Automatic Speech Emotion Recognition Using Machine Learning: Mental Health Use Case

Completed Research Paper

Samaneh Madanian
Auckland University of Technology
Auckland, New Zealand
Sam.madanian@aut.ac.nz

David Parry
Murdoch University
Perth, Western Australia, Australia
David.Parry@murdoch.edu.au

Olayinka Adeleye
Auckland University of Technology
Auckland, New Zealand
olayinka.adeleye@aut.ac.nz

Christian Poellabauer
Florida International University
Miami, FL, USA
cpoellab@fiu.edu

Farhaan Mirza
Auckland University of Technology
Auckland, New Zealand
Farhaan.mirza@aut.ac.nz

Shilpa Mathew
Auckland University of Technology
Auckland, New Zealand
wyc6987@autuni.ac.nz

Sandra Schneider
Saint Mary's College
Notre Dame, IN, USA
sandyl.schneider@gmail.com

Abstract

Human's emotional states affect their utterances which are generated through vocal cord vibrations. Accurate recognition of these emotional states encoded in human speech signals is critical and can be leveraged for mental health purposes. such as assisting practitioners in their assessments and decision-making, improving therapy effectiveness, safety monitoring of patients, and clinical training. Although there are existing works on speech emotion recognition, very few works address speech emotion recognition from a mental health perspective. This paper presents the results of our preliminary analysis that demonstrate the feasibility of automatic speech emotion recognition for mental health purposes. We used five machine learning paradigms for classifying emotions and evaluated their performance by focusing on their effectiveness in capturing human emotions using custom and benchmark databases, including TESS, EMO-DB, and RAVDESS. SVM demonstrated superior performance in overlapping settings based on F1-value and achieved 74% accuracy in RAVDESS and the custom datasets. We believe this research could be the initial step towards a fully implemented intelligent support service for mental health.

Keywords: Mental health, tele-mental health, speech analysis, automatic emotion recognition, machine learning.

Introduction

Emotions convey considerable information about the mental state of an individual and play a significant role in daily interpersonal human interactions. These mental states are typically named in senses such as emotions (e.g., fear, disgust, love) or perceptions (Oosterwijk et al. 2012) and they are essential for human rational and intelligent decision-making processes. Emotion recognition is a crucial aspect of mental healthcare and a core skill for any mental health practitioner (Minardi 2013). Furthermore, as mentioned by Copeland (2002), recording of emotional state may be useful in monitoring and treatment of mental illness. The awareness of the emotions of a mental health patient could provide insight into how to provide effective care for such a patient. However, it is difficult for mental health practitioners to manually recognise patient emotions. Practitioners mostly measure emotions based on either self-reports (Hasan et al. 2019) or subjective observations and decisions. This process of identifying human emotion requires high levels of skill and experience (Valstar et al. 2013) and remains dependent on the assessor's training and expertise.

In this regard, Automatic Emotion Recognition (AER), an aspect of affective computing (Picard 2000), has attracted significant interest. In this area, different modalities such as facial expressions (Jacintha et al. 2019), audio sounds (Schuller 2018), and physiological signals (Shu et al. 2018) have been explored for recognizing human emotions. Speech signals possess some intrinsic advantages that make them a preferred source for affective computing. When compared to facial expressions, which can be greatly altered by physical movement of the speakers or visual occlusions, due to glass or beard, speech qualities are rarely affected by these (Shah Fahad et al. 2021). Speech signals can be obtained more readily and economically than other biological signals, such as an electrocardiogram (ECG). Moreover, unlike facial emotion recognition, speech emotion recognition can use widely available recording technology, e.g., mobile phones, and it does not require full-face videos which people might not feel at ease, may be an issue in different cultural contexts, or may be difficult to capture, e.g., when masks are being used (such as in the case of COVID-19).

In addition, verbal communication is typically the first step in mental health assessment and monitoring. As speech signals carry valuable information, in parallel with other parameters, such as body language, automated systems that could process a patient's speech signals and give intelligent feedback could assist practitioners in their clinical assessments (Mitra et al. 2015). Therefore, most researchers in the affective computing domain are exploring automatic speech emotion recognition (ASER) solutions. ASER aims to recognise the underlying emotional states of a speaker from their speech signals (Schuller and Batliner, 1988); this is accomplished through the advancement of digital signal processing and machine learning (ML) paradigms. ASER can be used for various applications such as measuring learners' experience (Wang et al. 2018), identification of the emotional state of students in e-learning environments (El Hammoumi et al. 2018), or customer satisfaction (Ren and Quan 2012). ASER is being explored in supporting the assessment of patient depressive mood (Harati et al. 2018; Shinohara et al. 2021) and disorder (Cheng et al. 2020), stress (Lech and He 2014) and psychotherapy (Miner et al. 2020).

Although there are existing works on speech emotion recognition, only a handful of these works address speech emotion recognition from a mental health perspective. Considering the current health disparities in mental health treatment, redressing rather than intensifying equitable treatment across varied groups is necessary.

In this paper, we explore some ML techniques in speech emotion recognition for identifying a person's real emotion that could help mental health practitioners in their decision making, diagnosis and monitoring the treatment. The aim is to support psychiatrists and other mental practitioners in their clinical assessments to improve the quality of either in-person or remote mental health assessments. However, for such applications and purposes, lightweight and rapid AI support tools or applications are required. Therefore, the ML approach was chosen as it is potentially lightweight that makes its wider implementations more feasible and accurate, e.g. on smartphones.

We develop a proof-of-concept system that classifies and recognises emotions from speech based on different ML algorithms. Five conventional ML classifiers including Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), Gradient Boosting and Extra Trees are used for the recognition process; in the latter stages of the experiments. We also use the Bagging classifier. Then, their

performances are evaluated with a focus on how effective they are in capturing different human emotions using both custom and benchmark databases such as EMO-DB, RAVDESS, and TESS. Emotion corpora acquired from the Emo-DB database capture Angry, Boredom, Disgust, Fear, Happy, Neutral, and Sadness; the TESS database captures Angry, Excited, Frustrated, Happy, Neutral, and Sad, and the RAVDESS database captures Neutral, Calm, Happy, Sad, Angry, Fearful, Surprise and Disgust

Background and Related Works

Automatic identification and monitoring of mental health disorders such as depressive and neurological disorders from behavioural signals have been extensively studied in health science, psychology, and other related fields. Mental disorders such as anxiety, schizophrenia, and depression have been investigated and shown to have close ties with emotions (Chang et al. 2011). The emergence of AI paradigms and their applications in healthcare has transformed the operations of some health systems, making them more intelligent, especially in regulating human emotions and managing stress (Morris et al. 2010).

