Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 13, No. 1, March 2025, pp. 1~17
ISSN: 2089-3272, DOI: 10.52549/ijeei.v13i1.5506

Detection and Estimation of Schizophrenia Severity from Acoustic Features with Inclusion of K-means as Voice Activity Detection Function

Sheriff Alimi[1], Afolashade Oluwakemi Kuyoro[1], Monday Okpoto Eze[1] , Oyebola Akande[1]
[1]Department of Computer, Faculty of Computing and Engineering Sciences,
Babcock University, Ilisan-Remo, Ogun State, Nigeria

Article Info

Article history:

Received Mar 22, 2024
Revised Dec 28, 2024
Accepted Jan 11, 2025

Keywords:

Acoustic features
Enhanced K-means
Multi-layer Perceptron
Severity Estimation
Schizophrenia

ABSTRACT

Schizophrenia symptom severity estimation provides quantitative information that is useful at both the detection and treatment stages of the mental disorder, as the information helps in decision-making and improves the management of the illness. Very limited studies have been recorded for estimating the symptom severity as a regression task with machine learning, especially from speech recordings, which is the aim of this study coupled with detection. Acoustic features, which comprise frequency-domain and time-domain features, were extracted from 60 schizophrenia subjects and 59 healthy controls enrolled in this research. The acoustic features were used to train GridSearchCV-optimized XGBoost as a classifier. Three Multi-Layer Perceptron (MLP) networks, hyper-parameter-tuned by Bayesian Optimizer, were trained to predict the sub-type symptom severity from acoustic extracted features from the schizophrenia groups. The XGBoost classification model that discriminates between schizophrenia and healthy groups achieved a classification accuracy of 98.6%. The three MLP regression models yielded Mean Absolute Errors of 1.975, 2.856, and 1.555, as well as correlation coefficients of 0.888, 0.806, and 0.786 for predicting positive, negative, and cognitive symptom scores, respectively. Solution architecture for the deployment of the models for practical was suggested.

Corresponding Author:

Sheriff Alimi
Department of Computer Science, School of Computing and Engineering Sciences
Babcock University, Ilisan-Remo, Ogun State, Nigeria
Email: alimi0356@pg.babcock.edu.ng

1. INTRODUCTION

Schizophrenia is one of several mental illnesses with complex disturbances relating to perception, thinking, and social behavior [1] [2]. Roughly 21 million people in the world suffer from schizophrenia [3]. Schizophrenia assessment falls into two categories: classification and regression tasks. The classification is concerned with distinguishing schizophrenia from a healthy group, while the regression is concerned with estimating schizophrenia symptom severity. Unstructured data such as magnetic resonance imaging (MRI), electroencephalography (EEG), and voice recordings have been used with several machine learning algorithms for the classification task, and remarkable results were obtained in terms of classification accuracy [4], [5], [6]. The classification task involves feature extraction, which could be done manually (handcrafting) or automatically with a convolutional neural network (CNN) or auto-encoder [7], [8]. Sometimes, a feature selection stage is incorporated to improve the model's classification performance [9], [10].

The aspect of the use of machine learning for automated estimating or predicting schizophrenia symptom severity (regression task) has not experienced significant research effort, as seen in the case of the classification task. The recorded efforts are restricted to MRI and EEG, though [9] , and [11] are documented studies. It is important to note that voice recordings have not been used for severity estimation as a regression task. The best efforts are still limited to classification, which distinguishes between high- and low-severity symptom classes. The main purpose of this study is to use acoustic features extracted from recorded speech to distinguish between schizophrenia and healthy groups and also to estimate schizophrenia severity across the three subtypes: positive, negative, and cognitive symptoms with the incorporation of voice activity detection function to remove silent regions of the speech at the pre-processing stage.

Clinicians have used speech to assess the mental health of patients with schizophrenia disorder because speech offers a plethora of information that may be used to assess the speaker's mental state [12]. From several peer-reviewed publications, acoustic features taken from voice recordings have been effectively utilized to diagnose schizophrenia. The review literature is segmented into categories from speech or acoustic features, (1) Segregating between schizophrenia and healthy groups and (2) Classifying schizophrenia severity into two categories as low and high severity.

Standard dynamic volume value (SDVV), symmetric spectral difference level (SSDL), and quantization error and vector angle (QEVA) were extracted from speech recordings of 28 schizophrenia and 28 healthy controls while reading angry and afraid emotion text. A decision tree was used to discriminate between the two groups, which yielded a classification accuracy of 98.2% [13].With the extraction of acoustic features such as pitch, speech quality, length of voiced and unvoiced segments, number of voiced segments, variation in spectral slope, and loudness from speech recordings of 86 schizophrenia patients and 77 healthy controls and the use of SVM as a classifier to discriminate between the two groups, a classification accuracy of 82.8% was achieved [14].[6] employed sch-net deep learning for both automatic feature extraction and discrimination between schizophrenia and healthy speech samples, which recorded a classification accuracy of 97.68% but a better performance of 99.5% when applied to the LANNA speech database.

Speech recordings were obtained from 52 schizophrenia subjects(SZ) and 26 healthy controls(HC) who were enrolled in a study by Chakraborty [15]. 26 Low-level features which include MFCC (12), zero-crossing rate, pitch, and intensity among several others were obtained from speech signals, and the delta of these features was also computed. The final acoustic features which are 988 per subject were derived by statistical operations such as skewness, kurtosis, and standard deviation on the features above. The combination of $\chi 2$ and Adaboosted Decision Tree as feature selection and classifying algorithms respectively yielded the best classification accuracy of 84.62% for low and high negative symptom severity classes, while the combination of PCA and SVM yielded a correct classification of 79.49% between SZ and HC groups.

