university of groningen

faculty of science and engineering

University of Groningen

Voicing a Schizophrenic Mind

A natural language processing based approach to diagnose schizophrenia using speech

Behavioural and Cognitive Neurosciences
Cognitive Neuroscience and Cognitive Modelling (C-Track)

Karthik Charan Raghunathan
S4399617

Supervisor
Marieke Van Vugt
Cognitive Modelling group, University of Groningen.

umcg

July 31, 2022

Contents

1 Introduction

Speech is the ability to express one's thoughts, ideas, and emotions by means of articulate vocal sounds. It can be viewed as a tool to access the mental state of a person [1]. In clinical practice, impressions of speech have been the primary source of diagnosis for several mental disorders such as depression and schizophrenia spectrum disorders [2]. Speech abnormalities have been a prominent symptom of schizophrenia ever since its early definitions [3, 4]. Several characteristic symptoms such as alogia (i.e., poverty of speech) and flat affect (i.e., lack of emotional expressions) have been associated with people affected with schizophrenia [5]. This severe language disturbance is also known as Formal Thought Disorder (FTD) [3]. The disturbances are a part of the thought process themselves rather than the thought content. FTD is characterised by impaired verbal communication, i.e. disorganised speech [3] exhibiting loose associations, derailment and tangentiality (i.e. a sequence of unrelated or only remotely related ideas), or incoherence. There are several studies that elucidate a strong association between language impairment and cognitive decline in FTD [6, 7]. At a semantic level, patients affected with schizophrenia demonstrate difficulties with lexical selection in ongoing speech [4], reduced proactive inhibition [8, 9], and semantic priming (i.e., the phenomenon of response to a target word would be faster if the preceding word is semantically related; [10]). At a syntactic level, they display reduced syntactic priming [11], fewer discourse markers [12] and the use of syntactically less complex sentences [13]. Other language-based disturbances include diminished capacity to create coherent narratives [14], word approximations [8], monotonous intonation [8], and confused references [15, 16]. It is also noted that the speech of a schizophrenic person is more hesitant and contains frequent pauses characterised by alogia [17, 18].

Schizophrenic patients often experience psychotic episodes (i.e. relapse or psychosis) depending on the severity of the disorder. This is a symptomatic phase where a schizophrenic patient exhibits an abnormal perception of reality (accompanied by incidents such as hallucinations). The onset of a psychosis not only introduces an immediate mental and physical burden for the patient but also long-term deterioration of cognitive functioning and quality of life. Although there have been substantial advances in the field of neuroscience concerning schizophrenia, we are yet to yield biomarkers that can aid in automated diagnostics of the disorder, that are clinically relevant and can reliably predict the future onset of psychosis at the level of an individual patient. The dynamic nature of the relapse also accounts for the complexity in reliably predicting the onset of a psychosis.

1.1 Previous Research

In earlier studies, it was observed that fluency in speech is lower in schizophrenic patients than healthy controls. Elvevag et al. noticed this phenomenon using a verbal fluency task and few structured interviews among schizophrenic and healthy people and proved that this discourse can help reliably identify schizophrenic patients from healthy population [19]. Another interesting study found that the certain linguistic determiners such as speech-to-pause ratio, number of words per phrase etc. along with vector similarity between adjacent sentences in free speech could help identify clinically high risk (CHR) indivuduals that would possibly convert to psychosis [20]. Similarly, semantic coherence in language was found to be a potent predictor for the onset of psychosis between two different groups of CHR individuals [21]. Another study used graph-theory based tools to measure the connectivity between words to diagnose schizophrenia in individuals experiencing first episode of psychosis (FEP) [22]. Another different approach in diagnosis of schizophrenia and prediction of psychosis involves digital phenotyping. This method involves analysing a user's digital signatures to identify behavioural patterns that might help in diagnosis of several mental disorders. [23]

Recent developments in artificial intelligence-based natural language processing promise to complement clinical psychiatry in automated diagnosis and analysis of mental disorders. Language is assessed by quantitatively analysing phonetics, syntax, and semantics using automatic speech recognition and computing linguistic tools. Recently, several machine learning models have been devised that can classify patients with mental disorders, including schizophrenia, from healthy controls with accuracies ranging in the 80% to 90% range, leveraging the salient features of their speech [24, 25, 26]. Automated computerised language analysis has been previously used to assess schizophrenia based on various speech correlates.

Latent Semantic Analysis (LSA) is a technique in the natural language processing field that calculates the relationship between a set of paragraphs or sentences and the words contained within those sentences. It uses singular value decomposition, a mathematical technique, to scan unstructured data to find hidden relationships between terms and concepts. LSA is being extensively used to analyse the semantic structures in speech to identify reduced coherence associated with speech in patients affected with schizophrenia [19]. This high-dimensional analysis helps in differentiating patients affected with schizophrenia from healthy individuals based on subtle patterns and differences in speech patterns. Another technique includes structural speech analysis that leverages statistics-based language analysis for the computation of useful mathematical parameters to aid computerised language analysis. LSA combined with structural speech analysis was able to accurately differentiate between first-degree relatives of schizophrenia patients and unrelated healthy individuals [27]. This study analysed the word associations, verbal fluency, and various discourse in speech samples such as narrative speech (story telling) and expository speech (i.e., descriptions of abstract concepts). They exploited the groupings of words either within or across speech samples and also accessed the overall coherence in it. Their results suggests that computerised speech analysis can not only be used for discriminating a schizophrenic patient using their language but they are powerful enough to identify healthy relatives who may have subtle underlying genetic vulnerabilities to schizophrenia.

1.2 Acoustics: A different dimension

Recent studies suggest promising results in differentiating schizophrenics from healthy individuals using automated analysis of acoustic speech patterns, the semantic aspects such as complexity and coherence of the speech are well utilised in the automated assessments of the language [20, 21, 28, 29, 30, 31]. However, the crucial aspects that are often overlooked are the acoustic aspects (such as the pitch, loudness and spectral flux) of the speech. These articulatory components obtained from the speech signals can be exemplary predictors in classifying the individual patients [32].

Articulation and other aspects of speech production can be quantified using the acoustic signal. Sound waves can be decomposed into formants, which are acoustic resonance frequencies that indicate the position and movement of the articulatory organs during speech. Jaw/mouth openness and tongue height are indicated by F1 (First formant), whereas tongue orientation (front/back) and lip rounding are indicated by F2. F0, or fundamental frequency, is also an audio characteristic of pitch of the tone. In limited samples, such acoustic speech variables (F0 and F2 variability) have been linked to specific negative symptoms [33, 34] and used as differentiators between people with schizophrenia-spectrum disorders and healthy people, with overall classification accuracy ranging from 81 to 94 percent [35, 36]. In addition to this, classifying schizophrenic patients based on acoustic parameters can be computationally efficient due to the fact that there is little to no transformation or processing of data required.

1.3 Inter-individual variability: A new insight?

Studies exploring the speech aspects of subjects in order to identify patients with schizophrenia often compare the speech data of subjects to that of healthy controls. However, in the present study, we propose to focus on the inter-individual variability. We are interested in looking into the time-based variability of the acoustic features in the speech data of individual subjects.

Here, we propose a few machine learning classifiers and deep neural network models that leverage the acoustic components of speech to predict later onset of psychosis at an individual level. The main aim of the study is to position these acoustic parameters on a temporal scale and explore the value of considering these parameters as a time series for classification of relapse vs non-relapse cases. Based on this aim, we hypothesise that the acoustic features of voice recordings of schizophrenic patients can be a reliable biomarker for predicting if a patient is experiencing/will experience a relapse or not. Secondly, we hypothesise that changes in the acoustic features of an individual over time can help us predict a transition into a relapse.

We try to device a Support Vector Machine (SVM) and Random Forest (RF) classifiers and Deep Neural Networks that learn to classify the onset of psychosis in a schizophrenic patient using various nuances in speech signal such as pitch variability, mean length of utterance, clauses per utterance etc. Speech signals are time-varying signals with complex correlations with a range of different timescales. We chose the SVM classifier since it evaluates the notion of distance between the vectors in the feature space [37]. This property of the algorithm may help in discriminating the acoustic features of the 'Relapse' and 'Non-Relapse' classes in a higher dimensional feature space. The random forest classifier is a type of decision tree classifier that consists of a large number of individual decision trees. Each individual tree splits into a class prediction, and the class with the most votes becomes the model's prediction [38]. The random forest classifier was chosen due to its applicability and performance on complex datasets. The deep neural networks are advantageous in extracting the relevant features during the training phase of the model over the dataset. The layers of the models help extract the features for classifying the data in a complex dataset [39]. Considering the complexity of the problem at hand we chose to analyse it using deep learning models. With these acknowledgements, we propose the above mentioned machine learning and deep neural network models to better capture the time-based nuances in the data and hence classify the subjects with a competent accuracy.

2 Methods

2.1 Participants

A total of N=73 patients diagnosed with schizophrenia-spectrum disorder between 2017 and 2022 at the University Medical Center Groningen and University Medical Center Utrecht were included in this study. Data used in this study are from an ongoing HAMLETT trial. As an inclusion criterion, all subjects in this trial had their very first psychosis 3-6 months before inclusion. Speech recordings of the subjects were collected during their appointments with the clinician at an interval of 3 months, with the first appointment known as V1, second appointment as V2, and the third as V3. In this study, the data is restricted to the speech recordings until V3. To have an overview of the disorder, Positive and Negative Syndrome Scale (PANSS) scores of the patients were reported. For a major part of this study, we are interested in looking at the transition in speech parameters over a period of time and the rate of change of speech parameters over time. To facilitate this, the data from participants with less than two recordings were excluded from the analysis, with the exception of the first analysis, which brings down the sample size to 68 patients for the majority of the study. Since the first analysis involves classifying a patient into a ’Relapse’ or a ’Non-relapse’ class based on a single voice recording, the voice recordings of all the 73 participants were used for analysis. All participants in this study are over 16 years old and are native Dutch speakers.

2.2 Data

2.2.1 Speech recordings and preprocessing

A semi-structured interview, 5 to 30 minutes long (11 minutes on average), was conducted with a series of neutral, open-ended questions. An AKG-C544l cardioid microphone was used to record the speeches of the participants and the interviewer on two separate channels. The audio was digitally recorded on the Tascam DR40 solid-state recorder at a sample rate of 44.1 kHz. For more information on this methodology, refer to previous reports from our group [40, 41, 32]. The following steps were performed to eliminate crosstalk (i.e., speech of the interviewer that is recorded on the participant’s channel). (1) ”Comment on silence in the text grid” feature of PRAAT was used in the interviewer’s channel. (2) All audio segments resulting from the interviewer’s silence were selected in the participant’s channel, and (3) the resulting audio segment was concatenated into a new audio file containing only the participant’s audio.

2.2.2 Acoustic parameters

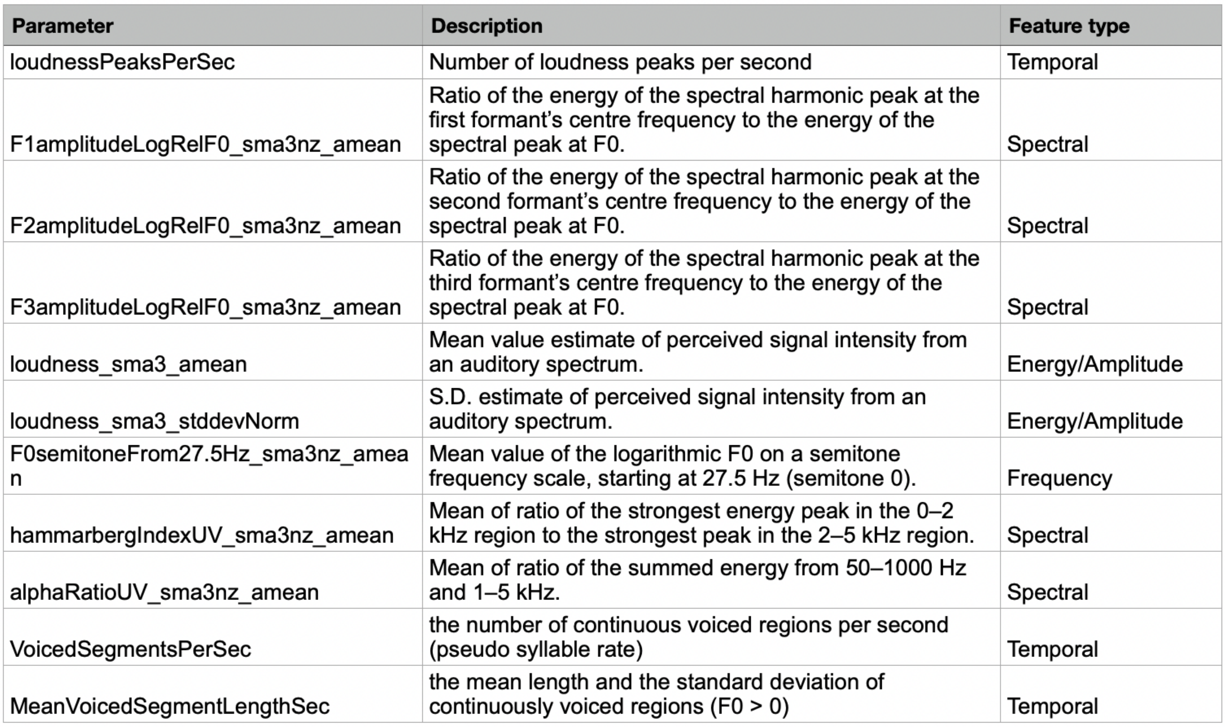
For this study, we chose the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [42], which acts as a basic standard acoustic parameter set for automatic voice analysis of clinical speech. This parameter set containing a total of 88 parameters was extracted from the speech recordings of the patients using OpenSMILE feature extraction toolkit, version 2.0 [43]. The feature set consists of 43 spectral parameters (e.g. Mel-frequency cepstral coefficients (MFCCs)), 24 frequency based parameters (e.g. fundamental frequency), 14 amplitude based parameters (e.g. intensity) and 6 temporal parameters (e.g. speech rate). The acoustic parameters that are extracted are the mean values for the entire segment of speech from each interview.