Understanding and recording the patient's emotional state is a key task in mental healthcare and it is a core skill for those who work in this field (Wilson and Carryer 2008). This task is vitally important as patients may have difficulty reporting their emotional state, especially in conditions such as schizophrenia (Tripoli et al. 2019), or patients may lack the ability to recognise their emotions, such as in alexithymia cases (Wilson and Carryer 2008). In many mild Traumatic Brain Injury (mTBI) cases or after traumatic events, children and adults may have a tougher time managing their emotions or are not able to communicate their feelings and emotions effectively to allow clinical assessment of their mental health. However, in mental health, measurement and assessment activities can be constrained or less reliable because of subjective observations and decisions (Valstar et al. 2013). To address this, there have been an increasing number of studies in emotion recognition in mental health. Various modalities have been explored for this purpose. However, capturing of affects can be quite challenging. For instance, sensors must be worn in the periphery of the body in some cases. Heart rate variability mainly captures stress and may be confounded by physical activity, facial, and other physiological gestures conveying rich emotions, however, they mostly require cameras pointing at the subject and sometimes may require real-time image analysis (Chang et al. 2011).

In contrast, vocal affect (emotional expression of speech) is easy to capture and has been shown to be an accurate modality for mental health evaluation. For instance, (Moore et al. 2004) showed 90% accuracy for depression from short-term speech data, and 70%-80% accuracy for a 2-way emotional classification for stress (Fernandez and Picard 2003). Existing work (Hall et al. 1995) in applied and preventive psychology showed that decreased verbal activity and monotonous/lifeless speech signal is indicative of depression or mental health-related issues. (Harati et al. 2018) indicated that there is a perceptible change in pitch, speech rate, loudness, energy, and articulation of mental health or depressed patients before and after treatment. Moreover, vocal affect and its relationship with the overall mood of mental health patients have been examined and validated in the affective computing domain (Stasak et al. 2016).

These developments motivate the research interest in speech emotion recognition for mental health. Speech emotion recognition system may support objective emotion detection to support practitioners in their decision making in early diagnosis and assessment stages and for diagnosis and monitoring of response to treatment in conditions when recognizing patients' emotions is important.

Speech Emotion Recognition in Mental Health

Most of the existing emotion recognition approaches are based on facial recognition and video data (Jacintha et al. 2019; Thome et al. 2016; Tripoli et al. 2019) or text analysis (Hasan et al. 2019) and only very few attempts have been made to use speech signals alone, especially in mental health. We discuss other related works in speech emotion recognition for mental health in this section.

The links between short-term emotions and long-term depressed mood states were leveraged to develop a predictive model using emotion-based features (Harati et al. 2018). The authors used auxiliary emotion datasets to train a deep neural network model, which was then applied to audio recordings of depression

disorder patients to find their low dimensional representation (LDR). The LDR is then used in the classification processes. Harati et al. (2018) research results indicate effective classification of depressed and improved phases of deep brain stimulation treatment. Similarly, Stasak et al. (2016) enhanced the classification of their automatic depression classification approach by 5% using speech emotion ratings. An emotional recognition system was developed by (Zisad et al. 2020) capable of recognising emotion from the speech of a neurological disorder patient using the RAVDESS dataset and a convolutional neural network (CNN) as the affective model. The model was able to correctly classify eight emotions including anger, fear, sadness, disgust, surprise, and happiness of a neurologically disordered patient. In (Chang et al. 2011), an affective and mental health monitoring library was developed for mobile applications. The library includes features for both mental health analysis and speech emotion recognition and can achieve up to 75% accuracy in two-class emotional classification of stress and neutral cases. In (Miner et al. 2020), the authors demonstrate that speech emotion recognition is feasible in psychotherapy by developing a HIPAA-compliant automatic speech recognition system that uses patient-therapist audio recordings collected in clinical trials. The emotion recognition solution shows a transcription word error rate of 25%, and a 83% prediction accuracy for depression-related utterances. We also acknowledged some studies that have attempted to use audio signals to identify speech characteristics, which can be used for diagnostics of mental health disorders such as depression (Mitra et al. 2015), or anxiety (Tsiakas et al. 2015). Despite all their potentials, these systems have some limitations: some are not designed for general use but just for the detection of a particular disorder and some used only a single algorithm for emotion detection that may affect the performance or the quality of the systems.

We also acknowledged some existing work in emotion recognition that have explored some of the ML algorithms we explored as baselines in this work. Tashev et al. (2017) proposed a system that aims to enhance user experience with combinations of a Gaussian Mixture Modelling (GMM) and neural network as feature extractors. SVM was used in (Fernandes et al. 2018) as a supervised learning approach for automatic emotion recognition. An improved version of the kNN algorithm is used for speech emotion recognition in (Feraru and Zbancioc 2013) that generated 62-67% accuracy for the Romanian language. SVM, RF, and Gradient Boosting were used for training ML models (Ghai et al. 2017). Although several works are available in the field of AER systems, these studies mostly refer to either overlapping or non-overlapping feature extraction or their emotion recognition part was constrained with two to three emotion categories, such as (Chatterjee et al. 2018). Also, very few efforts have been reported where the designed system used different future extraction techniques to measure their performance for AER. The classification accuracy is also low for current systems designed to detect more than two emotion types (Ooi et al. 2014).

Methods

In this research, we have treated speech emotion recognition as a classification problem, i.e., classifying speech data/signals to a particular class of emotion using algorithms. We explore five different ML paradigms for the classification process including SVM, KNN, RF, Gradient Boosting and Extra Trees for the recognition process. Figure 1 illustrates the flow of the recognition system. Raw audio datasets are first input into the models, then the data preprocessing is initiated which includes processes such as sampling, framing, augmentation, and normalization. After completion of the preprocessing, we extract the required features. For our experiment, we have used all the 34 features in the `pyaudioanalysis` API – see experiment section for details. We normalize the extracted features using 0-mean and 1-standard deviations.