Speech recordings were obtained from 21 healthy controls and 21 schizophrenia participants then and transcribed. Each transcription was represented as a word trajectory graph from which connectedness attributes were extracted. All these variables were converted Disorganization Index by their weighted combination with respect to the correlation of positive and negative symptom scales. Naïve Bayes was used as the classifier for discriminating (1) between HCs and SCs and (2) mild and severe negative symptoms. The first classifier for discriminating between schizophrenia disorder and healthy control groups achieved an accuracy of 97%. The classifier based on the disorganized index for distinguishing between mild and severe negative symptoms achieved 100% accuracy[16].Speech recordings were obtained from both the control and schizophrenia groups. Voice activity is carried out manually to eliminate silent segments with a visual inspection. With discriminant analysis, extracted acoustic variables were used (1) to discriminate between healthy and schizophrenia subjects, and (2) to classify positive, negative, and depressive symptomatology into low and high severity. Discriminating between schizophrenia and healthy controls with 12 variables yielded an accuracy of 95.2%. Distinguishing between low and high severity across positive, negative, and depressive symptomatology with 8, 10, and 10 features, respectively, produced accuracies of 71.9%, 75.9%, and 79.4% [17].

Prosody formant, source, and spectral acoustic features were obtained from a voice corpus of 26 healthy and schizophrenia subjects with the corresponding NSA-16 ratings. The relationship between the features and NSA scores was examined with statistical analysis and inter-group classification for schizophrenia into observable and non-observable was done with a machine learning classifier. Discrimination between schizophrenia and healthy controls was also conducted. The best discriminating accuracy between observable and non-observables is 79.6% while discriminating accuracy for healthy controls and schizophrenia was 81.3% using multilayer perceptron. Few of the acoustic features for the prediction of negative symptoms of schizophrenia from emotion have correlation coefficients >0.3 for some of the NSA criteria [18]. Speech recordings were obtained from both schizophrenia subjects (acute and chronic) and healthy control subjects based on the reading of emotionally neutral text. In addition, AMDP, SANS, PANSS, etc. ratings were also obtained from schizophrenia subjects. Non-verbal speech features such as mean pause duration, number of

pause durations, silence proportion, mean energy, F0 amplitude, and F0 dB bandwidth, etc were extracted. ANOVA/MANOVA identified 12 acoustic features with significant variance between the control and schizophrenia subjects, and, with the aid of discriminant function, a classification accuracy of 85.6% was achieved. In discriminating between low and high severity using the discriminant function, a classification accuracy of 78.6% with the aid of the following six acoustic features, which include total recording time, total length of utterance, and F0 contour was obtained [19]

Most of the reviewed studies focused on detecting schizophrenia from speech recordings, the best attempt at severity is still classification which classifies severity into high and low as seen in [19]. The research conducted by [20] and [9], SMRI and EEG have been used as regression tasks to estimate the severity of schizophrenia symptoms, with mean absolute errors of 2.71 and 1.44 and correlation coefficients of 0.811 and -0.625, respectively.

The main objective of the study was to use acoustic features or speech signals to estimate symptom severity as a regression task or continuous outcomes

2. MATERIALS AND METHODS

The methodologies comprise one method for the classification task and the second method for the regression task. The classification task is to distinguish between schizophrenia and healthy control, a binary task based on the input acoustic features. The focus of the second method are to build three regression models that estimate positive, negative, and cognitive schizophrenia symptom severity and their respective predictions in the ranges of 0-42, 0-42, and 0-96, respectively, based on the PANSS rating scale.

The classification task method consists of the following stages:

1. Data Acquisition
2. Voice Activity Detection and Feature Extraction
3. Feature Selection
4. Classifier Training and Model Optimization with GridSearchCV
5. Validation of the optimized classifier model

The regression ask method has the under-listed stages:

1. Data Acquisition
2. Voice Activity Detection and Feature Extraction
3. Data Augmentation
4. Training and Optimization of Regression Models with Bayesian Optimizer
5. Validation of Optimized Models

The same data acquisition, voice activity detection, and feature extraction procedures are employed for both classification and regression tasks, as shown in Figures 1 and 2, though presented separately to simplify the architecture.