2.3 Data Preprocessing

2.3.1 Feature Selection

Selecting the features that are fed into the model for training vastly influences the performance of a machine learning model. Reducing the number of features by eliminating the redundant and irrelevant features can improve the model training speed and reduce overfitting. For this study, we derived 11 acoustic features that present more useful information for identifying schizophrenia based on a combination of a study by Wolters et al. [44] and principal component analysis on the data. The features used are presented in Table 1**??**.

Table 1. Acoustic features for Classification: Relapse vs Non-Relapse



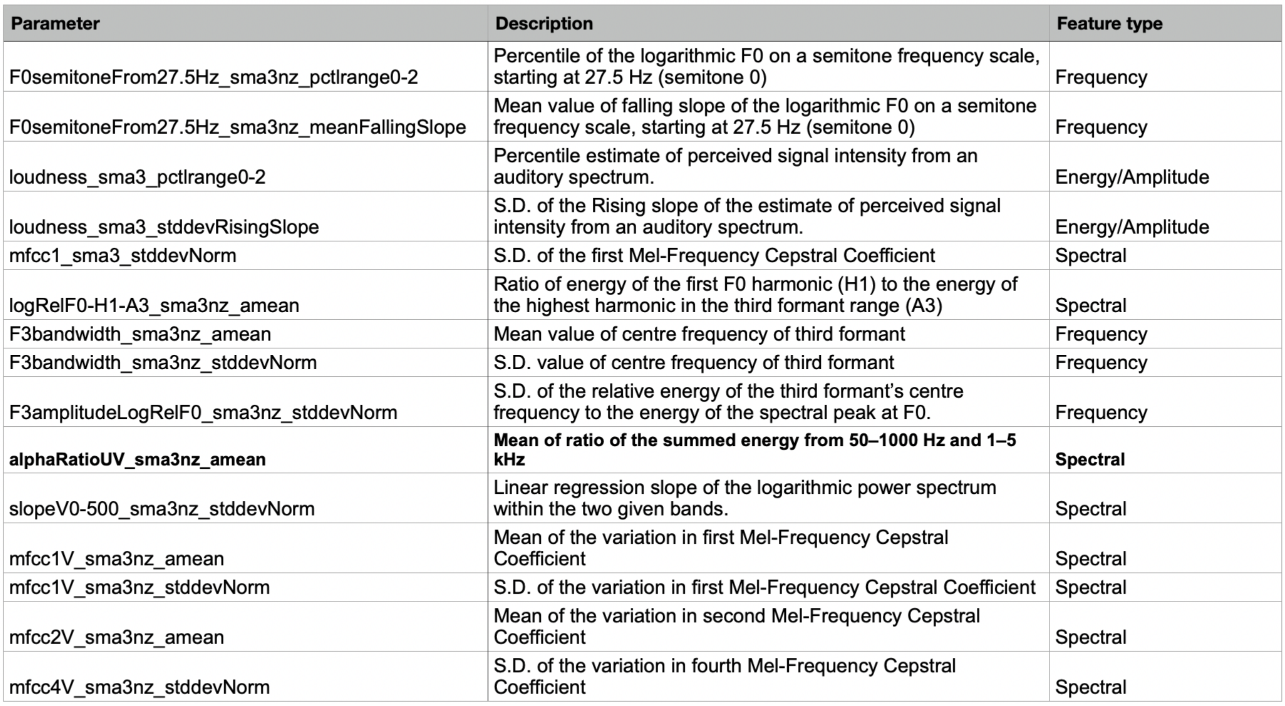
Parameter	Description	Feature type
loudnessPeaksPerSec	Number of loudness peaks per second	Temporal
F1amplitudeLogRelF0_sma3nz_amean	Ratio of the energy of the spectral harmonic peak at the first formant's centre frequency to the energy of the spectral peak at F0.	Spectral
F2amplitudeLogRelF0_sma3nz_amean	Ratio of the energy of the spectral harmonic peak at the second formant's centre frequency to the energy of the spectral peak at F0.	Spectral
F3amplitudeLogRelF0_sma3nz_amean	Ratio of the energy of the spectral harmonic peak at the third formant's centre frequency to the energy of the spectral peak at F0.	Spectral
loudness_sma3_amean	Mean value estimate of perceived signal intensity from an auditory spectrum.	Energy/Amplitude
loudness_sma3_stddevNorm	S.D. estimate of perceived signal intensity from an auditory spectrum.	Energy/Amplitude
F0semitoneFrom27.5Hz_sma3nz_amean	Mean value of the logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0).	Frequency
hammarbergIndexUV_sma3nz_amean	Mean of ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region.	Spectral
alphaRatioUV_sma3nz_amean	Mean of ratio of the summed energy from 50–1000 Hz and 1–5 kHz.	Spectral
VoicedSegmentsPerSec	the number of continuous voiced regions per second (pseudo syllable rate)	Temporal
MeanVoicedSegmentLengthSec	the mean length and the standard deviation of continuously voiced regions (F0 > 0)	Temporal

For the final analysis, since we are looking at the rate of change of these features at an individual level, it was important to reselect a set of features whose rate of change accounts more for identifying a relapse or remission. For this, we used Recursive Feature Elimination (RFE) using a decision tree classifier as the selection algorithm to select a set of 15 features. We started with 5 parameters and evaluated the model by adding 5 features for every iteration of RFE. The performance of the model did not improve after 15 features, hence we finalised with 15 features. The RFE algorithm trains a model with all the features at core and ranks the features by importance. Later, the features with the least importance are eliminated, and the model is fitted to the current set of features. This process is repeated until the desired number of features are selected. The list of features selected are presented in Table 2**??**. Features similar to Table 1 are highlighted in bold.

2.3.2 Data Augmentation

When presented with a dataset with a relatively huge imbalance in the data between the classes, the model might fail to learn a decision boundary to differentiate the minority class. The speech data of

Table 2. Acoustic features for Classification: Rate of Change of features over time



Parameter	Description	Feature type
F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	Percentile of the logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0)	Frequency
F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	Mean value of falling slope of the logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0)	Frequency
loudness_sma3_pctlrange0-2	Percentile estimate of perceived signal intensity from an auditory spectrum.	Energy/Amplitude
loudness_sma3_stddevRisingSlope	S.D. of the Rising slope of the estimate of perceived signal intensity from an auditory spectrum.	Energy/Amplitude
mfcc1_sma3_stddevNorm	S.D. of the first Mel-Frequency Cepstral Coefficient	Spectral
logRelF0-H1-A3_sma3nz_amean	Ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3)	Spectral
F3bandwidth_sma3nz_amean	Mean value of centre frequency of third formant	Frequency
F3bandwidth_sma3nz_stddevNorm	S.D. value of centre frequency of third formant	Frequency
F3amplitudeLogRelF0_sma3nz_stddevNorm	S.D. of the relative energy of the third formant's centre frequency to the energy of the spectral peak at F0.	Frequency
alphaRatioUV_sma3nz_amean	**Mean of ratio of the summed energy from 50–1000 Hz and 1–5 kHz**	**Spectral**
slopeV0-500_sma3nz_stddevNorm	Linear regression slope of the logarithmic power spectrum within the two given bands.	Spectral
mfcc1V_sma3nz_amean	Mean of the variation in first Mel-Frequency Cepstral Coefficient	Spectral
mfcc1V_sma3nz_stddevNorm	S.D. of the variation in first Mel-Frequency Cepstral Coefficient	Spectral
mfcc2V_sma3nz_amean	Mean of the variation in second Mel-Frequency Cepstral Coefficient	Spectral
mfcc4V_sma3nz_stddevNorm	S.D. of the variation in fourth Mel-Frequency Cepstral Coefficient	Spectral

the minority class (Relapse) accounts only for 15.8% of the entire dataset. This data might be surprising since the number of patients that relapsed is greater than the number of non-relapse patients. However, it is important to note that speech recordings of the relapse patients are labelled as 'Relapse' only for the recording that falls into the window within which the patients relapsed (V1, V2, or V3). Also, it is difficult to obtain speech recordings of a relapse patient since they often would not be able to make it to the appointment within that window.

But it is important to balance the data to get reliable predictions from the model. One way to solve this problem is to undersample the majority class, but since we have very little data in the minority class it might not be the best approach considering the amount of data needed by ML and DL algorithms to learn patterns. Hence, the best approach would be to oversample the data in the minority class, also referred to as data augmentation. We used an algorithm called Synthetic Minority Oversampling TEchnique (SMOTE) by Chawla et al. [45]. It involves selecting data points of the minority class that are close in the feature space (k - nearest neighbours), drawing a line between these values and impute data points at a random space on the line. The number of data points, i.e., the k-nearest neighbours, were chosen to be 4 based on experimentation with k = N = 1, 2, and 4. The number of samples in the minority class that was synthetically imputed depended on the number of samples needed to balance the minority class to that of the majority class. The dataset was split into training and test sets before oversampling the minority class and SMOTE was used only on the training set to avoid data leakage into the test set which might lead to misleading accuracies. As a standard norm, 33% of the dataset was split into the test set for all the modelling in this study.

2.4 Statistical Analysis

For categorical variables, a 2 test, and for continuous variables independent sample t-test were used to assess differences between groups in demographic characteristics. An independent sample t-test was used to assess the difference in groups based on the acoustic features of both the groups. For the dataset containing the rate of change of acoustic features over a period of time, a one-way ANOVA was used to examine the difference. A Tukey's post hoc test was performed to correct for multiple comparisons between the three groups. Alpha was set to 0.05 for all analyses.

2.5 Classifiers

2.5.1 Machine Learning Classifiers

In this study we focus on two common machine learning classifiers: the Support Vector Machine (SVM) and the Random Forest (RF) Classifier. Based on previous studies we selected the SVM classifier to be suitable for the problem at hand. Several studies also used decision tree based classifiers for similar classification problems. Hence, we chose to use the random forest classifier which is a collection of decision trees. Apart from other classifiers, the RF classifier was chosen due to its applicability and performance on complex datasets.

The SVM algorithm classifies the data points by finding a hyperplane in the N-dimensional feature space that has the maximum margin distance between the data points of either class. SVMs are effective in high-dimensional feature spaces. We use the radial basis function (rbf) kernel as it is the best suited function for non-linear datasets. The most important parameters of an SVM algorithm are 'gamma' and 'C'; where the gamma parameter defines how far the influence of a single training example reaches and the 'C' parameter defines the maximisation of the decision boundaries. The gamma was set to 'auto' for the algorithm to figure the optimal value. We iterated over the following values for the 'C' parameter: [100, 10, 1, 0.1, 0.01, 0.001] and chose C = 1 since it exhibited the best classification performance (evaluated based on accuracy and AUC scores) for the given data set out of all the other C values.

The RF algorithm consists of a number of individual decision trees that acts as an ensemble, each tree branches out predicting a specific class and the class with the highest votes becomes the prediction. After brute forcing for values of 10 to 500 with increments of 10 for the number of trees and 10 to 100 for the depth of each tree we arrived at an optimal value of 50 trees with a depth of 20 each. Due to the small sample size, we used cross validation to estimate the performance of the models on our dataset. We used RepeatedStratifiedKfold cross validation since a single run of StratifiedKfold might provide a noisy estimate of the model's performance. The RepeatedStratifiedKfold algorithm repeats the procedure N times which results in a better estimate of the model's performance. We generally used 20-fold cross-validation (less in case of data unavailability in minority class) and N = 3 repeats for our dataset. We tested the models using 5, 10, 15, 20 and 30-fold cross-validation and chose 20-fold since it elicits the best performance.

2.5.2 Deep Neural Networks

A deep neural network consists of layers of neurons that are modelled loosely after the human brain and they exhibit a very good performance in recognising patterns in a dataset. We used sequential feed forward neural networks for classification of our dataset. Depending on the analysis, we added 7-8 layers of neurons to each of the neural networks and the number of epochs and the batch size were estimated based on the performance of the model at the time of training. For the neurons in the hidden layers, we tried using both 'ReLU' and 'Tanh' activation functions; the 'ReLU' activation function exhibited better classification for our problem. For binary classification, the output layer consisted of one neuron with 'sigmoid' activation function and we used 'binary cross entropy' as the loss function. Whereas for multiclass classification, since our problem consists of three classes, the output layer consisted of three neurons with 'softmax' activation function. 'Categorical cross entropy' was used as the loss function to evaluate the multiclass network. For all the neural network models, we used 'Adam' optimizer as the optimisation algorithm. Adam algorithm associates and updates each parameter in the network with an individual learning rate. The individual learning rates adapt themselves during the training steps. This optimizer is well suited for noisy gradients [46].