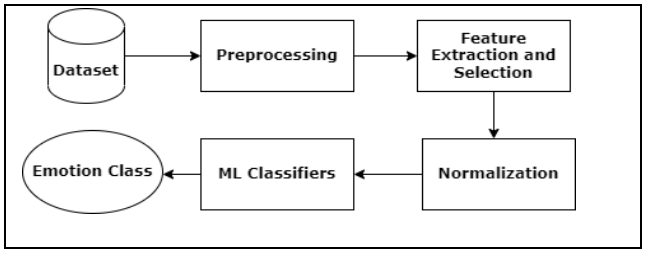

Figure 1. Recognition flow chart

The system is developed using five ML algorithms (Figure 2) for speech analysis of signals. The system we describe aims to automatically identify and recognize emotions from speech signals and classifies them into six basic emotional groups - anger, disgust, fear, happiness, pleasant surprise, sadness (Elfenbein and Ambady 2002), as well as neutral/no emotion. Using multiple algorithms can help in developing more effective systems as their comparison can enhance their applications based on mental health specific requirements and scenarios (Schuller 2018). In this work, we designed different models and compared the performance of different algorithms by a comparative study of overlapping and non-overlapping feature extraction and experiments. AER performance depends on three critical aspects: datasets, suitable speech features, and classification techniques to maximize the recognition accuracy of AER systems (Mustafa et al. 2018). The following section discusses how this work addressed these areas.

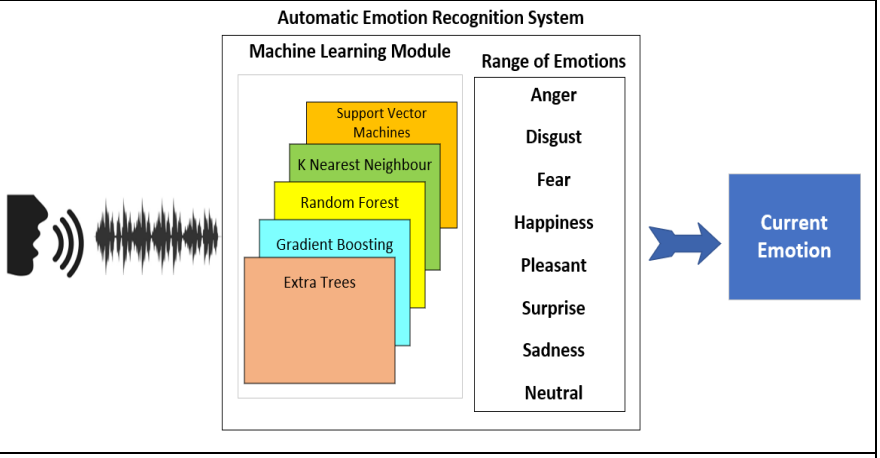

Figure 2. The overall structure of the proposed system

Experiments

This section presents the empirical performance of various classifiers used in the speech emotion recognition process, specifically using datasets downloaded from 4 different databases (detail discussed in data acquisition and pre-processing section) and our custom emotion dataset. We evaluate the speech emotion recognition process using selected emotions that are relevant to mental health. We run the experiments on an Anaconda environment with Python 3.7 on a PC with Intel i5-6500@3.2 GHz CPU, 16 GB RAM. Our objective is to answer the following research questions:

- How well do the classification models capture specific speech emotions relevant to mental healthiness? We do this by evaluating the emotion recognition accuracy of each model used in the experiment.
- Which of the acoustic features used in the experiment provide more accurate mental health speech emotion recognition? We do this by mapping specific emotion recognition accuracy given by each model to the acoustic features used in this work.
- We evaluate the impact that each model hyperparameters have on speech emotion recognition accuracy in both overlapping and non-overlapping cases. To do this, we use grid search to tune the hyperparameters.

We use different Python-based libraries for the experiment including `PyAudioAnalysis`, `Keras`, `Tensorflow`, `Numpy`, `Librosa`, `sklearn`, `Nlpaug` and `Matlibplot`. For the conventional ML models, we use `PyAudioAnalysis` with `sklearn` built-in functions to implement the models used in

the classification process. We also use `sklearn` for splitting our datasets into testing and training sets and generating the confusion matrix in the evaluation phase. `Tensorflow` was used to facilitate the backend of the system. For the deep-learning implementations, we use `Keras` for developing the three models that we base our SER implementation on. `Keras` provides some built-in functions such as layer definitions, optimizers and activation functions that enable easy implementation of the models. We employ `Numpy` for numerical analysis. *`PyAudioAnalysis`* enable us to load raw audio files and sample them with a specific sampling rate.

Data Acquisition

We use a custom dataset and datasets from 4 different databases. The databases and dataset attributes are described below:

- RAVDESS[1] (The Ryson Audio-Visual Database of Emotional Speech and Song). The dataset includes recordings of 24 male and female actors, vocalizing two lexically matched statements in a North American accent. The speech emotions captured in the data include fear, surprise, sad, anger, calm, happiness and disgust. This dataset has been highly rated for its emotional validity, reliability, and genuineness (Livingstone and Russo 2018). We acquired and used 1440 speech files with 1012 song files from the database in our experiment.
- TESS[2] (Toronto Emotional Speech Set modelled on the Northwestern University Auditory Test) (Dupuis and Pichora-Fuller 2010). We collected 2800 audio files from the TESS database following the recommendation of (Tin Lay et al. 2001). The dataset contains a set of 200 target words spoken by 2 actresses of ages 64 and 24 years. The speech emotions captured in the set include anger, fear, pleasant, surprise, happiness, sadness, disgust and neutral.
- EMO-DB[3] datasets contain utterances of males and females between ages 21-35. We collected 535 audio files spoken by 10 speakers (5 female and 5 male) capturing 7 acted emotions including anger, boredom, disgust, fear, happiness, sadness and neutral.
- Custom dataset: For testing the models, we took sample recordings of 15 candidates' utterances and we labelled the audio files with actors registered emotions as custom datasets. We converted the raw recording samples using 16kHz sample rate and mono channel method. We used the *`ffmpeg`*[4] tool to do the audio conversion.

We have opted for 'acted' databases because of their standard and the datasets are collected from experienced professional artists (Shah Fahad et al. 2021). They are also the commonly used emotional databases. The variation in the age groups can help us to develop a more generalized SER system, as the selected sample for the training model represents possible emotion states which can help avoid the case of selection bias.

[1] https://zenodo.org/record/1188976#.Yg8cEuhBw2w

[2] https://tspace.library.utoronto.ca/handle/1807/24487

[3] http://emodb.bilderbar.info/docu/

[4] https://ffmpeg.org

Data Preprocessing

Data preprocessing is the first step after data acquisition for training the classifiers. Some preprocessing techniques are used to extract features, while others are used to normalize the features so that variances in speakers and recordings do not affect the recognition process (Akçay and Oğuz 2020). All audio files used are sampled at rate 16kHz setting parameter sr (sample rate) to 16000 in the load function of the `PyAudioAnalysis` (Giannakopoulos 2015) and `librosa`[5] libraries used in our analysis. We explore these Python tools in facilitating feature extraction, segmentation, classification, and training models.Feature Selection

Feature selection is an important phase in emotion recognition processes, due to the high diversity in audio feature types. For feature selection, we concentrate on commonly used acoustic features proposed in similar studies, such as (Giannakopoulos 2015), or acoustic features with links to symptoms of mental health disorders such as Mel Frequency Cepstral Coefficients (MFCCs), and abrupt changes in the energy level of a speech signal (Cummins et al. 2015).