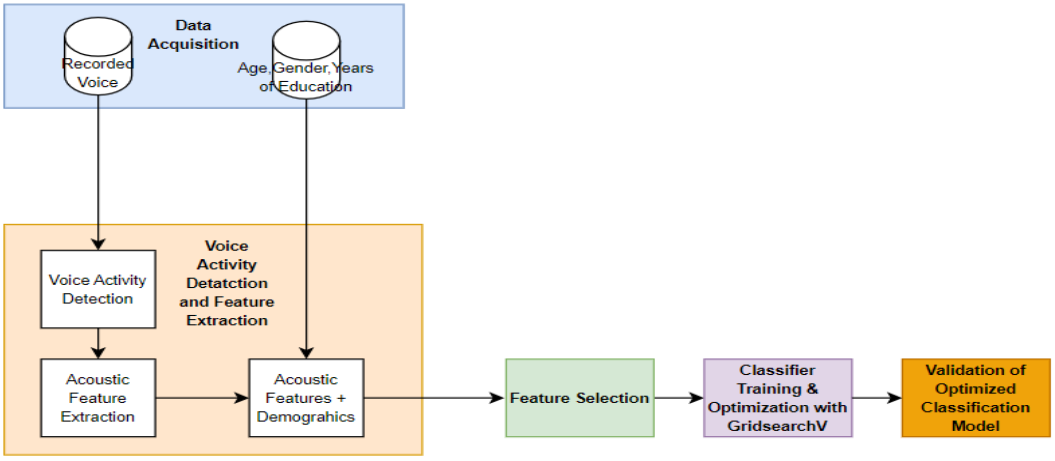

Figure 1. Classification Task Method

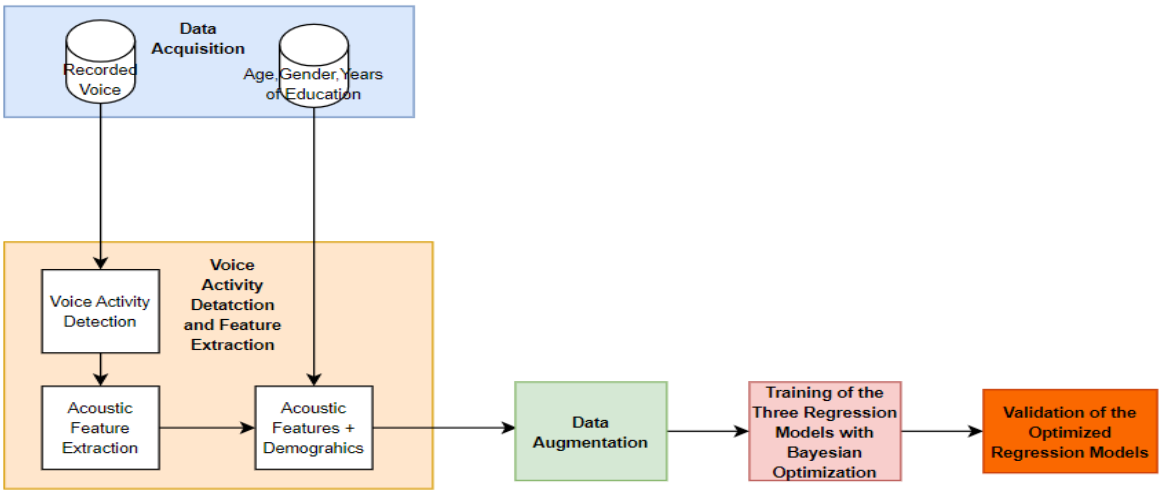

Figure 2. Regression Task Method

2.1 Data Acquisition

Sixty (60) schizophrenia subjects were enrolled from the Federal Neuro-Psychiatric Hospital, Yaba, Lagos, Nigeria while fifty (59) healthy controls were obtained from Babcock University Psychiatric Center, Ilisan, Ogun state, Nigeria. Participants with significant physical ailments or other underlying medical conditions were excluded from the sample group. The statistics of the participants are presented in Table 1 and Table 2 shows the severity scores of some of the research participants.

Each participant in each group was asked to describe themselves, and the conversation was recorded on audio, and additional information (age, years of education) was also collated. A Windows 10 laptop running the Audacity software and a headset microphone were used to record the speech using 16-bit linear PCM at a 44 kHz sampling rate.

For the schizophrenia group, two psychiatrists estimated the positive, negative, and cognitive symptom severity for each of them using the Positive Negative Syndrome Scale (PANSS) rating scale.

Table 1. The statistics of the participants

Categories	Gender	Count	Age Distribution ($\bar{x} \pm \sigma$)	Years of Education ($\bar{x} \pm \sigma$)
Control	Female	28	26.28 ± 9.27	16.428 ± 2.45
	Male	31	29.61 ± 9.17	18.1 ± 2.41
Schizophrenic	Female	28	41.67 ± 11.16	13.75 + 2.99
	Male	32	39 ± 11	15.09 ± 3.12

Table 2. Sample of Schizophrenia Symptom Severity Scores of Some of the Participants

SN	Gender	Age	Years of Education	Positive Scores	Negative Scores	Cognitive Scores	PANSS Total Score
1	F	29	17	10	8	21	39
2	F	57	12	18	23	22	63
3	F	55	6	8	16	22	46
4	F	27	14	20	17	22	59
5	F	54	12	23	16	28	67
6	M	25	17	8	31	22	61
7	F	30	17	33	40	46	119
8	M	19	12	12	25	28	65
9	M	43	12	11	19	27	57
10	F	22	9	13	43	37	93
11	M	23	15	9	28	20	57
12	F	30	17	21	8	28	57
13	M	18	12	26	29	22	77
14	M	33	17	24	43	30	97

2.2 Voice Activity Detection and Feature Extraction

The output of the stage is utilized for both classification and regression tasks.

2.2.1 Voice Activity Detection and Feature Extraction

Many speech applications incorporate voice activity detection to remove silence and noise from speech to improve the performance outcome of such systems. The voice activity detection is realized in two stages, as depicted in Figure 3.

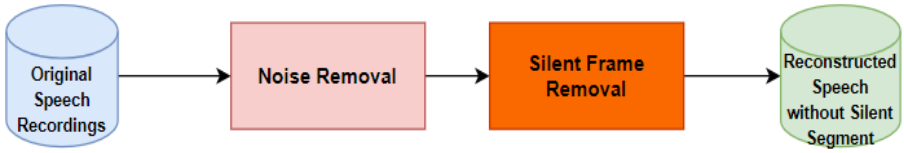

Figure 3. The voice activity detection process Voice Activity Detection Process

2.2.1.1 Noise Removal

Frequency band limiting and spectral subtraction were combined for noise removal. The human speech frequency ranges between 80 and 8,000 Hz[21], and an 8-node Butterworth Bandpass filter that allows the specified frequency components of the recorded speech to pass through while rejecting frequency components outside the band. The Bandpass filter was implemented with the Python Scipy library.

Spectral subtraction removes noise within the allowed frequency band. The output of the pass band filter is converted into a frequency domain using the Fast Fourier Transform (FFT). An estimate of the noise is obtained (the mean value of the total absolute mean of each frame's spectral level) and then subtracted from the signal. At any point in the signal where the noise estimated is greater, the value is set to zero. The signal is then converted back to the time domain by inverse Fast Fourier Transform (IFFT).

2.2.1.2 Silent Segment Removal

The designed approach utilized for silent frame/segment removal from each participant's speech is presented in Figure 4.