For analysing the transition of acoustic features over a period of time we also used a Convolutional Neural Network (CNN) model. CNNs are mainly used for image classification, they use matrix convolution over the numerical dataset. This convolution procedure extracts salient spatial difference-based features from the data. We ideated that using this method over the acoustic features at different timepoints would help us capture the changes in them over time. Generally, a two-dimensional convolution layer will be used for image classification. Here, we use a one-dimensional convolutional layer to convolve over the tabular numerical data. To realise this model, we appended two one-dimensional convolution layers with 512 filters each, over the deep neural network architecture described above. The data is arranged such that for each record, the column (2nd dimension) represents the time point and the 3rd dimension represents the acoustic features. For example, for 5 records with 11 acoustic features over 2 time points T and T-1, the dimensions of the dataset will be 5 x 2 x 11.

2.6 Predictive Analysis

2.6.1 Speech based diagnostic classifiers

To examine whether the acoustic features data can be a reliable biomarker for differentiating between a relapse vs a non-relapse schizophrenic patient, we fit the machine learning models over the acoustics data of the participants. We use 20-fold cross validation to evaluate the performance of the model based on the mean accuracy and mean AUC (Area Under Curve of the receiver operator characteristic (ROC) plot) scores across all the 20 folds.

2.6.2 Speech transition based diagnostic classifiers

Since we were analysing the transition of acoustic features over a period of time, participants with less than two speech recordings were excluded. Initially, we were interested in differentiating between the 'Relapse' and the 'Non-Relapse' class. Both the machine learning and the deep learning models were fitted over the data of the participants where the acoustic data at different points in time T, T-1 and T-2 (i.e. data from V1, V2 and V3) are appended sequentially together. That is, each participant is considered as an individual sample with acoustic features at three different points in time making up for 33 features per record (3 * 11 features). Since for each participant we have acoustics data from three different time points, the models will learn any relevant patterns that exist implicitly between the time points which would help us discriminate between classes based on a transition over a period of time. For participants with only two recordings, we replace the column values with zeros. We used this approach based on the assumption that the models will learn to ignore the zeros while searching for implicit patterns in the data.

An interesting paradigm was found during the analysis of the transition data; some patients exhibited a transition from a 'Relapse' state to a 'Non-Relapse' state. To analyse this paradigm, it was necessary to look into two time points, T and T-1 and the samples would be labelled as per the transition of the patient during this time frame. The above mentioned samples were labelled as ‘RelapsetoHealthy’ class, this classifies the samples into three classes, a set of patients that stays healthy (‘Non-Relapse’ class), a set of patients that transition from healthy to a Relapse (‘Relapse’ class) and a set of patients that transition from a Relapse back to a healthy condition/Remission (‘RelapsetoHealthy’ class). To facilitate this we arranged the data as samples containing acoustic features at two time points, namely T and T-1. The samples containing three recordings were split into two samples of T and T-1. For the second sample, the recording at T-1 is T and T-2 is T-1. This arrangement helps us obtain a few samples from each class. In this setup, for patients with only two recordings, the first recording is considered as the recording as T-1 and the next recording as T. Both machine learning classifiers and deep neural networks were trained over this data and their prediction performances were evaluated.

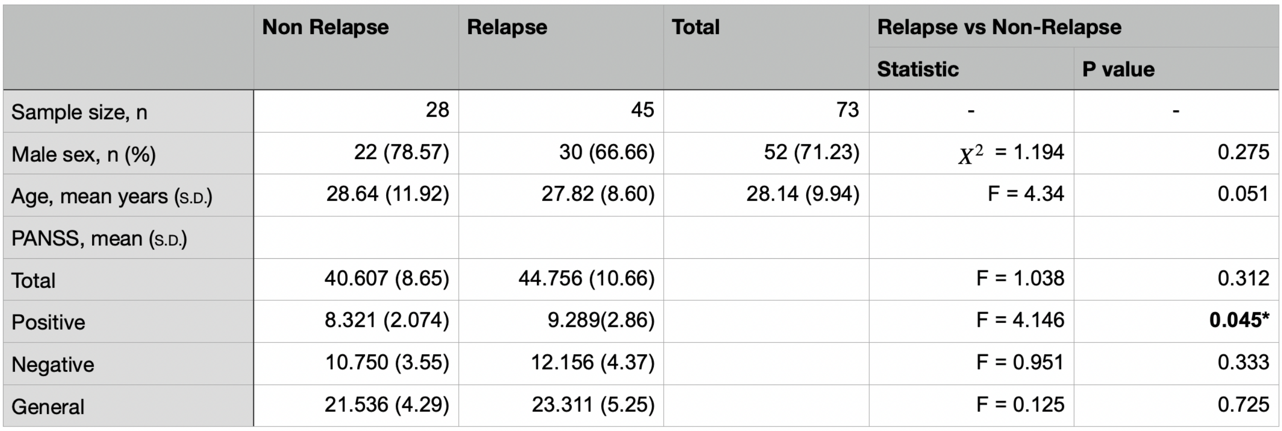
2.6.3 Diagnostic classifiers based on rate of change in acoustic features

Concerning our hypothesis that transition in acoustic features helps differentiate relapse patients from non-relapse patients, it might be interesting to look at the rate of change of the features over a period of time. To better capture the subtle variations in acoustic features over time, we propose a difference score method. In this method we calculate the absolute difference between the acoustic features at time points T and T-1. This will help us obtain a rate of change of the acoustic features over a period of time. As described in the section above, we used RFE to select 15 features that best account for classifying between the three classes based on the difference scores. The machine learning classifiers and deep neural network were trained on these scores and their performance were evaluated in a similar way as done with other procedures. It is to be noted that the CNN cannot be modelled here since the rate of change of features is a single entity over a period of time.

3 Results

3.1 Statistical Analysis

The schizophrenia-spectrum patients who have relapsed and patients who did not relapse did not differ significantly in age, sex, or total PANSS scores (see Table 3 1). However, the PANSS positive score had a significant difference. Based on the acoustic features, there was a statistically significant difference between the 'Non-Relapse' and 'Relapse' groups as determined by the independent samples t-test. Standard deviation of *HNRdBACF* and the mean *slope of the V5001500* differed significantly with a p-value of 0.001 and 0.005 respectively. There was a statistically significant difference between the 'Non-Relapse', 'Relapse' and the 'RelapsetoHealthy' groups based on the rate of change of acoustic features as determined by one-way ANOVA for three acoustic features, namely, standard deviation (S.D.) of *HNRdBACF* (p = 0.025), S.D. of *mfcc4v* (p = 0.011) and *equivalentSoundLevel* (p = 0.042). A Tukey post hoc test revealed that the 'Non-Relapse' group and the 'Relapse' group were statistically significantly different in S.D. of *HNRdBACF* (p = .018) and *equivalentSoundLevel* (p = .032). Based on S.D. of *mfcc4v* there was a statistically significant difference between the 'Non-Relapse' and 'RelapsetoHealthy' groups (p = .008) and the 'Relapse' and 'RelapsetoHealthy' groups (p = .042). There was no statistically significant difference between the 'Relapse' and 'Non-Relapse' group (p =0.985).



	Non Relapse	**Relapse**	**Total**	**Relapse vs Non-Relapse**	
				Statistic	**P value**
Sample size, n	28	45	73	-	-
Male sex, n (%)	22 (78.57)	30 (66.66)	52 (71.23)	X^2 = 1.194	0.275
Age, mean years (S.D.)	28.64 (11.92)	27.82 (8.60)	28.14 (9.94)	F = 4.34	0.051
PANSS, mean (S.D.)					
Total	40.607 (8.65)	44.756 (10.66)		F = 1.038	0.312
Positive	8.321 (2.074)	9.289(2.86)		F = 4.146	**0.045***
Negative	10.750 (3.55)	12.156 (4.37)		F = 0.951	0.333
General	21.536 (4.29)	23.311 (5.25)		F = 0.125	0.725

Figure 1: Demographics

These results highlight that the 'Non-Relapse' and the 'Relapse' groups are significantly different in two of these audio features, namely, the *HNRdBACF* (i.e. the Harmonics to Noise ratio) and the *equivalentSoundLevel* in decibels. The Harmonics to Noise ratio (HNR) gives out the relation of energy in harmonic components to energy in noise-like components [42], in simpler terms this parameter will account for the ratio of tone/excitement level in voiced segments to that of unvoiced segments of the speech samples which would directly relate to the monotonous nature in the speech of the subject. Similarly, the *equivalentSoundLevel* relates to the loudness in speech. This also relates to studies on speech in schizophrenic patients which is monotonous and in a lower tone than usual [40, 47]. The 'Non-Relapse' group and the 'RelapsetoHealthy' group (i.e. subjects that went into remission) differed significantly in the standard deviation of *mfcc4v*. The higher order Mel-Frequency Cepstral Coefficients (MFCC) describes the fine-grained energy distribution in the speech sample i.e., the phonetic content [42]. This makes sense in a way that there would not be much difference in voiced and unvoiced segments in the speech of both the groups since both the groups are not experiencing relapse during the time window of the interview. However, due to the previous relapse, the remission group may exhibit few phonetic disturbances in their speech compared to the group that has never experienced a relapse.

3.2 Speech based diagnostic classification: Relapse vs Non-Relapse

Previous studies have proven that simple classifiers such as SVM and RF can differentiate between schizophrenic patients from healthy controls based on acoustic features extracted from their speech. Hence, we were interested in knowing if these classifiers can deliver a similar performance in differentiating a relapse patient from a non-relapse patient based on their acoustic features. The models were trained and evaluated using 20-fold cross-validation with 142 samples out of which 120 were of Non-Relapse class and remaining 22 were of Relapse class. The SVM classifier trained on the acoustic data with class imbalance exhibited a mean accuracy of 84.3%. The similarly trained RF classifier generated a mean accuracy of 82.7%. At the first glance these accuracies might look promising, however, accuracies can be misleading when the test set is small and if there is an imbalance in the dataset. The imbalance in the classes introduces bias in the models' classification. The accuracies will be high since there are more samples of the majority class in the test set and the model is biased towards predicting the majority class. This results in high number of correct predictions for majority class which inturn boosts the accuracy even though the model exhibits a poor performance in classifying the minority class (which has less samples in the test set). To better evaluate the model we plotted the Receiver Operating Characteristics (ROC) graph and looked at the Area under the Curve (AUC) which is the measure of true positive rate vs the false positive rate. The SVM classifier had an AUC of 0.52 with a specificity of 100% and sensitivity of 0%. And the random forest classifier had an AUC of 0.59 with a specificity of 97.2% and a sensitivity of 5.8%. The AUC scores indicate that the models did not learn to classify among the classes and rather classifying them at random. We need to use specificity and sensitivity to evaluate the ability of the model to recognize non-relapse and relapse classes respectively. Both the models exhibit a very high specificity and very poor sensitivity. This signifies that the model has only learnt to predict the Non-Relapse class and has failed to learn any associative pattern regarding the Relapse class. To verify this we plotted the confusion matrix of both the models which is presented as Figure 2. It can be seen that all the samples in the test set are predicted to be of Non-Relapse class by the SVM model and only a very few samples were predicted to be of Relapse class by the random forest classifier. The models exhibit such behaviour due to the imbalance in the dataset. To counteract this effect, as described in the previous section, we use SMOTE to oversample the minority class. After augmentation, the number of samples in the minority class were equal to that of the majority class with 120 samples. Both the classifiers were again trained and validated using 20-fold cross validation on the new class-balanced dataset. The SVM classifier generated a mean accuracy of 64.14% with an AUC of 0.64 and the RF classifier generated a mean accuracy of 72.1% with an AUC of 0.52.