This phase is very important in emotion recognition processes. It improves model generalization and reduces the chances of overfitting by increasing interpretability and reducing computational costs. For each speech recording, we extracted audio features (audio frames) using `PyAudioAnalysis library` in the frequency and the time domain. 34 features including MFCC, Energy, Entropy of Energy, Zero crossing rate and Chroma (see table 1 for the description of the key features used) were extracted from the raw speech signals on a short-term basis and stored in the ARFF format. We consider entropy of energy as a measure of such energy change in speech signal energy levels. It is computed by dividing each frame into sub-frames of fixed duration. Then, the final feature value is computed as the entropy of that sequence of (normalized) sub-energies. We also consider spectral entropy, which is similar to the entropy of energy but applied to the frequency domain. We also explore chroma vector, a 12-dimensional representation of spectral energy of the speech signal. Gainnakopoulos et al., (2014), used chroma vector in their experimental design for real time depression estimation using mid-term audio feature and they showed it can discriminate between different emotional states of mental health patients.

Feature Extraction

The feature extraction was based on Discrete Fourier Transformations of samples, but for MFCC, cepstral features were extracted after applying Inverse Discrete Fourier Transformations. We follow the parameter selection process described in (Giannakopoulos 2015) for the features and extraction processes. Short term analysis was used to extract features on which the audio files are divided into short-term windows or frames 20ms to 100ms in size (Fernandes et al. 2018). In this research, the window size was set as 50ms for each frame. Experiments were conducted on classifiers to observe the changes with overlapping and non-overlapping on short-term windows. In this experiment, to create an overlapping window a frame size of 50ms and frame step of 25ms was selected (50% overlapping). For non-overlapping windows, frame size and step were set to 50ms.

Since the shape of the extracted features would not be the same and the range would not be specific without normalization, we normalize the features after extraction from the files by subtracting each feature from maximum one to have a consistent shape across the set. This will prevent the unstructured features from reducing the accuracy of and recognition rate (Zisad et al. 2020). Augmentation of the audio database typically generates new audio files by performing some kind of special operation on the original database, such as injecting noise, altering pitch, changing voice tract, and changing speed. The `NoiseAug` function from the `nlpaug package` is used to inject noise into all of the files in this paper.

Feature Name	Description
Energy	*The sum square of the signal values, normalized by the respective frame length*
Entropy of Energy	*The entropy of subframes' normalized energies. It can be interpreted as a measure of abrupt changes.*
Zero-Crossing Rate	*The rate of sign-changes of the signal during the duration of a particular frame*
MFCCs	*A cepstral representation, where frequency bands are not linear but distributed according to the Mel-scale*
Chroma Vector	*A 12-element representation of the spectral energy where the bins represent the 12-equal tempered pitch classes of western type music.*
Spectral Centroid	*The centre of gravity of the spectrum*
Spectral Entropy	*Entropy of the normalized spectral energies for a set of sub-frames.*

Table 1. Speech features used in this paper.

Emotion Classification Process

The `audioTrainTest.py` component of the `PyAudioAnalysis` has the implementation of all the classifiers we used for our experiment. The `featureTrainTest` method of the component was used for classifier training using `Sklearn` functionality. Experiments were run on all five classifiers to observe the effect of overlapping and non-overlapping feature extraction. Each classifier was trained with different parameters that enable algorithms to tune with the dataset. Different parameters with different values were selected for each classifier. Then, based on grid search the best parameter for each classifier was identified. Then, cross-validation was used to select an optimal parameter with the best parameter value for classifiers. The details of parameters and their values are presented in Table 2. `PyAudioAnqlysis` was used to train models. We normalized the feature sets using 0-mean and 1-standard deviation.

Classifier	Parameter	Parameter Values
SVM	Parameter C	0.001, 0.01, 0.5, 10.0, 20.0
RF	Number of trees	10, 25, 50, 100, 200, 500
KNN	Number of nearest neighbors	1, 3, 5, 7, 9, 11, 13, 15
Gradient Boosting	Number of boosting stages	10, 25, 50, 100, 200, 500
Extra Tree	Number of trees	10, 25, 50, 100, 200, 500

Table 2. Classifiers and Parameters Values in this Research

Metrics and Performance Measurement

To maintain consistency with other existing works in this domain, we report the classification results and performance of the models using the following metrics given in (Schuller et al. 2009). It includes the

accuracy $Ac(i)$, precision $Pc(i)$, recall $Rc(i)$, and the F-Score $Fc(i)$ computed for each class $c(i)$ ($i = 1, \ldots, N$) . Where N is the number of emotional classes. We define the metrics as follows:

$$Ac(i) = \frac{tp(i) + tn(i)}{tp(i) + tn(i) + fp(i) + fn(i)}$$

$$Pc(i) = \frac{tp(i)}{tp(i) + fp(i)}$$

$$Rc(i) = \frac{tp(i)}{tp(i) + fn(i)}$$

$$Fc(i) = 2\frac{Pc(i) \times Rc(i)}{Pc(i) + Rc(i)}$$

tp(i), tn(i), fp(i), fn(i) represent the numbers of true positive, true negative, false positive and false negative of the classification outcome respectively. The F-1 score measures the harmonic mean of precision and recall and was averaged over all the classes of emotions used in the experiment.

We perform cross-validation and select the optimal parameters for the evaluation to compute the average F1 value. To select the best parameter for each model, 10 validation experiments for each parameter were performed. For each experiment, 90% of features were used for training. After the 10th experiment, average accuracy and F1 values are computed. We base our comparison of the classifiers on these values with different parameters.

Results and Analysis

We performed our experiments in two different phases. First, we only used the TESS dataset, and developed and trained the ML models based on 10-fold cross-validation. For these classifiers, after determining the best parameters, F1 values were extracted for models' performance comparison. Experiments were conducted to study the effect of overlapped and non-overlapped feature extraction on the performance of models. The F1 score results of our experiments are shown in Figure 3. In this figure, the performance of all the classifiers with different parameter values for both overlapping and nonoverlapping cases are shown.