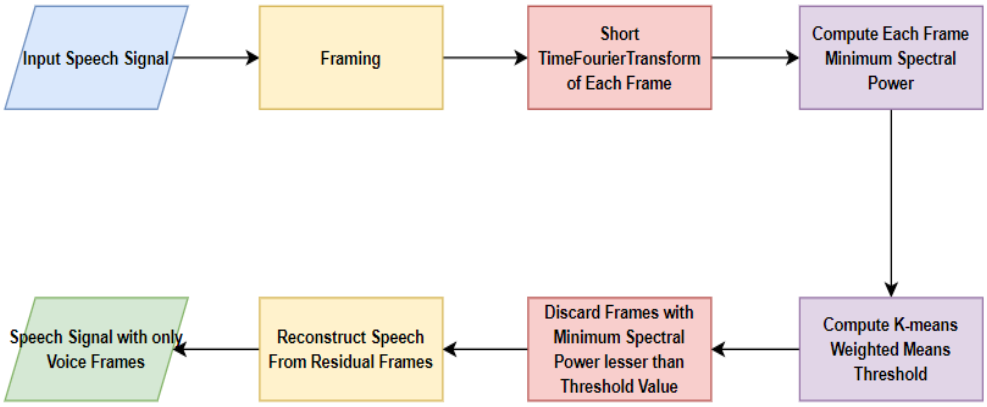

Figure 4. Silent Segments Removal Functionality Design Diagram

Each speech signal was broken into frames of 2048 samples, with Short Time Fourier Transform (STFT), each frame was converted to the frequency domain, and minimum spectral power was computed for each of the frames. A K-means thresholding (an extension of the k-means clustering algorithm introduced in this study) was employed to classify frames as either silent frames or voiced frames with each frame's spectral power as input. The K-Mean Thresholding Algorithm details are provided below.

Algorithm 1: K-Means (Weighted) Thresholding algorithm

(1) Randomly Select 2 points or centroids υ_1 and υ_2 from the input dataset (list voice frame minimum spectral power)
(2) Assign each frame of the frame spectra's mean to the closest centroid, to form 2 clusters
(3) Compute mean the means of each cluster to form new centroids υ_1 and υ_2

While (*the number of iterations is less than the predefined times*) do
Repeat steps 2 and 3.

End

(4) Obtain the number of data points $N_1$and N_2 assigned to centroids υ_1 and υ_2 respectively, and compute the weighted average as expressed in equation (1).

$$WeightedThreshold_{(KmeansWTh)} = \frac{N_1 * \upsilon_1 + N_2 * \upsilon_2}{N_1 + N_1} \quad (1)$$

Frames with a minimum spectral power equal to or higher than the threshold were regarded as belonging to the speech regions of the voice recordings, whereas frames with a minimum spectral power lower than the k-means-weighted threshold were labeled as "silent frames" (without voice content) and were eliminated. Only the voiced frames were used to reconstruct the speech signals.

2.2.2 Feature Extraction

Each participant's reconstructed speech signals (silent segment removed) were split into 8 segments, and features were extracted from each segment. This translates to 472 and 480 data sizes for the healthy controls (59 subjects) and schizophrenia group (with 60 subjects, respectively). Time and frequency domain features were extracted from each of the voice segments.

Mel-frequency Cepstral coefficients (MFCC) and their first and second-order derivatives, MFCC-delta and MFCC-delta-delta, are the features of the frequency domain. The MFCC was calculated by applying Mel-Filter Banks on the power spectrum of each speech frame, taking the output logarithm, and then applying the Discrete Cosine Transform (DCT).

A stack of four 13 MFCC, four 1st order derivatives of 13 MFCC, and four 2nd order derivatives of 13 MFCC are obtained per voice segment. The extraction of MFCC and its derivative was done using the librosa library in Python.

For time domain features, a total of 18 pitch-related features were extracted using the Parselmouth library in Python and the features are meanF0, stdevF0, f1-mean, f2-mean, and f3-mean, f4-mean, hnr(harmonic-to-noise ratio), localJitter, localabsoluteJitter, rapJitter, ppq5Jitter, ddpJitter, localShimmer, localdbShimmer, apq3Shimmer, aqpq5Shimmer, apq11Shimmer, and ddaShimmer.

Statistical features, six (6) in total were also computed: skewness, standard deviation, kurtosis, zero-crossing, the ratio of duration of voice segments to overall voice duration, and average mean power of the voice segment. Phonatory intensity variation diversity (PIVD) was also computed which brings the total of time domain features to twenty-five (25).

Equations (2) to (7) show how some of the time-domain acoustic features were computed.

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \quad (2)$$

$$jitter(local) = \frac{jitta}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100\% \quad (3)$$

$$Shimmer(local) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100\% \quad (4)$$

$$Shimmer(local, dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20 \times log\left(\frac{A_{i+1}}{A_i}\right)| \quad (5)$$

$$Shimmer(apq3) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - (\frac{1}{3}\sum_{n=i-1}^{i+1} A_n)|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100\% \quad (6)$$

$$Shimmer(apq5) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i-(\frac{1}{5}\sum_{n=i-2}^{i+2}A_n)|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100\% \quad (7)$$

There are 156 MFCC-related features in the frequency domain per segment, which includes 52 MFCC, 52 MFCC-delta, and 52 MFCC-delta-delta features. 156 features is a high dimension, which was reduced to 15 Independent Components (ICs) using Independent Components Analysis (ICA), with 5 ICs for the MFCC and each of its derivatives.3.

2.3 Feature Selection, Classifier Training, and Optimisation and Validation

This section focused on the downstream stages of the classification task which are discussed sequentially in detail.

2.3.1 Feature Selection

The features extracted comprise (1) 15 MFCC-based features after the application of ICA, (2) 18 pitch-related features (time domain), (3) 6 statistical features, (4) 1 intensity variation feature, Phonatory Intensity Variation diversity (PIVD), and (5) 3 demographic features (age, gender, and years of education). This brings the total number of features to 43.

For the classification task, segregating between schizophrenia and healthy groups, feature selection operations are performed to optimize classification performances.

Recursive Feature Elimination (RFE) and XGBoost were independently used in the selection of features that optimize segregation between schizophrenia and the healthy control group. XGBoost has an innate capacity to identify the significance of features throughout the training phase, optimize memory usage, and significantly reduce computation time. It has been used as a feature selection in several classification tasks and has helped improve classification outcomes in several studies [22][23][24][25]. Recursive Feature Elimination (RFE) is a widely used and effective relevant feature selection algorithm and is considered a good choice for this study[26].