As noted, the accuracies of the models have dipped, but looking into the AUC scores (presented in Figure 3 and 4) and the confusion matrix (presented in Figure 5) it can be clearly argued that the models have performed comparatively better and the dip in accuracy is due to the models predicting less samples by chance compared to the previous training. Both the models have predicted more 'Relapse' class samples using the augmented dataset even though RF elicited a better prediction of the 'Non-Relapse' class in the previous training. We can note that the AUC of the SVM algorithm has increased drastically from 0.52 to 0.64. Contrastingly, the AUC of the RF algorithm has gone down from 0.59 to 0.52. However, both the AUC scores still point out that the performance of the models are poor even with a class balanced dataset. It can be concluded that the models were not successful

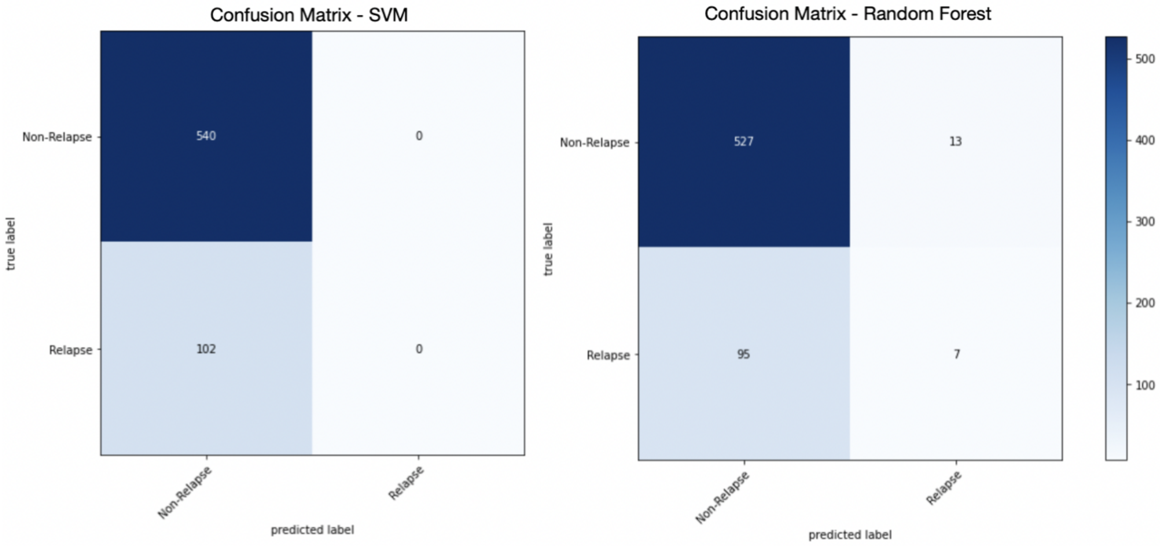


Figure 2: Confusion matrix of SVM and Random Forest models for ’Relapse’ vs ’Non-Relapse’ classification

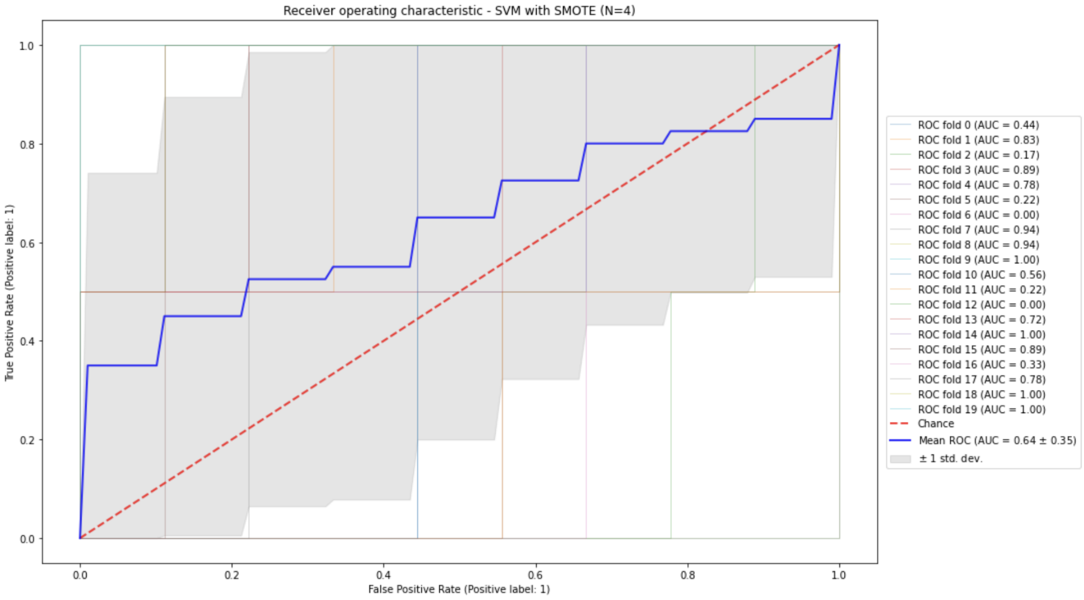


Figure 3: ROC of SVM model with SMOTE (N=4) with the AUC of each fold of cross validation and the mean AUC

in differentiating between a relapse vs a non-relapse patient based on the acoustic features derived from their speech recordings.

3.3 Speech transition based diagnostic classification

3.3.1 Classification between Relapse vs Non-Relapse samples

In order to differentiate Relapse patients from Non-Relapse patients we thought it is interesting to look at the transition of the acoustic features over time at the level of an individual patient. To realise this we arranged the data in such a way that each record corresponds to a single patient and the

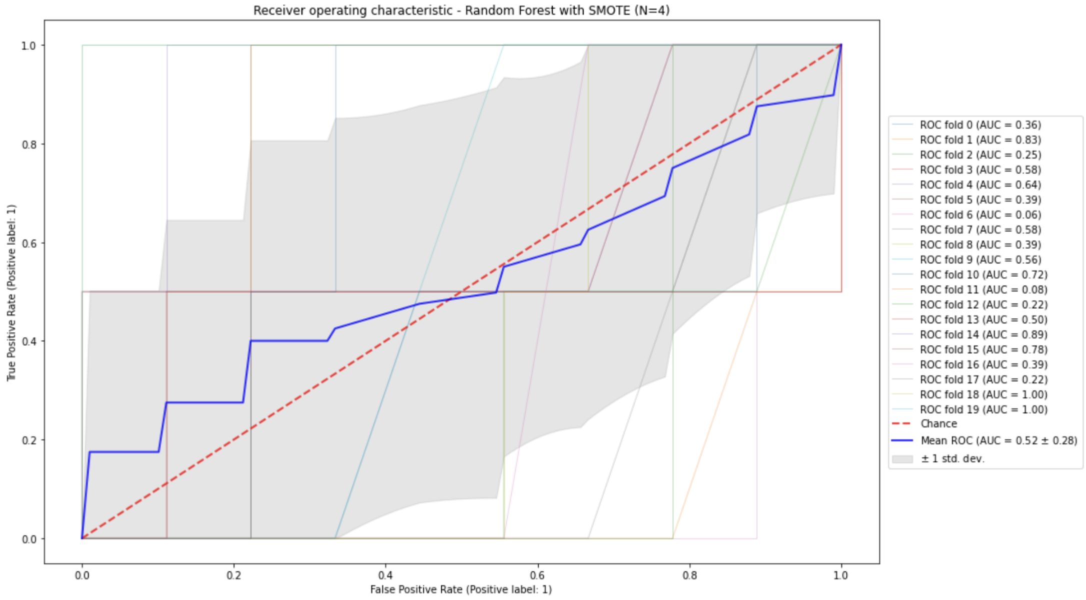


Figure 4: ROC of RF model with SMOTE (N=4) with the AUC of each fold of cross validation and the mean AUC

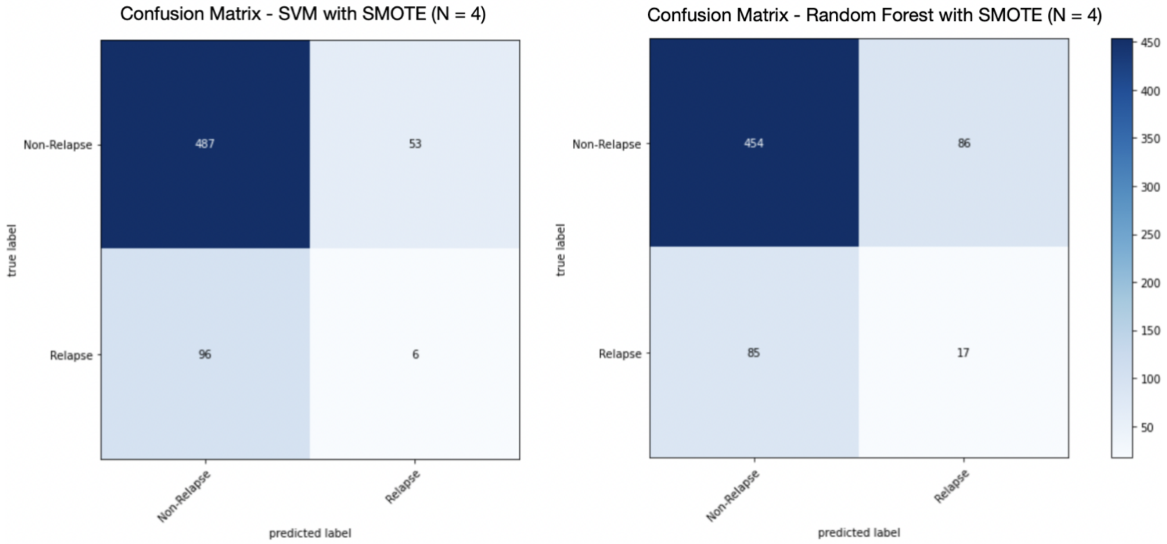


Figure 5: Confusion Matrix of SVM and RF model with SMOTE (N=4) for 'Relapse' vs 'Non-Relapse' classification

columns represent the selected acoustic features of three different speech recordings (i.e. 33 columns corresponding to 11 acoustic features derived from voice recordings from each of the three appointments [V1, V2 and V3], also referred here as transition dataset) as described in the methods section of this paper. Contrasting to the previous case, since each record corresponds to an individual patient, we have 44 samples for the 'Relapse' class and 24 samples for the 'Non-Relapse'. This was expected given the fact that the number of patients that relapsed is greater than that of the patients that did not experience a relapse.

Machine Learning Classifiers: The 20-fold cross validated SVM model had a mean accuracy of

63.33% and an AUC of 0.51 and the 20-fold cross validated RF model had a mean accuracy of 59.16% and an AUC of 0.57. Both models were trained on a class balanced dataset using SMOTE (N = 4). This data signifies that the model failed to learn an associative pattern to differentiate a Relapse patient from a Non-Relapse patient. To further evaluate the specifics of how well the model was able to distinguish these two classes based on the transition of the acoustic features over time, we looked into the confusion matrices of the models shown in the Figure 6 below. It can be clearly seen from the plots that the SVM model was highly biased to the relapse class and predicts almost all the datapoints as being a ’Relapse’ class. This can be due to the fact that the transition dataset has more number of ’Relapse’ class samples than ’Non-Relapse’ class samples. Comparitively, the similarly trained random forest model exhibits lesser bias to any particular class, however, it can also be concluded that the model did not learn to discriminate the classes accurately based on the fairly equal amount of misclassification seen on the confusion matrix.

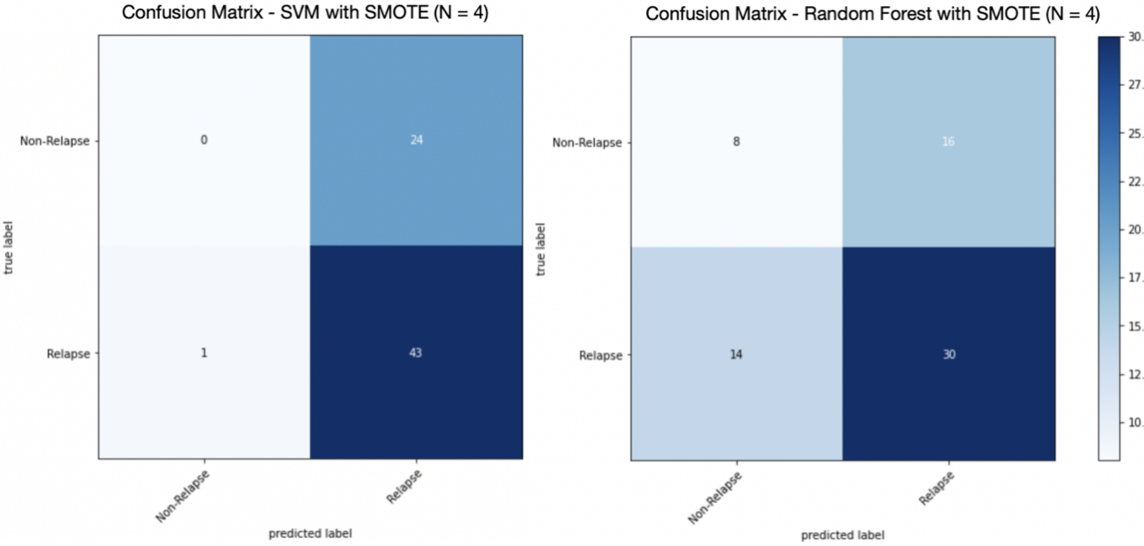


Figure 6: Confusion Matrix of SVM and RF model with SMOTE (N=4) for the transition dataset for ’Relapse’ vs ’Non-Relapse’ classification

Deep Learning Models: Since deep learning algorithms can cope up with some imbalance in the data [2], we trained the deep neural network model on the data with class imbalance. The 8-layered feed-forward neural network (NN) trained over 400 epochs with a batch size of 64 delivered an accuracy of 60.87% with an AUC of 0.46 with a maximum training accuracy of 97.8%. The CNN designed as described in the previous section, on top of the 8-layered neural network generated an accuracy of 52.17% with an AUC of 0.48 when trained over 350 epochs with a batch size of 64. The CNN attained a maximum accuracy of 93.3% during training. The specificity and the sensitivity of the deep neural network was 65.91% and 62.43% respectively. And the CNN model attained a specificity of 50% and a sensitivity of 48.72%. It can be interpreted from these findings that the deep learning model can cope with the imbalance in the dataset since both the sensitivity and the specificity are around the same range and do not exhibit a huge difference as seen on the ML models with class imbalance. However, the CNN model failed to distinguish between the classes and exhibited a poor performance. To examine the classification performance of the model, we plotted the confusion matrix which is presented in Figure 7. Examining the confusion matrices, it can be concluded that even though there isn’t a huge bias towards a particular class there was a fair amount of samples that has been misclassified. This highlights the fact that the models did not particularly learn any relative underlying pattern to accurately discriminate between the ’Relapse’ and ’Non-Relapse’ classes.