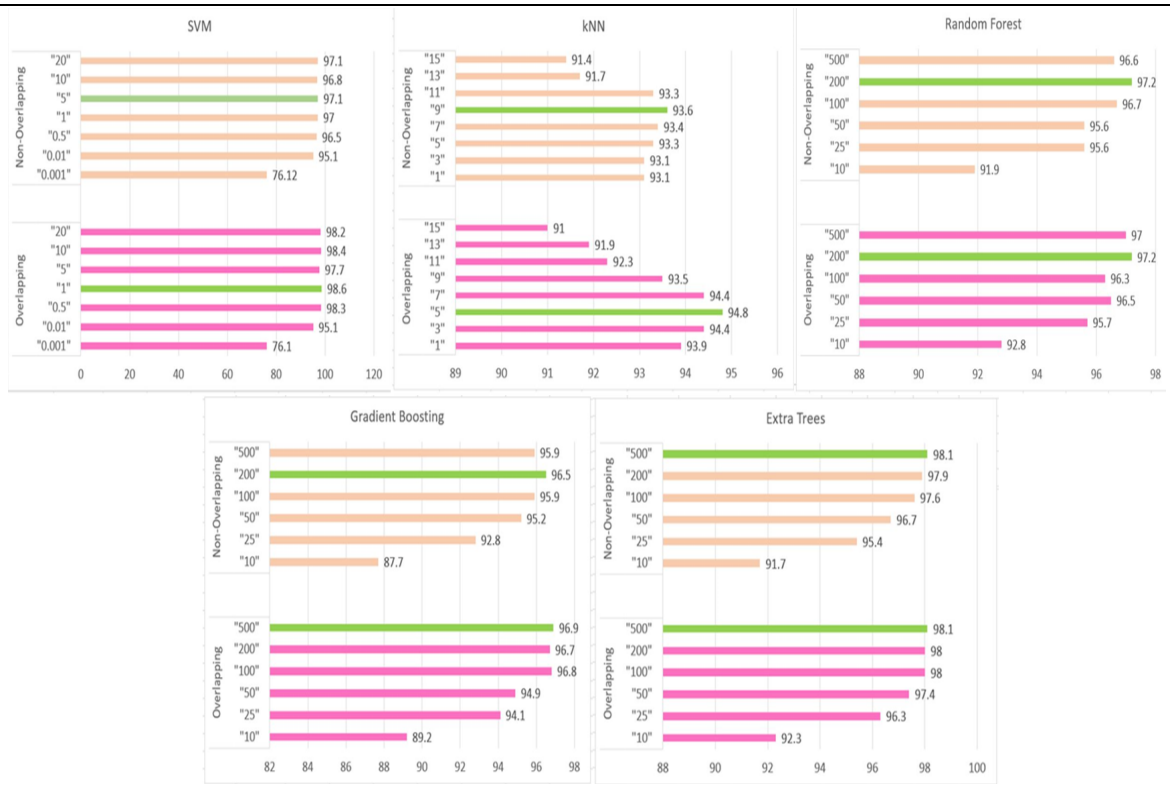

Figure 3. Different algorithms and their F1 values comparison

The performance of SVM in our experiment was low when the C value was low [0.001], in both overlapping and non-overlapping settings. However, it gave good performance results when the 'C' value was set between 0.5 and 20, so this value was chosen. kNN performance in the case of overlapping feature selection was high when K values were between 7 and 3. However, in the non-overlapping case, the best performance was shown with K values ranging from 11 to 5. Therefore, it was concluded the model was able to accurately classify speech files with a smaller number of neighbours taken into consideration in case of overlapping. The computational cost is also observed to increase with higher K values. In all the cases we examined, overlapped feature extraction gives better performance. In RF, no significant improvement in performance was observed in the case of overlapping and non-overlapping feature extraction when the number of trees (N) was increased above 100. This implies RF was able to achieve higher accuracy at a lower value of N (number of trees) in the case of overlapping for our selected dataset. To obtain the optimal number of trees in Gradient Boosting, several experiments were done. The settings for the experiments include setting the learning rate to 0.1, the maximum depth of tree to three and the minimum sample split to two. The performance of the model was observed to increase by increasing the number of trees. We observed however that there was a limit to this, no significant rise in performance when the number of trees increased above 100 in case of overlapping and 200 in the case of non-overlapping. Extra trees select a random value to perform split while constructing trees; this could be the reason for its better performance compared to other classifiers. It is also observed that, in cases of overlapping and non-overlapping, the model's performance increased by increasing the number of trees.

Based on the observation, kNN had the lowest performance and SVM had the best performance in overlapping feature extraction, while in the non-overlapping Extra Tree demonstrated better performance. In the case of overlapped feature selection, SVM produced roughly the same performance as extra trees (98.6% vs. 98.1%). In non-overlapped settings, SVM had the performance of 97.1% and extra trees achieved 98.1%. In the case of overlapping, the F1 value for kNN is 94.8% while for non-overlapping F1 was 93.6%. For Gradient Boosting, overlapping and non-overlapping gave F1 values of 96.9% and 96.5% respectively. Both overlapped and non-overlapped cases using RF, gave a result of 97.2% for the F1 value. As demonstrated in Figure 4, it can be observed that there was a minor difference in model performance based

on the F1 value for overlapped and non-overlapped settings and the overlapped feature extraction resulted in an increase in accuracy for most models.

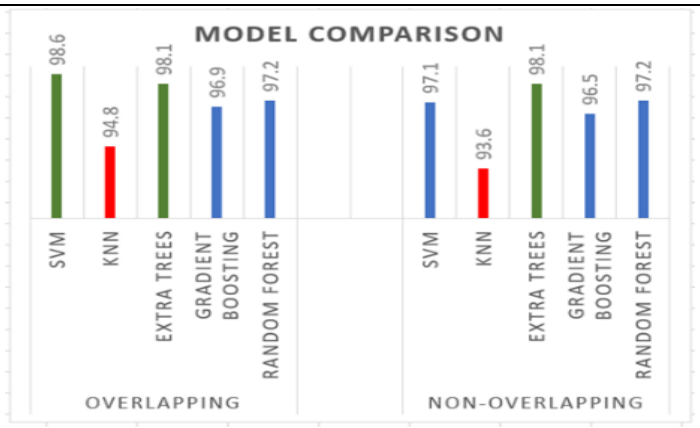

Figure 4. Overall comparison of models' performance based on F1 value

In the second phase of our experiment, we merged the RAVDESS, and TESS datasets based on their common emotions in both datasets to create a large dataset for the model training process. This helps us in developing ML models which not only fit the training data but also generalizes well on test/validation data. Moreover, we used a custom speech emotion dataset from 15 candidates (160 files), which capture all the 7 emotions considered in this experiment. Then the merged dataset was split in ratio 8:2 for training and testing respectively. To avoid oversampling, we balance the datasets to avoid sampling issues (e.g., oversampling). In this stage, we also added and used a Bagging classifier which is popular for emotion recognition as an ensemble ML algorithm (Kamble and Sengupta 2022). For Bagging parameters we used, Maximum Features, Number of Estimators and Maximum samples with the following values, derived from the grid search, 0.5, 50 and 1, respectively.