For each feature selection algorithm, the number of features to be selected was varied, and the corresponding set of features that produced the best and most reliable results was obtained. The output features from each algorithm were combined, and their overall performance was validated by the XGBoost algorithm.

2.3.2 Classifier Training, Optimisation, and Validation

XGBoost is the choice of the classifier that was adopted because of its high performance. The features selected by RFE and XGBoost were combined to train the classifier. The dataset was divided into training and test datasets in a ratio of 70 to 30 respectively. This translates to training set 666 and the test set 286 (952 in total with the control group being 472 and schizophrenia group being 480).

The GridSearchCV was applied to optimize(tune hyperparameters) the classification performance of the XGBoost classifier by selecting the best hyperparameters. The search space for the hyper-parameters is listed below.

```
'n_estimator'  = {100,200,300,400}
'max_depth' = {3,4,5,6}
'min_child_weight ' = {1,3,5} 3
'subsample '= {0.4,0.6,0.8,0.9,1.0}
colsample_bytree' = {0.8, 0.9, 1.0}
```

The evaluation metrics used for the classifier are accuracy, F1-score, precision, and recall.

2.4 Data Augmentation, Training, and Optimisation of Regression Models and Validation

This section describes the data augmentation process, the training and optimization of the three regression models for estimating schizophrenia symptom severity, and the validation process.

2.4.1 Data Augmentation

Eight feature sets per participant with schizophrenia make up the 480-person dataset from the schizophrenia group. Large datasets are necessary for deep learning to discover intricate correlations between dependent variables and input data. Data augmentation was used to double the size of the dataset applied [27][28] and it has been applied in many regression tasks to improve performance [29][28][30][31].

The jittering augmentation technique is adopted, by adding random noise to the original dataset to generate synthetic data [32] as expressed in equation 8. The random number N_{random} ranges from 0 to 10% of the feature standard deviation (σ).

$$X_{synthetic} = X_{original} + N_{random} \quad (8)$$

2.4.2 Training, and Optimisation of Regression Models and Validation

The regression models to predict positive, negative, and cognitive symptom severity scores were realized using three Multilayer Perceptron Networks (MLP). All 43 features were employed in the training and evaluation of MLPs, the dataset is split up into training and test sets at a ratio of 80% to 20%. Since accurate predictions are prioritized in regression tasks, a model's accuracy is enhanced by having 80% more training data and being less susceptible to overfitting. The optimum hyper-parameters for each of the three MLP regressors are chosen using Bayesian optimization, and the parameter search space is the same for all three, as shown below:

```
{'num_layers': (1, 15),
 'units_per_layer': (5, 64),
 'learning_rate': (0.001, 0.1)}
```

The evaluation metrics for the regression models are correlation coefficients and Mean Absolute Error (MAE) expressed with equations 9 and 10 respectively.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (9)$$

The difference between the rankings of the two observations is denoted by d_i , while the number of observations is represented by n.

$$\text{Mean Absolute Error (MAE)} = \frac{\sum_i^N |Y_i - Y^*_i|}{N} \quad (10)$$

3. RESULTS AND FINDINGS

Voice files totaling 90 minutes and 43 seconds of voice speech were recorded from 59 healthy controls (from 102 participants) at Babcock University; likewise, 83 minutes and 32 seconds of voice recordings were taken from an interview conducted with 60 schizophrenia subjects (4 patients declined participation in this research).

The results and findings section is divided into four sub-sections: voice activity detection, extracted features, training, and validation of models.

3.1 Voice Activity Detection

Figure 5 displays a sample voice signal with a duration of 41.32 seconds before applying the suggested voice activity detection method. The output of the speech activity detection technique is shown in Figures 6 and 7. The first shows the expected reconstructed voice signal without silent portions, which lasts for 25.68 seconds, while the second shows the silent section, which lasts for 15.62 seconds.

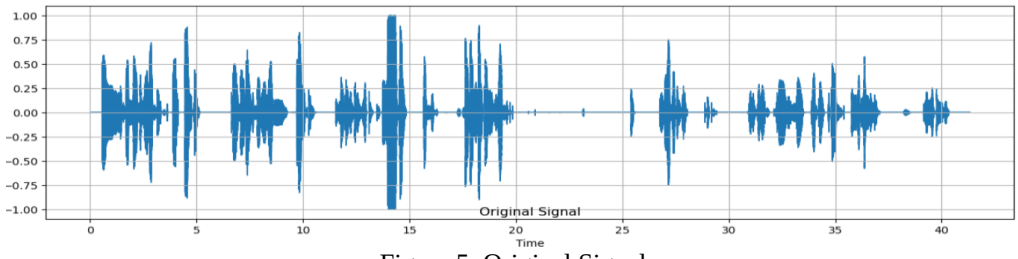

Figure 5. Original Signal

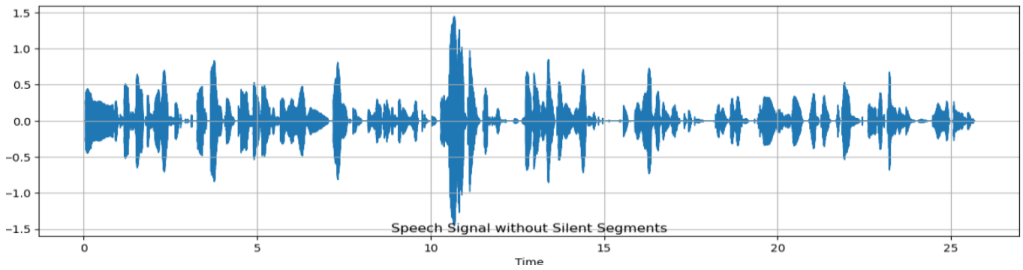

Figure 6. Voiced Segment

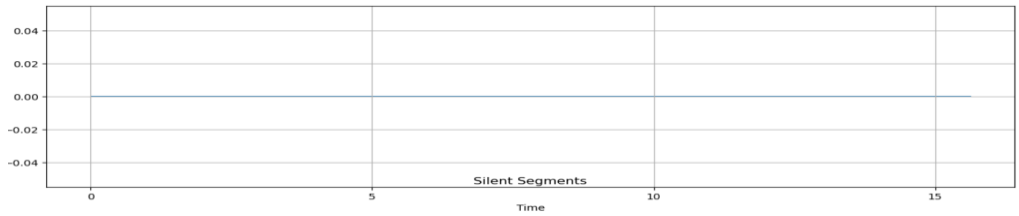

Figure 7. Silent Segment

By visual analysis of Figure 7, the amplitudes of all points are zero, the silent segment of the original speech signal. In like manner, the voiced speech segment is similar to the original voice signal by pattern except for the absence of the silent segment. With visual analysis, the implemented voice activity detection algorithm effectively removes silent frames or silent segments.