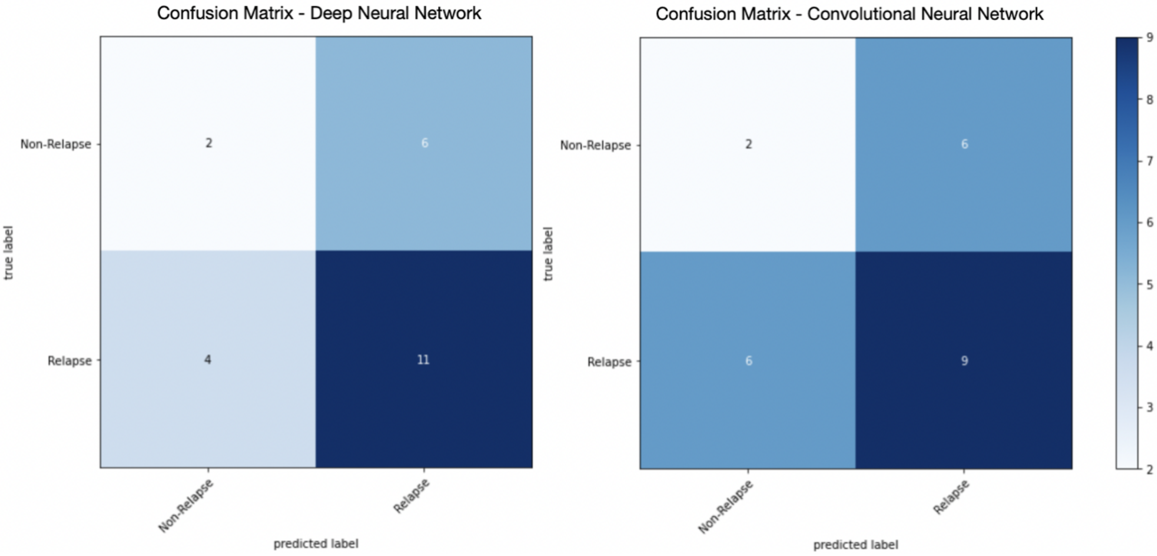


Figure 7: Confusion Matrix of NN and CNN for the transition dataset for ’Relapse’ vs ’Non-Relapse’ classification

In order to test if there are any significant differences in the performance of the models on a class balanced dataset, we trained both the models on the data set with data augmentation using SMOTE (N = 4). The deep neural network trained over 400 epochs with a batch size of 32 delivered an accuracy of 65.22% with an AUC of 0.57. The CNN generated an accuracy of 52.17% with an AUC of 0.57 when trained over 300 epochs with a batch size of 64. Comparatively, the deep neural network has an improved accuracy and the CNN model does not exhibit any changes in accuracy. Examining further on the AUC scores, it can be seen that the performance of the models have been quite better compared to the class imbalanced dataset. The specificity and the sensitivity of the deep neural network model was 66.67% and 72.59% respectively. Similarly, The CNN model attained a specificity of 53.85% and a sensitivity of 57.95%. There is a drastic improvement in the sensitivity of both the models, this is expected since there are more samples of the ‘Non-Relapse’ class for the model to learn from in the current dataset.

3.3.2 Classification between Relapse vs Non-Relapse vs Relapse-to-Healthy (Remission) samples

As the next step we would like to discriminate between three classes, i.e. ‘Non-Relapse’ vs ‘Relapse’ vs ‘RelapsetoHealthy’. The rearranged dataset with acoustic features at T and T-1 time frame consisted of 54 ‘Non-Relapse’ samples, 36 ‘Relapse’ samples and 12 ‘RelapsetoHealthy’ samples. Since this is a multi-class problem (3 classes), the best performance metric to analyse will be the weighted F1 scores rather than the AUC scores which is more suited for binary classification. Machine Learning Classifiers: The SVM model had a mean accuracy of 53.3% and a weighted F1 score of 0.38 over 10-fold cross-validation. Since the number of samples in the minority class is only 12, the maximum folds for cross validation is limited to 12. After the split, the data was augmented using SMOTE (N = 4) to balance the class distributions. Data Augmentation cannot be performed before splitting the dataset because augmented data in the test dataset will lead to data leakage into the model training, i.e., the augmented data will enable the model to classify similar data in the test set very well, which leads to false accuracies. The RF model exhibited a mean accuracy of 53.2% and a weighted F1 score of 0.46 in a similar cross-validation procedure.

Deep Learning Models: The 8-layered feed-forward neural network delivered an accuracy of 29.41% when trained over 350 epochs with a batch size of 32. The maximum training accuracy of this model was 83.3%. The CNN trained over 500 epochs with a batch size of 64 generated an accuracy of 32.35% and a maximum training accuracy of 82.4%. Both the models did not learn any patterns in the data that would help classify the samples into different classes. Neither the ML classifiers nor the DL models obtained any significant results over the dataset with the current arrangement (with T-1 and T time-points for ‘Non-Relapse’, ‘Relapse’, and ‘RelapsetoHealthy’ classes). When the test data was introduced, the models exhibited chance accuracies (i.e. SVM - 53.3%, random forest - 53.2%, deep neural network - 29.41%, and CNN - 32.35%). Due to this no further evaluations were performed on the performance metrics of the model.

3.3.3 Diagnostic classifiers based on rate of change in acoustic features; Relapse vs Non-Relapse vs Relapse-to-Healthy (Remission) samples:

As an extension of our second hypothesis, we decided to examine the rate of change of acoustic features over a period of time. To facilitate this, we obtained the difference scores from the acoustic features of the patients as described in the methods. The Recursive Feature Elimination algorithm was used to select 15 features that are more relevant in classifying the samples into their respective classes. The resulting subset of 15 acoustic features are presented in Table 2 in the previous section.
Machine Learning Classifiers: The SVM model had a mean accuracy of 75.8% and a weighted F1 score of 0.66 in 10-fold cross-validation. The RF model had a mean accuracy of 76.1% and a weighted F1 score of 0.66. Both the classifiers were trained over class balanced dataset using SMOTE (N = 4) and exhibit higher accuracies compared to the other classifications based on the transition of acoustic features. To further examine the performance of the models we plotted the confusion matrices with the sum of all classification across each fold of cross validation.

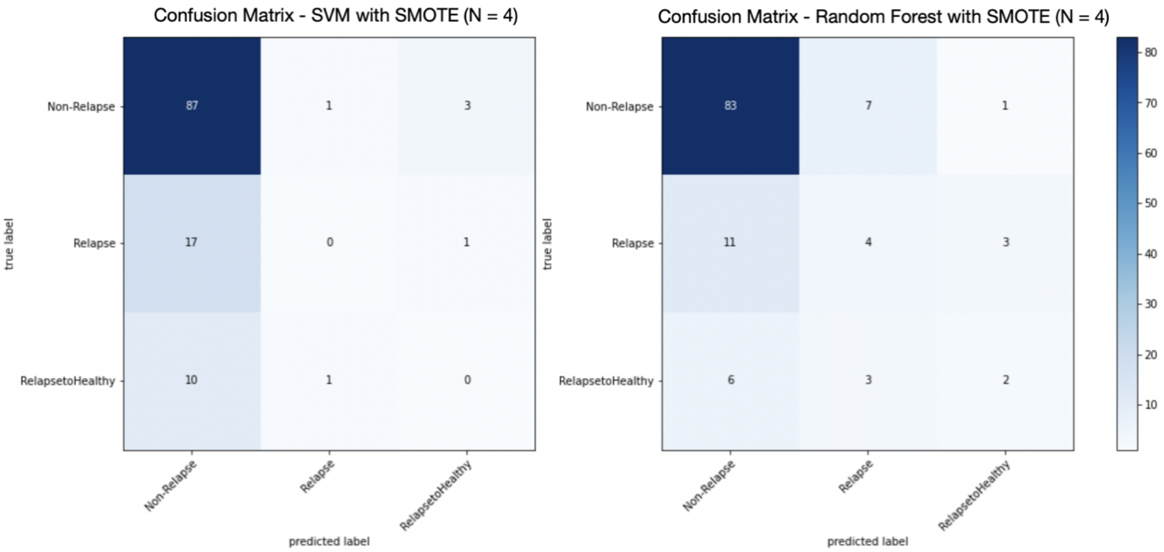


Figure 8: Confusion Matrix of SVM and RF model with SMOTE (N=4) for the difference score dataset for ’Non-Relapse’ vs ’Relapse’ vs ’RelapsetoHealthy’ classification

From the confusion matrix in Figure 8 presented above, it is clear that even though both the models have exhibited a slightly better accuracy compared to the other classifiers based on transition of acoustic features, the performance of the models cannot be considered significant. The models have learned to classify 'Non-Relapse' class fairly well, however, both the other classes were classified randomly.

Deep Learning Models: The 7-layer feed-forward NN that was trained over 300 epochs with a batch size of 64 delivered an accuracy of 70% with a maximum training accuracy of 98.7%. Since this is a multi-class classification problem, we plotted the ROC and obtained the AUC using the 'One vs Rest' approach, i.e. each class is evaluated by considering the other two classes as a single class which essentially considers it as a binary classification problem. The model scored a multi-class mean AUC of 0.61. The per-class AUC scores are as follows: 'Non Relapse' class vs rest: 0.66 ,'Relapse' class vs rest: 0.60, 'RelapsetoHealthy' class vs rest: 0.58.

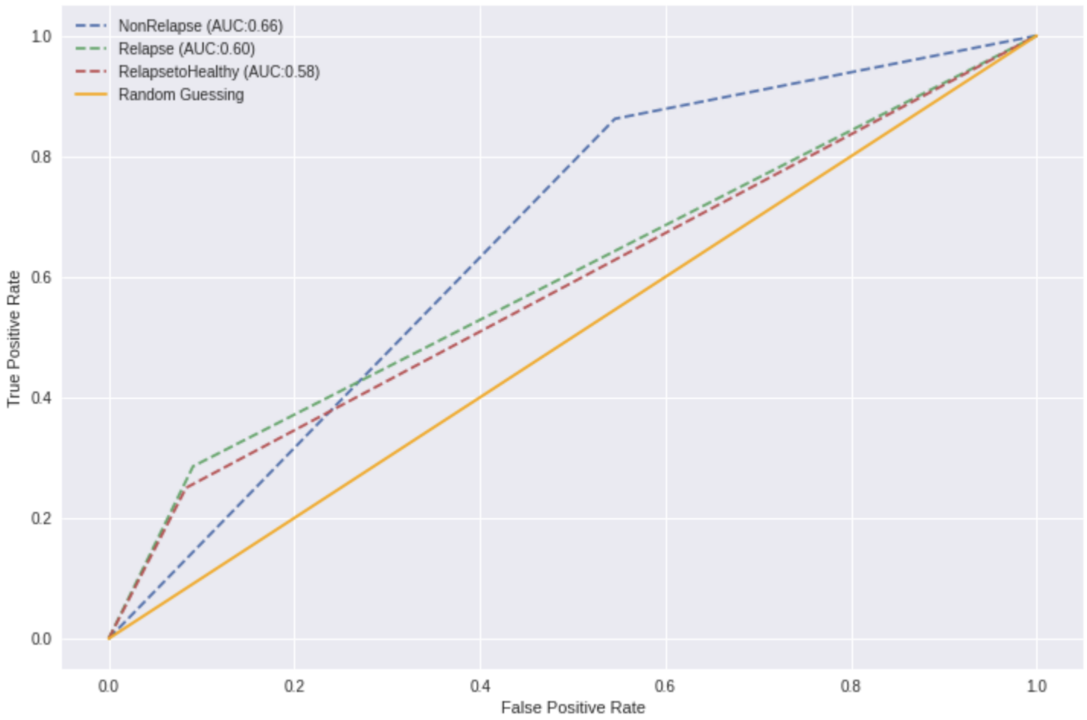


Figure 9: Multi-class ROC plot with 'One vs Rest' AUC of Deep Neural Network for difference score dataset for 'Non-Relapse' vs 'Relapse' vs 'RelapsetoHealthy' classification

The multi-class ROC plot with 'One vs Rest' AUC of the different classes of the model is given in Figure 9. The confusion matrix of this model is presented in Figure 10. It can be noted that the model has not learnt to discriminate among the three classes based on the difference score dataset. The deep neural network was also trained over a class-balanced difference score dataset using SMOTE with N = 4. The model yielded an accuracy of 77.5% with a multi-class mean AUC score of 0.68 when trained over 300 epochs and a batch size of 64. The model generated a maximum training accuracy of 99.4%. In order to examine the model, we plotted a multi-class ROC with 'One vs Rest' approach for the AUC as shown in Figure 11. The 'Non-Relapse vs Rest' curve obtained an area of 0.77 and the 'Relapse vs Rest' curve obtained an area of 0.81. Both of the curves exhibit significant performance. However, the 'RelapsetoHealthy vs Rest' curve only exhibited an area of 0.47 of the graph.

