we performed data augmentation in deep learning but there was no much improvement in machine learning. This concludes that deep learning models are highly dependent on the amount of data but this is not true for machine learning classifiers. In the machine learning approach, we got highest accuracy on Random Forest and lowest accuracy on Naive Bayes classifier while other classifiers performed averagely on both the datasets. In the deep learning approach, highest accuracy was achieved by 2D CNN for both English and Kannada datasets.

REFERENCES

[1] Bisio, Igor, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone. "Gender-driven emotion recognition through speech signals for ambient intelligence applications." *IEEE transactions on Emerging topics in computing* 1, no. 2: 244-257, 2013.

[2] Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, and Mohamed Ali Mahjoub. "Speech Emotion Recognition: Methods and Cases Study." In *ICAART (2)*, pp. 175-182, 2018.

[3] Ingale, Ashish B., and D. S. Chaudhari. "Speech emotion recognition." *International Journal of Soft Computing and Engineering (IJSCE)* 2, no. 1: 235-238, 2012.

[4] Ghai, Mohan, Shamit Lal, Shivam Duggal, and Shrey Manik. "Emotion recognition on speech signals using machine learning." In *2017 international conference on big data analytics and computational intelligence (ICBDAC)*, pp. 34-39. IEEE, 2017.

[5] Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-5. IEEE, 2018.

[6] Deshmukh, Girija, Apurva Gaonkar, Gauri Golwalkar, and Sukanya Kulkarni. "Speech based emotion recognition using machine learning." In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 812-817. IEEE, 2019.

[7] Iqbal, Aseef, and Kakon Barua. "A real-time emotion recognition from speech using gradient boosting." In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1-5. IEEE, 2019.

[8] Kumar, S. Sravan, and T. RangaBabu. "Emotion and gender recognition of speech signals using SVM." *Emotion* 4, no. 3: 71, 2015.

[9] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech emotion recognition using deep learning techniques: A review." *IEEE Access* 7 (2019): 117327-117345.", 2019.

[10] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." In *Fifteenth annual conference of the international speech communication association*, 2014.

[11] Cheng, Huihui, and Xiaoyu Tang. "Speech emotion recognition based on interactive convolutional neural network." In *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 163-167. IEEE, 2020.

[12] Darshan, K. A., and Dr BN Veerappa. "Speech Emotion Recognition." *International Research Journal of Engineering and Technology,* 2020.

[13] Sun, Ting-Wei. "End-to-end speech emotion recognition with gender information." *IEEE Access* 8: 152423-152438, 2020.

[14] Burkhardt, Felix, Astrid Paeschke, Miriam Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. "A database of German emotional speech." In *Ninth european conference on speech communication and technology*, 2005.

[15] Hui Jonathan. "Speech recognition-feature extraction mfcc & plp.", https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9, 2019.

[16] Jing, Zhang, and Chen Xiao-mei. "A research of improved algorithm for GMM voiceprint recognition model." In *2016 Chinese Control and Decision Conference (CCDC)*, pp. 5560-5564. IEEE, 2016.

[17] Moldagulova, Aiman, and Rosnafisah Bte Sulaiman. "Using KNN algorithm for classification of textual documents." In *2017 8th International Conference on Information Technology (ICIT)*, pp. 665-671. IEEE, 2017.

[18] Navada, Arundhati, Aamir Nizam Ansari, Siddharth Patil, and Balwant A. Sonkamble. "Overview of use of decision tree algorithms in machine learning." In *2011 IEEE control and system graduate research colloquium*, pp. 37-42. IEEE, 2011.

[19] Patel, S. V., and Veena N. Jokhakar. "A random forest based machine learning approach for mild steel defect diagnosis." In *2016 IEEE international conference on computational intelligence and computing research (ICCIC)*, pp. 1-8. IEEE, 2016.

[20] Ghosh, Sourish, Anasuya Dasgupta, and Aleena Swetapadma. "A study on support vector machine based linear and non-linear pattern classification." In *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 24-28. IEEE, 2019.