3.2 Extracted Acoustic Features

Samples of extracted pitch-related features, meanF0, stdevF0, hnr (harmonic-to-noise ratio), localJitter, localabsoluteJitter, rapJitter, ppq5Jitter, ddpJitter, localShimmer, localdbShimmer, apq3Shimmer, aqpq5Shimmer, apq11Shimmer, and ddaShimmer (a total of 14 features) are presented in Table 3. There is also a field for the class label, where class label 0 is the healthy control group and class label 1 is the schizophrenic group.

Table 4 shows samples of additional pitch-related features, F1-mean, F2-mean, F3-mean, and F4-mean, computed using the Parselmouth library. Statistically related features that speak to asymmetry and tailedness (outlier occurrence) of the data, such as skewness, kurtosis, and standard deviation, plus PIVD (Phonatory intensity variation diversity), and zero-crossing, are presented in Table 5.

Table 3. Pitch Related Features for the two groups

Class	meanF0	stdevF0	hnr	localJitter	localabsoluteJitter	rapJitter	ppq5Jitter	ddpJitter	localShimmer	localdbShimmer	apq3Shimmer	apq5Shimmer	apq11Shimmer	ddaShimmer
1	193.3	13.95	24.76	0.011	5.68E-05	0.0037	0.00446	0.01112	0.041	0.521	0.010	0.017	0.039	0.029
1	186.7	25.166	22.65	0.016	8.69E-05	0.007	0.008	0.020	0.058	0.695	0.014	0.024	0.057	0.044
1	205.97	37.397	26.751	0.010589	5.16E-05	0.003	0.0041	0.009571	0.040255	0.565416	0.008	0.014	0.036	0.025
0	168.3	60.88	12.22	0.024	0.0001	0.0118	0.0115	0.0353	0.09587	1.009322	0.0334	0.044	0.091	0.102
0	170.5	35.54	12.915	0.01881	0.00011	0.009	0.009	0.026	0.097	1.0038	0.033	0.046	0.088	0.099
0	166.35	58.30	12.78	0.021	0.000125	0.009	0.01	0.029	0.094	1.011	0.033	0.042	0.075	0.100

Table 4. F1 to F4 Mean and Mean Power of the two groups

Class	F1-mean	F2-mean	F3-mean	F4-mean	Mean-Power
1	391.0461	1231.85	2453.248	3486.896	0.01689937
1	352.4319	1188.287	2269.945	3382.033	0.02061915
1	399.7812	1181.064	2243.737	3299.759	0.02220404
1	378.6293	1221.089	2217.192	3345.546	0.01780628
0	425.7124	1232.06	2295.109	3280.029	0.00791787
0	515.3805	1220.959	2287.378	3164.986	0.00641012
0	412.5769	1195.906	2263.175	3178.462	0.00654464
0	404.3075	1081.632	1936.848	3247.719	0.00780092

Table 5. Statistical -Related Features of the two groups

Class	Zero-crossing	Stdev	Skewness	Kurtosis	PIVD
1	0.032	0.125	-0.196	0.716	0.026
1	0.026	0.112	-0.226	2.303	0.019
1	0.028	0.149	-0.299	0.363	0.033
1	0.024	0.133	-0.242	0.555	0.031
1	0.032	0.145	-0.185	0.335	0.028
0	0.074	0.088	-1.482	6.475	0.028
0	0.061	0.074	-1.609	6.719	0.013
0	0.078	0.088	-1.359	5.334	0.014
0	0.093	0.080	-1.545	7.315	0.012

3.3 Feature Selection, Classifier Training, Optimisation and Validation

This section discussed the outcomes of feature selection, binary classifier training, its optimization, and its validation.

3.3.1 Feature Selection

Fifteen MFCC-based features were obtained per speech recording after ICA was utilized for dimension reduction. Eighteen pitch-related features and six statistical features were also computed; in addition to that, one intensity variation (PIVD) feature was also included. Three demographics (age, gender, and years of education) were also added, which brings the total number of features to 43.

Figure 8 is the graph of accuracy versus the number of features selected by XGBoost feature selection, and it is observed that the best classification result was produced with the selection of five (5) features.

XGBoost Selected Features (5 features): The features are kurtosis, active-ratio, localShimmer, F1mean, and skewness.