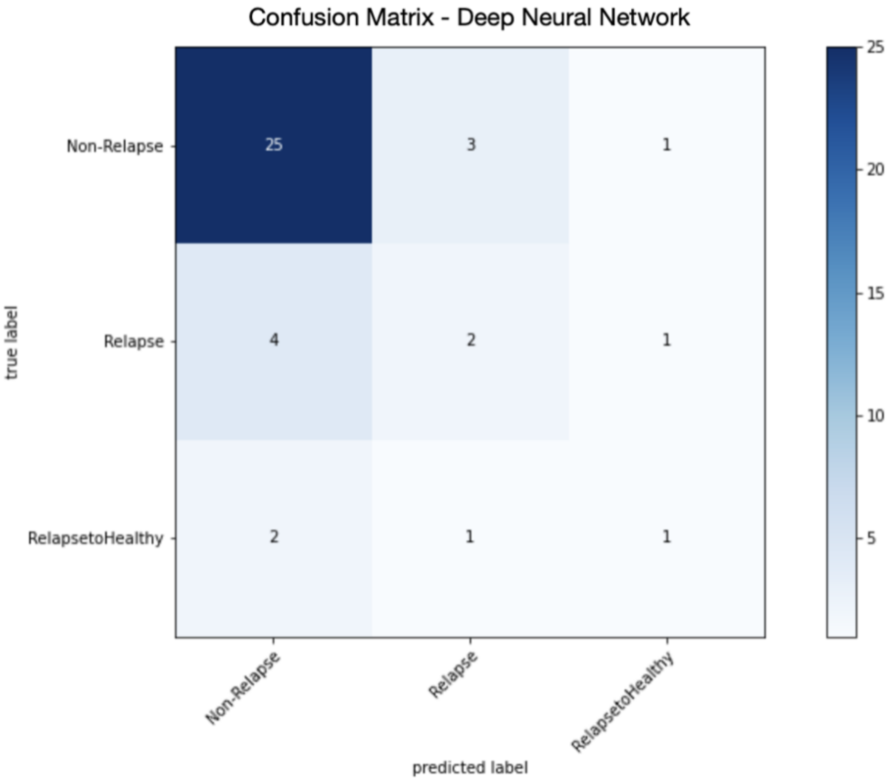


Figure 10: Confusion Matrix of Deep Neural Network for the difference score dataset for 'Non-Relapse' vs 'Relapse' vs 'RelapsetoHealthy' classification

This is expected considering the fact that the number of samples of the 'RelapsetoHealthy' class only accounts for half of that of the 'Relapse' class.

Since the ROC plot shows promising results, we further analysed the performance of the model by plotting the confusion matrix. The confusion matrix is presented in Figure 12. From the figure, it can be interpreted that the model can differentiate 'Non-Relapse' and 'Relapse' class from the dataset with a very good precision. Out of 29 'Non-Relapse' samples, 26 samples were correctly classified by the model. Similarly, 5 out of 7 'Relapse' samples were identified by the model. The deep neural network has learnt implicit patterns to identify these classes and discriminate between them. However, the model failed to identify any relevant patterns for the 'RelapsetoHealthy' class. Based on these data, it can be concluded that our hypothesis that "Changes in the acoustic features of an individual over time can help us predict a transition into a relapse". It is also noteworthy that deep neural networks provide a better performance as compared to the machine learning classifiers in classifying between a relapse patient and a non-relapse patient using the difference scores, i.e., the rate of change of acoustic features over a period of time.

To further analyse how well the models can discriminate between classes, the least populated class 'RelapsetoHealthy' was excluded and models were trained on class-balanced dataset to classify only between 'NonRelapse' and 'Relapse' labels. The RF classifier with 10- fold cross- validation generated a mean AUC of 0.73. The deep NN achieved an AUC of 0.75. Both these scores represents a good discrimination of features between the class labels. Overall, these AUC scores demonstrates a pretty good discrimination. However, the cumulative confusion matrix of both the models show that there is a fair amount of misclassification of the 'Relapse' class as presented in the figure below

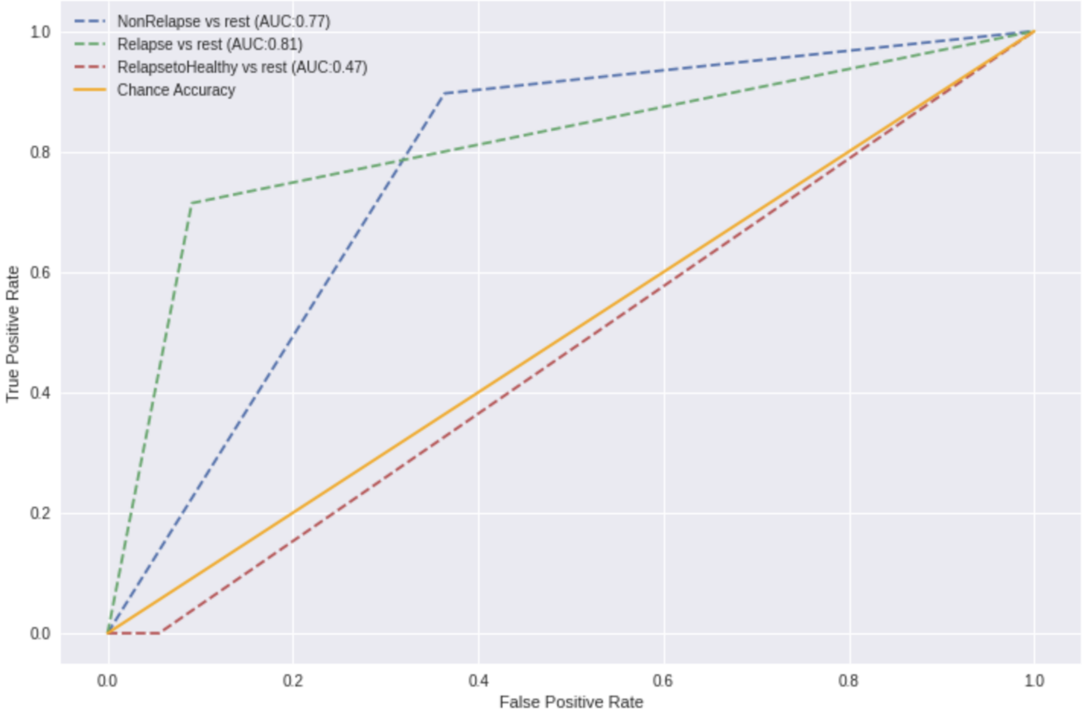


Figure 11: Multi-class ROC plot with ‘One vs Rest’ AUC of Deep Neural Network for difference score dataset with SMOTE (N = 4) for ’Non-Relapse’ vs ’Relapse’ vs ’RelapsetoHealthy’ classification

(see Figure 13). A summary of the performance scores of the models for all the predictive analysis performed is presented in Table 4 **??**. A good classification performance score is highlighted in bold. All the analysis presented are done on class balanced data until it is specified explicitly in the table.

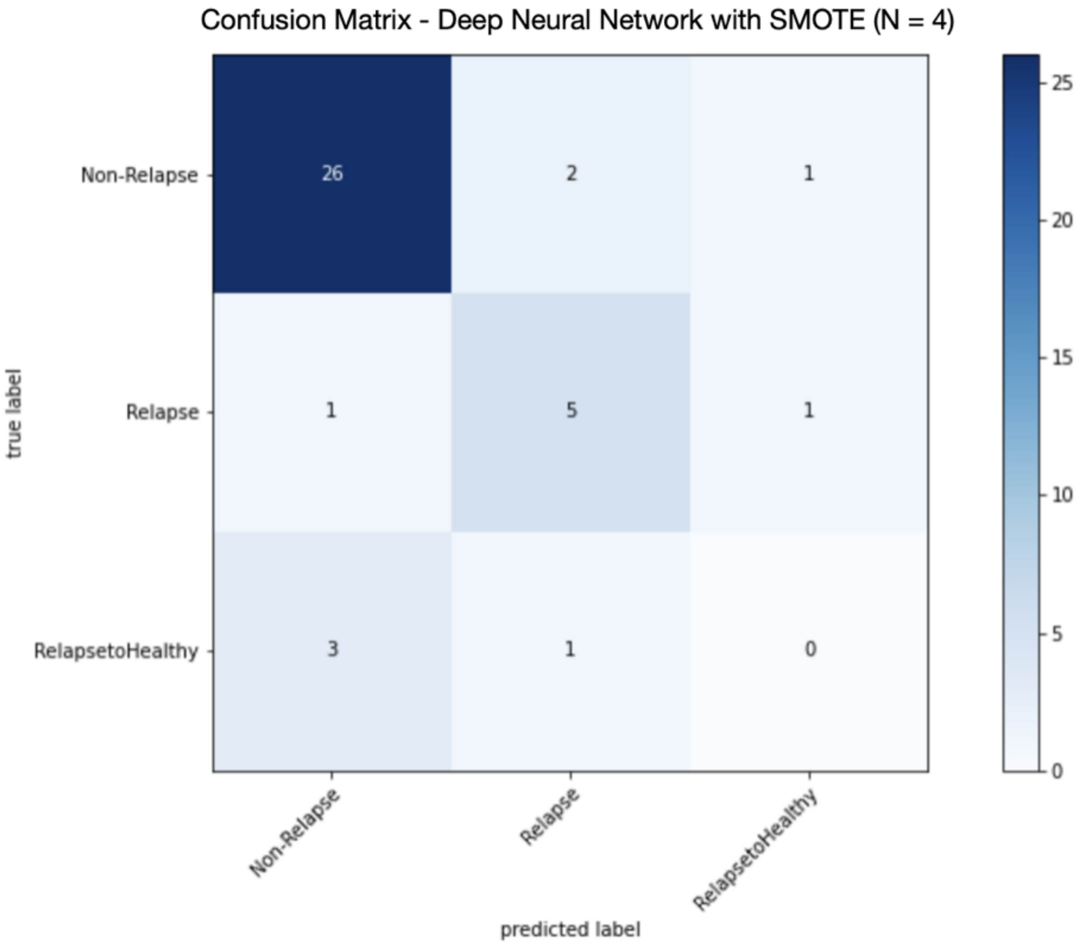


Figure 12: Confusion Matrix of Deep Neural Network for the difference score dataset with SMOTE (N = 4) for 'Non-Relapse' vs 'Relapse' vs 'RelapsetoHealthy' classification

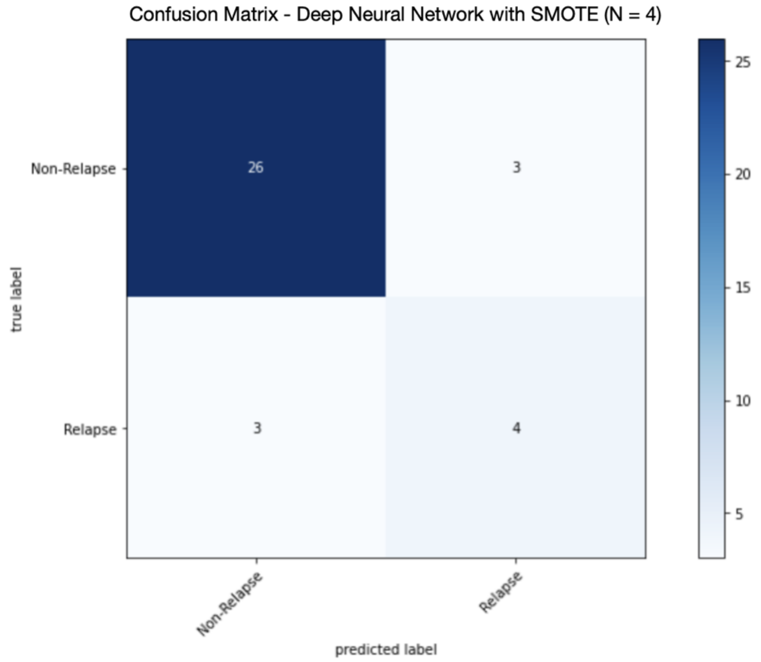


Figure 13: Confusion Matrix of Deep Neural Network for the difference score dataset on binary classification between 'Non-Relapse' and 'Relapse' with SMOTE (N = 4)

Table 4. Summary of performance scores of the models

Predictive analysis	**Model**	**AUC (binary)**	**Weighted F1 score**	**One vs Rest AUC**		
				Non-Relpase	**Relapse**	**RelapsetoHealthy**
Relapse vs Non-Relapse (without SMOTE)	SVM	0.52	-	-	-	-
	Random Forest	0.59	-	-	-	-
Relapse vs Non-Relapse	SVM	0.64	-	-	-	-
	Random Forest	0.52	-	-	-	-
Speech transition based diagnostic classification (Relapse vs Non-Relapse)	SVM	0.51	-	-	-	-
	Random Forest	0.57	-	-	-	-
	Deep Neural Network	0.46	-	-	-	-
	CNN	0.48	-	-	-	-
Speech transition based diagnostic classification (Relapse vs Non-Relapse vs RelapsetoHealthy)	SVM	-	0.38	-	-	-
	Random Forest	-	0.46	-	-	-
	Deep Neural Network	0.18	-	-	-	-
	CNN	0.21	-	-	-	-
Classification based on rate of change in acoustic features (Relapse vs Non-Relapse vs RelapsetoHealthy)	SVM	-	0.66	-	-	-
	Random Forest	-	0.66	-	-	-
	Deep Neural Network (without SMOTE)	0.61	-	0.66	0.60	0.58
	Deep Neural Network	0.68	-	**0.77**	**0.81**	0.47

4 Discussion

In our study, we hypothesised that the acoustic features of voice recordings of schizophrenic patients can be a reliable biomarker for predicting if a patient is experiencing/will experience a relapse or not. Additionally, we also hypothesised that changes in the acoustic features of an individual over time can help us predict a transition into a relapse. In order to test these hypotheses we devised SVM and random forest machine learning classifiers and a few deep neural networks and convolutional neural networks and calculated their performance in classifying the dataset containing acoustic features of patient's speech recordings into 'Non-Relapse' and 'Relapse' classes.