Figure 5 shows the performance of the models with respect to the testing sets.

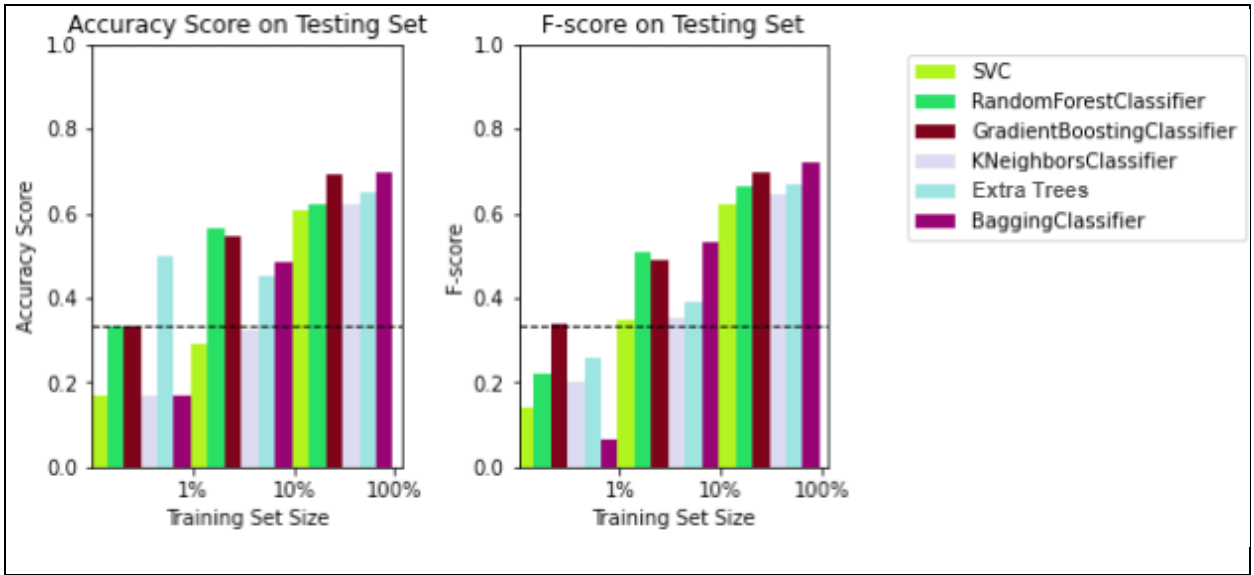

Figure 5. Accuracy and F1-Score of the Models Using Different Training Sizes

The performances of the models increased as the size of the training set increased. In this setting, SVC outperforms the other models with 89.426% predictive accuracy, while Gradient Boosting and RF classifiers perform extremely well, achieving 89.098% and 81.214% accuracy respectively. In Figure 6, we evaluated the models using a subset of the training dataset. The F1-score and accuracy significantly increased for all the models.

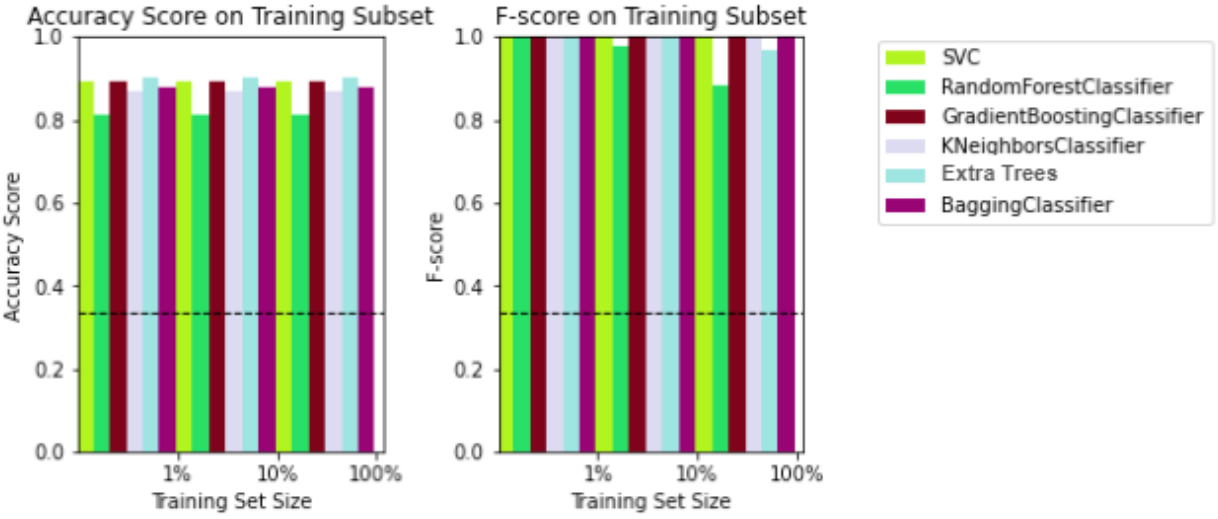

Figure 6. Accuracy and F1-Score of the Models Using Subset of Training Data

Discussion and Conclusion

We have treated SAER as a classification problem, i.e. classifying speech data/signals to a particular class of emotion using algorithms for which different classification algorithms could be used.

In our research, we successfully developed and compared the performance of several ML models for automatic speech emotion recognition. We achieved this by evaluating the emotion recognition accuracy of each model using common emotional databases with a custom dataset in both overlapping and non-overlapping feature extraction settings to improve the system performance and generalisation. We found that overlapping feature extraction slightly increased accuracy. The custom dataset helps us in developing models with improved generalization and reduced overfitting by increasing interpretability and reducing computational costs. This could be potentially important in developing a smart recognition system.