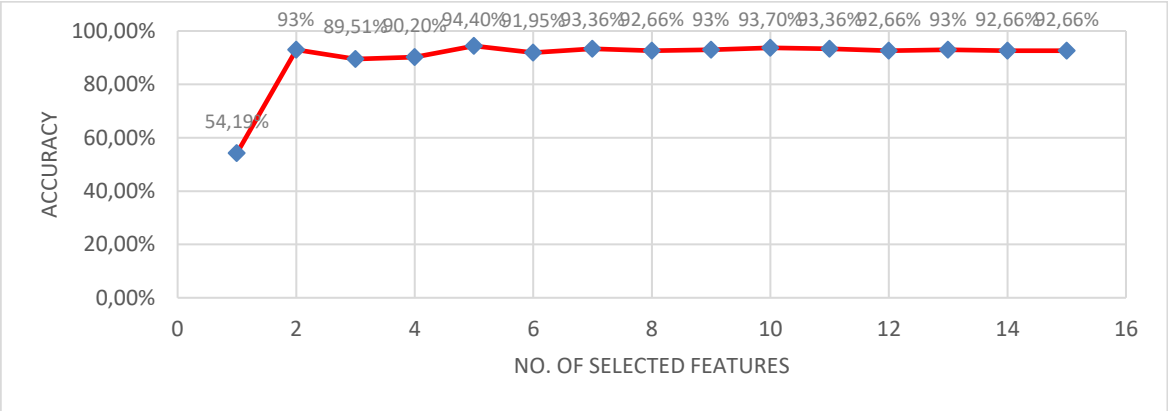

Figure 8. Number of selected features by XGBoost against Accuracy

For RFE feature selection, Figure 9 is the graph of accuracy versus the number of selected features. A reliable best classification accuracy was obtained when six (6) features were selected. From observation, better performances were obtained when one and two features were selected, respectively, but they were discarded due to the likely consequences of the feature's outlier effect.

RFE-selected features (6 features): The features are the 3rd and 4th IC (Independent Component) of mfcc, ddpJitter, active-ratio, skewness, and mean-power.

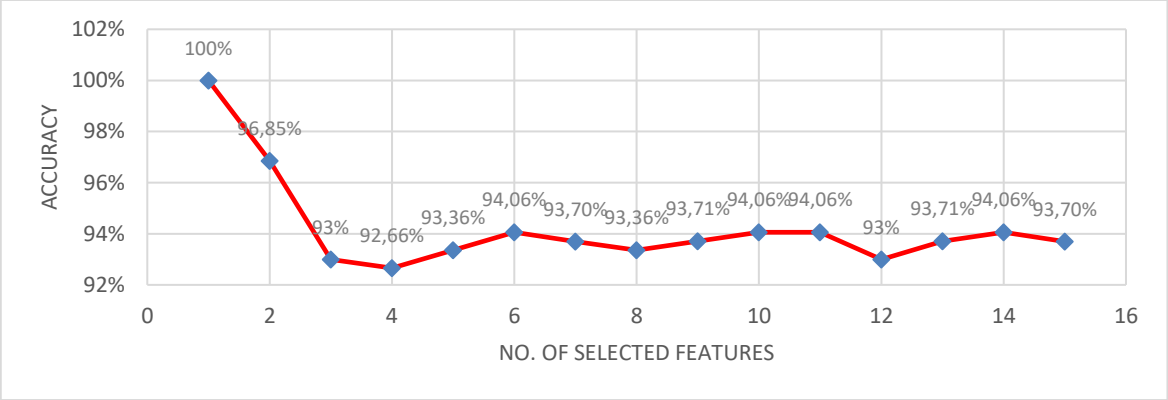

Figure 9. Number of selected features by RFE against Accuracy

The features selected by RFE and the XGBoost feature selector were combined and presented to XGBoost as selected features (9 unique features, as there were overlaps) to validate performance. The performance output was better compared to the set of features selected by each of the feature selection algorithms as presented in Figure 10. The combined features produce a slightly better classification accuracy of 94.76%.

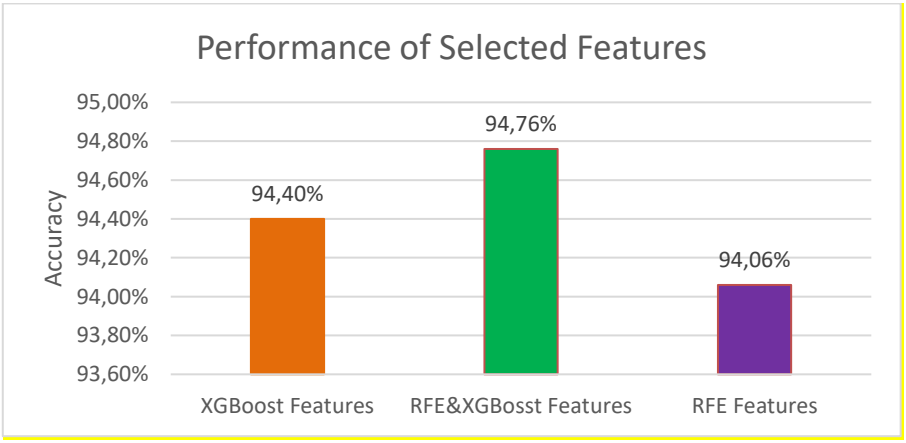

Figure 10. Performance of the Various selected sets of features

Combination of XgBoost and RFE-selected features (9 features): The combined features are: kurtosis, active-ratio, localShimmer, 3rd and 4th IC (Independent Component) of mfcc, ddpJitter, skewness, mean-power and f1mean.The features common RFE and XGBoost feature selectors are active-ratio and skewness.

3.3.2 Classifier Training, and Optimisation and Validation

The nine combined features were used to train the XGBoost classifier, whose hyperparameters were optimized by GridSearchCV.The confusion matrix in Figure 11 is the outcome of the performance of the classifier against the test dataset with a true positive of 143 and a true negative of 139. The misclassified data points out of 286 validation/test sets are 4, with 3 reported as false positives and 1 against false negatives.

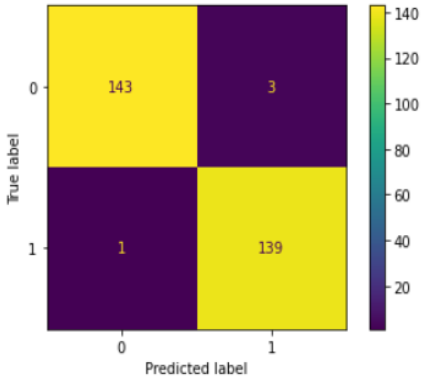

Figure 11. XGBoost Classifier Confusion Matrix

From the confusion matrix of Figure 11, the computation of precision, recall, f1-score, and accuracy yielded results of 97.89%, 99.29%, 98.58%, and 98.6%, respectively presented in Figure 12.