4.1 Key findings and Interpretation

We obtained 11 acoustic features that highly account for identifying schizophrenic patients [44] from the speech recordings of the patients and calculated the classification performance of machine learning algorithms in classifying a 'Relapse' vs a 'Non-relapse' class on a dataset containing 180 'Non-Relapse' samples and 34 'Relapse' samples. The machine learning models that were trained to classify the samples into 'Relapse' and 'Non-Relapse' classes based on the acoustic features derived from the speech of the patients exhibited very poor classification performance. Based on the confusion matrices (see Figure 2) of the models over 20-fold cross-validation on a class-balanced dataset, it can be visualised that the model learnt to predict a 'Non-Relapse' class (i.e. the majority class) fairly well but the performance in classifying a 'Relapse' class (i.e. the minority class) is very poor. Based on these results we reject our first hypothesis that a relapse can be reliably predicted based on the acoustic features derived from the speech of a schizophrenic patient.

To test whether a transition into a relapse can be predicted using the changes of acoustic features over a period of time, we trained both machine learning and deep learning models over the acoustic features data. Each record in this dataset corresponds to a single patient with acoustic features of all the three speech recordings as the data in each record. Both the ML and DL models demonstrated a poor performance over the test dataset. The AUC scores of the models using a class-balanced dataset were in the 0.5 to 0.6 range. The confusion matrices helped visualise the distribution of correct vs incorrect classification and it can be clearly interpreted that the model did not learn any underlying pattern in the dataset.

An extension of the above analysis includes labelling a new class 'RelapsetoHealthy' where the patients that had a transition from a relapsed state to a non-relapse state were addressed. This is a multi-class classification problem with 3 classes to be distinguished. Both the machine learning classifiers and the deep neural networks failed to learn the associations to classify the classes. Since all the models exhibited extremely poor accuracies, no further analysis was carried out on the performance of these models such as AUC scores or confusion matrices.

Finally, we used the difference scores of the acoustic features to differentiate between relapse and non-relapse subjects. We thought that even though the acoustic features of the subjects are highly heterogeneous, the rate of change of these acoustic features within the individuals experiencing a transition may exhibit similar signatures. Both the SVM and the random forest classifiers that were trained on the difference scores of the acoustic features using 10-fold cross-validation failed to reliably distinguish the samples to their corresponding classes.

However, the 7-layer feed-forward deep neural network that was trained with 300 epochs and a batch size of 64 on the difference score dataset learnt to predict the ‘Non-Relapse’ and ‘Relapse’ samples in the dataset. The model correctly classified 26 out of 29 ‘Non-Relapse’ samples and 5 out of 7 ‘Relapse’ samples. The model did not learn to classify the ‘RelapsetoHealthy’ class, but this is justifiable since there were only 7 records of this class in the training dataset for the model to learn from. Based on this result, we can accept the hypothesis that changes in the acoustic features of an individual over time can help us predict a transition into a relapse.

Contrastingly, when the least populated class ‘RelapsetoHealthy’ was excluded and the deep neural network model was trained on class-balanced dataset to classify only between ‘NonRelapse’ and ‘Relapse’ labels it exhibited a fair amount of misclassification for the ‘Relapse’ class. This implies that there are subtle variations between the ‘Relapse’ class and the ‘RelapsetoHealthy’ class which helped the previous model learn to classify the data with good precision. The ‘Non-relapse’ class performance was the same as the previous model.

An interesting interpretation from our results is that, based on the distribution of the samples, the models were fairly able to predict both ‘Relapse’ and ‘Non-Relapse’ classes. In the figure below (see Figure 14), we can see that during different analyses, a random forest classifier was able to predict different classes. This may imply that the models can find an implicit association for both the classes but they are limited by the distribution of the classes in the dataset. Our results also highlight that the deep NNs perform fairly better compared to their ML alternatives. It can be understood that the DL models demonstrate better performance since the complex nuances in the data can be learnt better by the DL models. DL models require more data to learn the class based associations. However, what is surprising is that the DL models exhibited this performance with the same data as the ML models. This result signifies that DL models are a better fit for our problem.

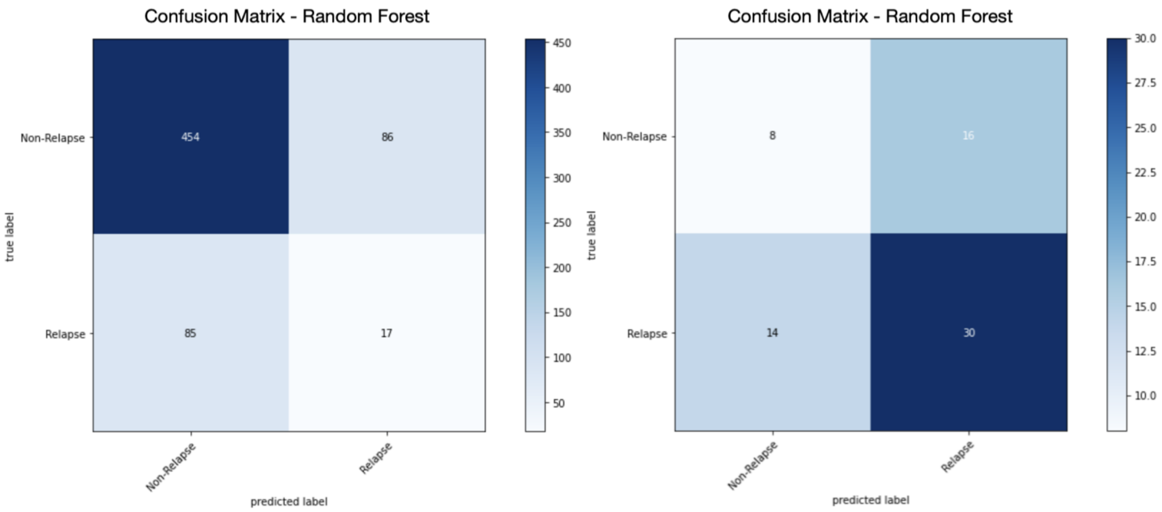


Figure 14: Confusion Matrix of Random Forest model with different arrangements of data for ’Relapse’ vs ’Non-Relapse’ classification

There are a few cohort studies that define the characteristics of relapse patients in schizophrenia [48, 49], however, studies based on predicting a relapse are pretty rare. One such study uses deep learning to predict psychotic relapse using behavioural changes based on mobile sensing [23]. The approach used to predict a relapse is highly different from our study, but the results are fairly similar. The sample size of the study was quite large, the study had a 144 dimensional hourly mobile sensing data of 63 subjects for each day for 28 days. The Neural Network demonstrated a very low F2 score, which implies a poor performance in classifying the data. Contrastingly, one of our deep neural networks demonstrated a very good performance in classifying relapse patients in our dataset. Our results with predicting the transition into a relapse based on rate of change of acoustic features over a period of time matches with one such study that predicts the First Episode of Psychosis (FEP) [24]. However both the studies look at different aspects of predicting schizophrenia, these two studies are highly comparable. It demonstrated an accuracy of 77.5% in detecting schizophrenia in FEP which is similar to the results obtained by our study.

4.2 Limitations of the study

One interesting limitation of this study is that the information about the exact point in which the patients' relapse is not taken into account. The patient may relapse in a couple of days after their appointment in the V2 time frame, or the patient may relapse after two months after their appointment in the V2 time frame. Both these scenarios will lead to encoding the speech recording at that appointment to be labelled as 'Relapse' class. Accounting for this paradigm may help improve the accuracy of the models.

The main limitation of this study revolves around the scarcity of the data available for the analysis and oversampling of the minority class to obtain a class-balanced dataset with a highly heterogeneous and noisy dataset. Classification algorithms deliver a poor performance in both these scenarios. Also, differentiating a relapse patient from a non-relapse patient within a group of schizophrenic patients is a very complex problem. In previous studies, both the classes belong to a distinctive group, i.e., one set of people are healthy controls and the other group is Schizophrenic. However, in our problem, both the groups belong to the same overall group, i.e., they are Schizophrenic patients. This implies that both the sub-groups of people will be experiencing some physical impairments in the language subsystems in the brain and hence express difficulties in language [41].

4.2.1 Scarcity of Data

Considering the complexity of the problem, data of 73 subjects is extremely low for the classifiers to reliably discriminate between different labels. Also, there is a huge imbalance in the dataset with the minority class ('Relapse') accounting for only 15.8% of the total dataset. This can be contradictory since the demographics represent a larger number of subjects identified to be relapsed. However, most of the subjects who relapsed failed to make it to the appointment. Therefore, the available recordings of these subjects will be from previous appointments during which they had not relapsed yet. Due to this, the speech recordings at these appointments will be labelled as the 'NonRelapse' class.

Since the main interests of the study are time-based transitions within an individual's speech, subjects with single recordings were excluded for certain analysis such as speech transition based diagnostic classification and diagnostic classification based on rate of change of acoustic features over a period of time. After excluding participants with single recording we were left with the data of 68 subjects. The impact of this reduction in the dataset cannot be comparatively measured since both the studies are quite different in their approaches.

4.2.2 Oversampling the minority class

We used SMOTE to oversample the underrepresented class to promote the classification algorithm to recognise the minority class during validation. SMOTE oversamples the minority classes by generating random synthetic data points between the domain gaps of the existing minority samples. SMOTE implicitly assumes that the distribution of per class instances are sufficiently homogenous in some domain space around the minority class [45]. While this approach works well for a vast majority of datasets, our data is highly heterogeneous and noisy. The data instances of each class are highly intertwined in our dataset which makes it difficult to randomly sample an instance for a specific class.

4.2.3 Deriving methods of the acoustic parameters

The semi-structured interview used during the recording of the speech is on the whole a very good quantitative measure of language of the patient since it involves thought provoking neutral ended questions such as *"What would you do if you won a lottery of 1 million euros?"*. Using such paradigms we can coherently relate the language of the patients to their underlying cognitive thought patterns. However, the acoustic parameters (i.e. eGeMAPS features) derived from the speech is actually the mean value of the parameters for the entirety of the interview. This measuring method may nullify any specific characteristic patterns in the acoustic feature values for certain thought provoking questions due to their respective values during general interactions during the interview. It is also a relatively difficult process to compute and analyse the acoustic parameters at a non-average level since the interviews are semi-structured and vary from 5 to 30 minutes in length within individuals. However, a time-window based computation of these parameters (for e.g. a window of 5 minutes) might help us yield new insights.

4.3 Future Work

4.3.1 Data collection and synthetic data generation

As a future trend in this study, it is vital to collect more data. This is simultaneously being facilitated by the ongoing HAMLETT study since the dataset obtained for this study is based on the HAMLETT study [50]. As a part of this study, we searched for online speech and GeMAPS datasets of schizophrenia patients to extend our dataset. However, no such datasets are available. One approach for synthesising more data based on the original dataset despite the heterogeneous and noisy nature of the data will be to use Conditional Generative Adversarial Networks (Ct GANs) [51]. Previous studies have successfully augmented noisy speech data based on Ct GANs which tends to be a promising direction [52]. Using Ct-GANs for data augmentation is out of scope for this project since GANs need larger samples of data to learn from during the training phase. GANs need several thousands of samples for the algorithm to perform well [53]. Data collected in the future may pave a path to successfully augment more data using such complex algorithms.

4.3.2 Extending the data over multilingual datasets

In order to facilitate a larger dataset, it would be a good option to explore multi-lingual datasets. Studies confirm that salient features in the acoustic parameters can still be relevant across different languages [54]. It is indicated that there are similarities between entirely different languages in the way that they manifest depression [47]. This can be relevant for other mental disorders as well which could help in the collection of vast datasets.

4.3.3 More complex models

Another interesting direction in which this study can be extended is by devising more complex models for classifying the data. Hidden Markov Models (HMM) is one such algorithm which can capture transitions in state using the previous state probabilities. This model is advised given the fact that the acoustics data over a period of time agrees with the basic assumptions of a HMM such as markovianity, output independence, and stationarity. HMM are also feasible with a limited amount of data, however, they require equally distributed samples among all the states. Hence, on further data collection this might be an interesting model to look at.

As it is evident from our results that CNN models help capture the temporal-spatial differences between the acoustic features to an extent, a composite model of CNN and HMM can thus be used to enhance the classification performance. Such hybrid models have been used previously to solve complex temporal sequence transition problems [55]. Other deep NN architectures such as Recurrent Neural Networks (RNN) and Long-Short Term Memory (LSTM) NNs that provide certain memory mechanisms to classify the current state based on a previous state in time would also be recommended provided there is more data for the models to learn from.