In comparison with (Ooi et al. 2014) research we have achived a better performance and accuracy. We successfully classified six emotions plus nutral, while in (Ooi et al. 2014) classification accuracy was low for detecting more than two emotions. Likewise, the presented work by (Chatterjee et al. 2018) constrained with two to three emotion categories. We have also reached a better accuracy in comparison with (Ghai et al. 2017) models with SVM, Random Forest and Gradient Boosting classifiers, as it is shown in our Figure 5.

However, it should be noted that, the aim of our research was not only to develop the ML models for emotions recognition; but, the primary goal was to provide insight into the effectiveness of using ML in capturing and recognising emotions from speech and leveraging it for generic mental health applications. The extracted speech features to recognise the emotions were supported by the literature in computer science and mental health domains.

The results of this research could be the baseline to guide the future design and evaluation of speech emotion recognition support systems for mental health practitioners. This system could be used as a supportive tool for mental health practitioners, especially in situations where patients are impaired in

recognizing their emotions or when the practitioners need to monitor patients' treatment progress. This gives practitioners the ability to monitor the emotional state of their patients more frequently and act more effectively in emergency situations, such as where there is a risk of suicide. Furthermore, the system can be used to see how the patients react to their treatments or the prescribed medication. Such a system may address practitioners' unwillingness and lack of uptake of tele-mental health services (Ojha and Syed 2020) such as telepsychiatry, teleconsultation and telediagnosis. Since one of the issues in telemental health is that physicians are not able to fully monitor the patient's reaction and body language to precisely decide about their emotional state assessment. These approaches g may allow an increase in support for people in the community and improve equity of access to mental health services.

Despite the success of the current research, further research will be required to investigate the usability of speech emotion recognition in real-world mental health treatment and assessment while meeting the requirements of practitioners. Particularly, the usability of the system can be investigated in facilitating telemental health which has become particularly important due to the COVID pandemic. Systems developed for SER may also be useful in training new mental health practitioners as they provide additional support in emotion detection for those in the early stages of their careers.

We noted that this preliminary research has some limitations. Although normalization and noise addition techniques were implemented on these data, it is not efficient enough to classify real-time normal or random speech records. To tackle this issue, we will need a larger dataset, with more diverse voices and larger computational capacity. We planned to use clinical datasets with auxiliary data in our future work. Empirical research, comparing SER outputs with self-reported emotions and expert evaluations are critical for validation of this approach if it is to be used in clinical practice. Other issues to consider in the future include cases where people deliberately mask emotions for example when using sarcasm or irony. Larger-scale datasets will be helpful to enhance the system performance, especially by increasing the diversity of the speakers' backgrounds. Audacity software used in (El Hammoumi et al. 2018) can be used to record real-time speech and use that to test models. Real-time systems will need to deal with security and capacity issues. Furthermore, empirical research and expert involvement could be useful to determine the effectiveness of results and how to improve the analysis produced by speech-based emotion recoition systems.

Acknowledgements

This search is supported by AUT Contestable Grant Award 2021, AUT Faculty of Design and Creative Technology Summer Studentship 2021-2022 and AUT School of Engineering, Computer and Mathematical Sciences Emerging Researcher Grant

References

Akçay, M. B., and Oğuz, K. 2020. "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers," *Speech Communication* (116), pp. 56-76.

Chang, K.-h., Fisher, D., Canny, J. F., and Hartmann, B. 2011. "How's My Mood and Stress?: An Efficient Speech Analysis Library for Unobtrusive Monitoring on Mobile Phones," *BODYNETS*.

Chatterjee, J., Mukesh, V., Hsu, H., Vyas, G., and Liu, Z. 2018. "Speech Emotion Recognition Using Cross-Correlation and Acoustic Features," *16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, Athens, Greece pp. 243-249.

Cheng, X., Wang, X., Ouyang, T., and Feng, Z. 2020. "Advances in Emotion Recognition: Link to Depressive Disorder."

Copeland, M. E. 2002. "Wellness Recovery Action Plan," *Occupational Therapy in Mental Health* (17:3-4), pp. 127-150.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. 2015. "A Review of Depression and Suicide Risk Assessment Using Speech Analysis," *Speech Communication* (71), pp. 10-49.

Dupuis, K., and Pichora-Fuller, M. K. 2010. "Toronto Emotional Speech Set (Tess)." Retrieved 15 July 2021, from https://tspace.library.utoronto.ca/handle/1807/24487

El Hammoumi, O., Benmarrakchi, F., Ouherrou, N., El Kafi, J., and El Hore, A. 2018. "Emotion Recognition in E-Learning Systems," *6th International Conference on Multimedia Computing and Systems*, Rabat, Morocco: IEEE, pp. 1-6.

Elfenbein, H. A., and Ambady, N. 2002. "On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis," *Psychological Bulletin* (128:2), pp. 203-235.

Feraru, M., and Zbancioc, M. 2013. "Speech Emotion Recognition for Srol Database Using Weighted Knn Algorithm," *Proceedings of the International Conference on Electronics, Computers and AI*, Pitesti, Romania, pp. 1-4.

Fernandes, V., Mascarehnas, L., Mendonca, C., Johnson, A., and Mishra, R. 2018. "Speech Emotion Recognition Using Mel Frequency Cepstral Coefficient and Svm Classifier," *2018 International Conference on System Modeling & Advancement in Research Trends*, pp. 200-204.

Fernandez, R., and Picard, R. W. 2003. "Modeling Drivers' Speech under Stress," *Speech Communication* (40:1), pp. 145-159.

Ghai, M., Lal, S., Duggal, S., and Manik, S. 2017. "Emotion Recognition on Speech Signals Using Machine Learning," *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, pp. 34-39.

Giannakopoulos, T. 2015. "Pyaudioanalysis: An Open-Source Python Library for Audio Signal Analysis," *PLOS ONE* (10:12), p. e0144610.

Hall, J. A., Harrigan, J. A., and Rosenthal, R. 1995. "Nonverbal Behavior in Clinician—Patient Interaction," *Applied and Preventive Psychology* (4:1), pp. 21-37.

Harati, S., Crowell, A., Mayberg, H., and Nemati, S. 2018. "Depression Severity Classification from Speech Emotion," *Annu Int Conf IEEE Eng Med Biol Soc* (2018), pp. 5763-5766.

Hasan, M., Rundensteiner, E., and Agu, E. 2019. "Automatic Emotion Detection in Text Streams by Analyzing Twitter Data," *International Journal of Data Science and Analytics* (7:1), pp. 35-51.

Jacintha, V., Simon, J., Tamilarasu, S., Thamizhmani, R., Thanga yogesh, K., and Nagarajan, J. 2019. "A Review on Facial Emotion Recognition Techniques," *International Conference on Communication and Signal Processing*, pp. 0517-0521.

Kamble, K. S., and Sengupta, J. 2022. "Ensemble Machine Learning-Based Affective Computing for Emotion Recognition Using Dual-Decomposed Eeg Signals," *IEEE Sensors Journal* (22:3), pp. 2496-2507.

Lech, M., and He, L. 2014. "Stress and Emotion Recognition Using Acoustic Speech Analysis," in *Mental Health Informatics,* M. Lech, I. Song, P. Yellowlees and J. Diederich (eds.). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 163-184.

Livingstone, S. R., and Russo, F. A. 2018. "The Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," *PLOS ONE* (13:5), p. e0196391.

Minardi, H. 2013. "Emotion Recognition by Mental Health Professionals and Students," *Nurs Stand* (27:25), pp. 41-48.

Miner, A. S., Haque, A., Fries, J. A., Fleming, S. L., Wilfley, D. E., Terence Wilson, G., Milstein, A., Jurafsky, D., Arnow, B. A., Stewart Agras, W., Fei-Fei, L., and Shah, N. H. 2020. "Assessing the Accuracy of Automatic Speech Recognition for Psychotherapy," *npj Digital Medicine* (3:1), p. 82.

Mitra, V., Shriberg, E., Vergyri, D., Knoth, B., and Salomon, R. M. 2015. "Cross-Corpus Depression Prediction from Speech," *International Conference on Acoustics, Speech and Signal Processing*, pp. 4769-4773.

Moore, E. I. I., Clements, M., Peifer, J., and Weisser, L. 2004. "Comparing Objective Feature Statistics of Speech for Classifying Clinical Depression," *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 17-20.

Morris, M. E., Kathawala, Q., Leen, T. K., Gorenstein, E. E., Guilak, F., Labhard, M., and Deleeuw, W. 2010. "Mobile Therapy: Case Study Evaluations of a Cell Phone Application for Emotional Self-Awareness," *J Med Internet Res* (12:2), p. e10.

Mustafa, M., Yusoof, M. A. M., Don, Z. M., and Malekzadeh, M. 2018. "Speech Emotion Recognition Research: An Analysis of Research Focus," *International Journal of Speech Technology* (21:1), pp. 137-156.

Ojha, R., and Syed, S. 2020. "Challenges Faced by Mental Health Providers and Patients During the Coronavirus 2019 Pandemic Due to Technological Barriers," *Internet interventions* (21), pp. 100330-100330.

Ooi, C. S., Seng, K. P., Ang, L.-M., and Chew, L. W. 2014. "A New Approach of Audio Emotion Recognition," *Expert Systems with Applications* (41:13), pp. 5858-5869.

Oosterwijk, S., Lindquist, K. A., Anderson, E., Dautoff, R., Moriguchi, Y., and Barrett, L. F. 2012. "States of Mind: Emotions, Body Feelings, and Thoughts Share Distributed Neural Networks," *NeuroImage* (62:3), pp. 2110-2128.

Picard, R. W. 2000. *Affective Computing*. Cambridge, Mass.: MIT Press.

Ren, F., and Quan, C. 2012. "Linguistic-Based Emotion Analysis and Recognition for Measuring Consumer Satisfaction: An Application of Affective Computing," *Information Technology and Management* (13:4), pp. 321-332.

Schuller, B. W. 2018. "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM* (61:5), pp. 90-99.

Shah Fahad, M., Ranjan, A., Yadav, J., and Deepak, A. 2021. "A Survey of Speech Emotion Recognition in Natural Environment," *Digital Signal Processing* (110), p. 102951.

Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., Toda, H., Saito, T., Tanichi, M., Yoshino, A., and Tokuno, S. 2021. "Depressive Mood Assessment Method Based on Emotion Level Derived from Voice: Comparison of Voice Features of Individuals with Major Depressive Disorders and Healthy Controls," *International Journal of Environmental Research and Public Health* (18:10).

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., and Yang, X. 2018. "A Review of Emotion Recognition Using Physiological Signals," *MDPI Sensors* (18:7).

Stasak, B., Epps, J., Cummins, N., and Göcke, R. 2016. "An Investigation of Emotional Speech in Depression Classification," *INTERSPEECH*.

Tashev, I. J., Wang, Z. Q., and Godin, K. 2017. "Speech Emotion Recognition Based on Gaussian Mixture Models and Deep Neural Networks," *Information Theory and Applications Workshop*

Thome, J., Liebke, L., Bungert, M., Schmahl, C., Domes, G., Bohus, M., and Lis, S. 2016. "Confidence in Facial Emotion Recognition in Borderline Personality Disorder," *Personality Disorders: Theory, Research, and Treatment* (7:2), pp. 159-168.

Tin Lay, N., Foo Say, W., and Silva, L. C. D. 2001. "Speech Based Emotion Classification," *10th International Conference on Electrical and Electronic Technology*, pp. 297-301 vol.291.

Tripoli, G., Quattrone, D., Gayer-Anderson, C., Rodriguez, V., Ferraro, L., Cascia, C. L., Sartorio, C., Seminerio, F., Barbera, D. L., Morgan, C., Sham, P., Forti, M. D., and Murray, R. 2019. "O4.8. Can You Spot Emotions? Facial Emotion Recognition and Genetic Risk for Psychosis " *Schizophrenia Bulletin* (45:Supplement_2), pp. S172-S172.

Tsiakas, K., Watts, L., Lutterodt, C., Giannakopoulos, T., Papangelis, A., Gatchel, R., Karkaletsis, V., and Makedon, F. 2015. *A Multimodal Adaptive Dialogue Manager for Depressive and Anxiety Disorder Screening: A Wizard-of-Oz Experiment*. Corfu, Greece:

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. 2013. "Avec 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in: *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. Barcelona, Spain: Association for Computing Machinery, pp. 3–10.

Wang, L., Hu, G., and Zhou, T. 2018. "Semantic Analysis of Learners' Emotional Tendencies on Online Mooc Education," *MDPI Sustainability* (10:6), p. 1921.

Wilson, S. C., and Carryer, J. 2008. "Emotional Competence and Nursing Education: A New Zealand Study," *Nurs Prax N Z* (24:1), pp. 36-47.

Zisad, S. N., Hossain, M. S., and Andersson, K. 2020. "Speech Emotion Recognition in Neurological Disorders Using Convolutional Neural Network," *Brain Informatics,* M. Mahmud, S. Vassanelli, M.S. Kaiser and N. Zhong (eds.), Cham: Springer International Publishing, pp. 287-296.