4.4 Conclusion

In conclusion, based on the current dataset, GeMAPS acoustic features extracted from voice recordings of patients suffering from schizophrenia cannot be a reliable biomarker for predicting if a patient will undergo relapse or not. A more robust dataset is required to train the models for the underrepresented classes which may increase the performance of the classifiers. However, analysis of the rate of change in the acoustic features within individuals over a period of time does prove to be a promising direction in predicting a relapse. Although the machine learning classifiers devised to predict the onset of a relapse based on the difference scores exhibited only a fairly good classification, the performance of these classifiers were significantly better than that of the other models used for evaluation. The deep neural network trained on this data demonstrated a very good classification between the 'Non-Relapse' and 'Relapse' data, in view of which we chose to accept the hypothesis that the changes in the acoustic features of an individual over time can help us predict a transition into a relapse. It is also notable that the deep NN performed fairly better than that of the ML classifiers even though NNs require more data to learn as compared to ML models. Hence, deep NNs are better suited for the complexity of this problem. In view of the shortcomings of this study, in the future work, we would include the data of when exactly in a period of three months a patient has relapsed and we will collect more voice data from schizophrenic patients. In addition, we will further improve the classification performance by implementing more complex models that better fit the complexity of the problem provided we collect a more robust dataset. Future research is intended to further investigate change in acoustic features of a schizophrenic patient over time.

Bibliography

[1] G. Bedi, G. A. Cecchi, D. F. Slezak, F. Carrillo, M. Sigman, and H. de Wit, "A window into the intoxicated mind? speech as an index of psychoactive drug effects," *Neuropsychopharmacology*, vol. 39, no. 10, p. 2340–2348, 2014.

[2] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, "The DSM-5: Classification and criteria changes," *World Psychiatry*, vol. 12, no. 2, p. 92–98, 2013.

[3] "Dementia praecox or the group of schizophrenias," *Journal of the American Medical Association*, vol. 145, no. 9, p. 685, 1951.

[4] E. J. Tan, E. Neill, and S. L. Rossell, "Assessing the relationship between semantic processing and thought disorder symptoms in schizophrenia," *Journal of the International Neuropsychological Society*, vol. 21, no. 8, p. 629–638, 2015.

[5] M. Alpert, R. J. Shaw, E. R. Pouget, and K. O. Lim, "A comparison of clinical ratings with vocal acoustic measures of flat affect and alogia," *Journal of Psychiatric Research*, vol. 36, no. 5, p. 347–353, 2002.

[6] T. Wensing, E. C. Cieslik, V. I. Müller, F. Hoffstaedter, S. B. Eickhoff, and T. Nickl-Jockschat, "Neural correlates of formal thought disorder: An activation likelihood estimation meta-analysis," *Human brain mapping*, vol. 38, no. 10, pp. 4946–4965, 2017.

[7] S.-J. Kim, J.-C. Shim, B.-G. Kong, J.-W. Kang, J.-J. Moon, D.-W. Jeon, S.-S. Jung, B.-J. Seo, and D.-U. Jung, "The relationship between language ability and cognitive function in patients with schizophrenia," *Clinical Psychopharmacology and Neuroscience*, vol. 13, no. 3, p. 288, 2015.

[8] M. A. Covington, C. He, C. Brown, L. Naçi, J. T. McClain, B. S. Fjordbak, J. Semple, and J. Brown, "Schizophrenia and the structure of language: the linguist's view," *Schizophrenia research*, vol. 77, no. 1, pp. 85–98, 2005.

[9] B. A. Maher, T. C. Manschreck, J. Linnet, and S. Candela, "Quantitative assessment of the frequency of normal associations in the utterances of schizophrenia patients and healthy controls," *Schizophrenia Research*, vol. 78, no. 2-3, p. 219–224, 2005.

[10] G. R. Kuperberg, N. Delaney-Busch, K. Fanucci, and T. Blackford, "Priming production: Neural evidence for enhanced automatic semantic activity preceding language production in schizophrenia," *NeuroImage: Clinical*, vol. 18, p. 74–85, 2018.

[11] K. Dwyer, A. S. David, R. McCarthy, P. McKenna, and E. Peters, "Linguistic alignment and theory of mind impairments in schizophrenia patients' dialogic interactions," *Psychological Medicine*, vol. 50, no. 13, p. 2194–2202, 2019.

[12] J. A. Willits, T. Rubin, M. N. Jones, K. S. Minor, and P. H. Lysaker, "Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia," *Schizophrenia Research*, vol. 197, p. 365–369, 2018.

[13] A. Ozcan, G. Kuruoglu, K. Alptekin, and S. Ozsoy, "An analysis of complex sentence structures in patients with schizophrenia," *Global Journal of Foreign Language Teaching*, vol. 6, no. 4, p. 228–235, 2016.

[14] L. van Schuppen, K. van Krieken, and J. Sanders, “Deictic navigation network: Linguistic viewpoint disturbances in schizophrenia,” *Frontiers in Psychology*, vol. 10, 2019.

[15] G. R. Kuperberg, “Language in schizophrenia part 1: An introduction,” *Language and Linguistics Compass*, vol. 4, no. 8, p. 576–589, 2010.

[16] T. Ditman and G. R. Kuperberg, “Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia,” *Journal of Neurolinguistics*, vol. 23, no. 3, p. 254–269, 2010.

[17] E. J. Clemmer, “Psycholinguistic aspects of pauses and temporal patterns in schizophrenic speech,” *Journal of Psycholinguistic Research*, vol. 9, no. 2, p. 161–185, 1980.

[18] A. S. Cohen, K. R. Mitchell, and B. Elvevåg, “What do we really know about blunted vocal affect and alogia? a meta-analysis of objective assessments,” *Schizophrenia research*, vol. 159, no. 2-3, pp. 533–538, 2014.

[19] B. Elvevåg, P. W. Foltz, D. R. Weinberger, and T. E. Goldberg, “Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia,” *Schizophrenia research*, vol. 93, no. 1-3, pp. 304–316, 2007.

[20] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, C. M. Corcoran, and et al., “Automated analysis of free speech predicts psychosis onset in high-risk youths,” *npj Schizophrenia*, vol. 1, no. 1, 2015.

[21] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, and G. A. Cecchi, “Prediction of psychosis across protocols and risk cohorts using automated language analysis,” *World Psychiatry*, vol. 17, no. 1, pp. 67–75, 2018.

[22] N. B. Mota, M. Copelli, and S. Ribeiro, “Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance,” *npj Schizophrenia*, vol. 3, no. 1, pp. 1–10, 2017.

[23] B. Lamichhane, J. Zhou, and A. Sano, “Psychotic relapse prediction in schizophrenia patients using a mobile sensing-based supervised deep learning model,” 2022.

[24] A. Figueroa-Barra, D. Del Aguila, M. Cerda, P. A. Gaspar, L. D. Terissi, M. Durán, and C. Valderrama, “Automatic language analysis identifies and predicts schizophrenia in first-episode of psychosis,” *Schizophrenia*, vol. 8, no. 1, pp. 1–8, 2022.

[25] F. Carrillo, N. Mota, M. Copelli, S. Ribeiro, M. Sigman, G. Cecchi, and D. Fernandez Slezak, “Automated speech analysis for psychosis evaluation,” *Lecture Notes in Computer Science*, p. 31–39, 2016.

[26] J. Fu, S. Yang, F. He, L. He, Y. Li, J. Zhang, and X. Xiong, “Sch-net: A deep learning architecture for automatic detectionnbsp;of schizophrenia,” *BioMedical Engineering OnLine*, vol. 20, no. 1, 2021.

[27] B. Elvevåg, P. W. Foltz, M. Rosenstein, and L. E. DeLisi, “An automated method to analyze language use in patients with schizophrenia and their first-degree relatives,” *Journal of neurolinguistics*, vol. 23, no. 3, pp. 270–284, 2010.

[28] T. B. Holmlund, C. Chandler, P. W. Foltz, A. S. Cohen, J. Cheng, J. C. Bernstein, E. P. Rosenfeld, and B. Elvevåg, “Applying speech technologies to assess verbal memory in patients with serious mental illness,” *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[29] N. B. Mota, R. Furtado, P. P. Maia, M. Copelli, and S. Ribeiro, “Graph analysis of dream reports is especially informative about psychosis,” *Scientific Reports*, vol. 4, no. 1, 2014.

[30] L. Palaniyappan, N. B. Mota, S. Oowise, V. Balain, M. Copelli, S. Ribeiro, and P. F. Liddle, “Speech structure links the neural and socio-behavioural correlates of psychotic disorders,” *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 88, p. 112–120, 2019.

[31] N. Rezaii, E. Walker, and P. Wolff, “A machine learning approach to predicting psychosis using semantic density and latent content analysis,” *npj Schizophrenia*, vol. 5, no. 1, 2019.

[32] J. N. de Boer, A. E. Voppel, S. G. Brederoo, H. G. Schnack, K. P. Truong, F. N. Wijnen, and I. E. Sommer, “Acoustic speech markers for schizophrenia-spectrum disorders: A diagnostic and symptom-recognition tool,” *Psychological Medicine*, p. 1–11, 2021.

[33] S. Bernardini, A. Ferraresi, and M. Miličević, “From epic to eptic—exploring simplification in interpreting and translation from an intermodal perspective,” *Target. International Journal of Translation Studies*, vol. 28, no. 1, pp. 61–86, 2016.

[34] M. A. Covington, S. A. Lunden, S. L. Cristofaro, C. R. Wan, C. T. Bailey, B. Broussard, R. Fogarty, S. Johnson, S. Zhang, M. T. Compton, and et al., “Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders,” *Schizophrenia Research*, vol. 142, no. 1-3, p. 93–95, 2012.

[35] F. Martínez-Sánchez, J. A. Muela-Martínez, P. Cortés-Soto, J. J. G. Meilán, J. A. V. Ferrándiz, A. E. Caparrós, and I. M. P. Valverde, “Can the acoustic analysis of expressive prosody discriminate schizophrenia?,” *The Spanish journal of psychology*, vol. 18, 2015.

[36] Y. Tahir, Z. Yang, D. Chakraborty, N. Thalmann, D. Thalmann, Y. Maniam, N. A. binte Abdul Rashid, B.-L. Tan, J. Lee Chee Keong, and J. Dauwels, “Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia,” *PLoS One*, vol. 14, no. 4, p. e0214314, 2019.

[37] W. Aoudi and A. M. Barbar, “Support vector machines: A distance-based approach to multiclass classification,” in *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, pp. 75–80, 2016.

[38] “Sklearn.ensemble.randomforestclassifier.”

[39] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.

[40] J. De Boer, M. Van Hoogdalem, R. Mandl, J. Brummelman, A. Voppel, M. Begemann, E. Van Dellen, F. Wijnen, and I. Sommer, “Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts,” *NPJ schizophrenia*, vol. 6, no. 1, pp. 1–10, 2020.

[41] J. N. de Boer, M. van Hoogdalem, R. C. Mandl, J. Brummelman, A. E. Voppel, M. J. Begemann, E. van Dellen, F. N. Wijnen, and I. E. Sommer, “Language in schizophrenia: Relation with diagnosis, symptomatology and white matter tracts,” *npj Schizophrenia*, vol. 6, no. 1, 2020.

[42] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[43] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[44] M. Wolters, K. Nicodemus, and A. Cohen, "Clinically interpretable acoustic meta-features for characterising the effect of mental illness on speech and voice," Nov. 2017.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, 2002.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[47] C. Demiroglu, A. Beşirli, Y. Ozkanca, and S. Çelik, "Depression-level assessment from multilingual conversational speech data using acoustic and text features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–17, 2020.

[48] I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, and J.-P. Onnela, "Relapse prediction in schizophrenia through digital phenotyping: A pilot study," *Neuropsychopharmacology*, vol. 43, no. 8, p. 1660–1666, 2018.

[49] K. T. Jørgensen, M. Bøg, M. Kabra, J. Simonsen, M. Adair, and L. Jönsson, "Predicting time to relapse in patients with schizophrenia according to patients' relapse history: a historical cohort study using real-world data in sweden," *BMC psychiatry*, vol. 21, no. 1, pp. 1–12, 2021.

[50] M. J. Begemann, I. A. Thompson, W. Veling, S. S. Gangadin, C. N. Geraets, E. van't Hag, S. J. Müller-Kuperus, P. P. Oomen, A. E. Voppel, M. Van Der Gaag, *et al.*, "To continue or not to continue? antipsychotic medication maintenance versus dose-reduction/discontinuation in first episode psychosis: Hamlett, a pragmatic multicenter single-blind randomized controlled trial," *Trials*, vol. 21, no. 1, pp. 1–19, 2020.

[51] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[52] Y. Qian, H. Hu, and T. Tan, "Data augmentation using generative adversarial networks for robust speech recognition," *Speech Communication*, vol. 114, p. 1–9, 2019.

[53] F. U. Nuha and Afiahayati, "Training dataset reduction on generative adversarial network," *Procedia Computer Science*, vol. 144, p. 133–139, 2018.

[54] H. Lin, L. Deng, D. Yu, Y.-f. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary asr," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[55] Q. Guo, F. Wang, J. Lei, D. Tu, and G. Li, "Convolutional feature learning and hybrid cnn-hmm for scene number recognition," *Neurocomputing*, vol. 184, p. 78–90, 2016.