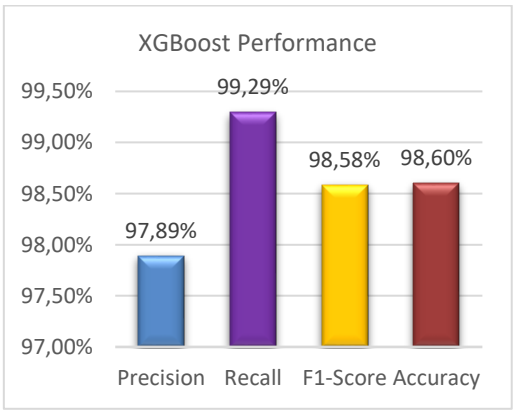

Figure 12. XgBoost Performance Metrics on the Optimal Selected Hyper-parameters.

3.4 Training, Optimisation, and Validation of Regression Models

With Bayesian optimization, the optimally selected architectures for the positive, negative, and general psychological MLP models have learning rates of 0.001, 0.0967, and 0.1, and the numbers of their respective hidden layers are 8, 10, and 15.

After the training and validation with the test dataset, the positive, negative, and general psychopathology MLP regressors reported a correlation coefficient of 0.888, 0.806, and 0.786 in that order (see Figure 13). The MAE for MLP positive, negative, and cognitive or general psychopathology severity regressors are 1.975, 2.856, and 1.555, respectively as presented in Figure 14. The average 0.827 correlation coefficient of the MLP models' predictions indicates a strong connection with the actual values.

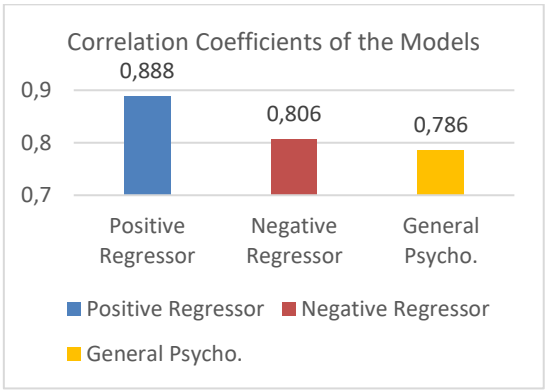

Figure 13. Correlation Coefficients of the Regression Models

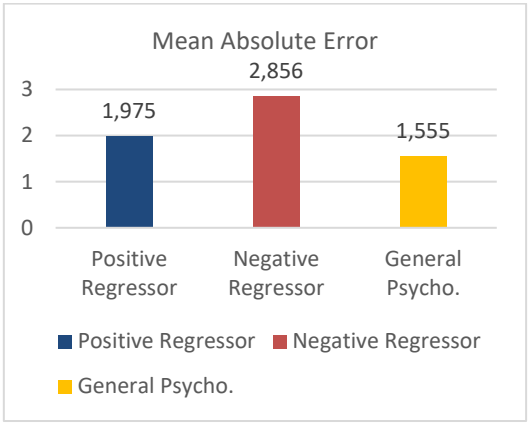

Figure 14. MAE for Regression Models

Figures 15 to 17 are the graphs of the predicted and actual values for the three symptom categories, which show similar trends, that reinforce the correlation coefficients.

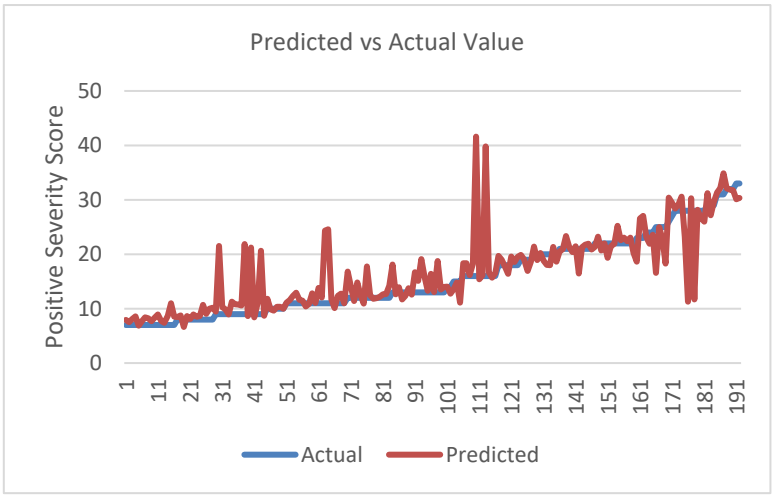

Figure 15. Predicted Positive Symptom Severity Scores vs the Actual as rated by Psychiatrist

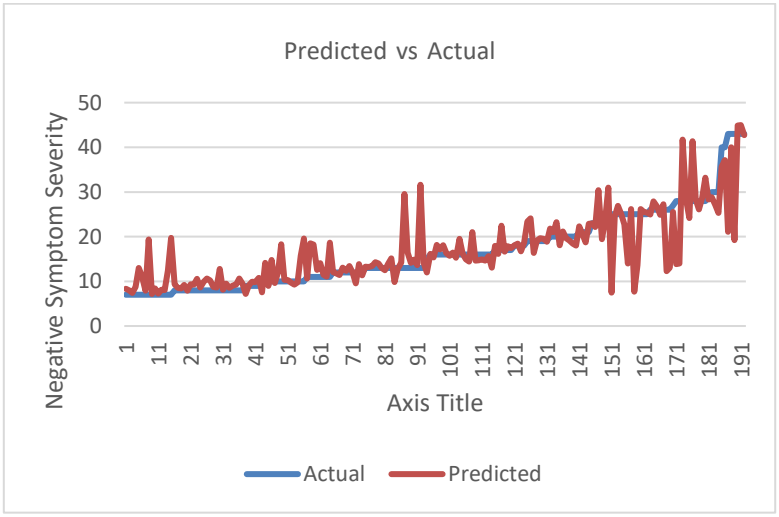

Figure 16. Predicted Negative Symptom Severity Scores vs the Actual as rated by Psychiatrist

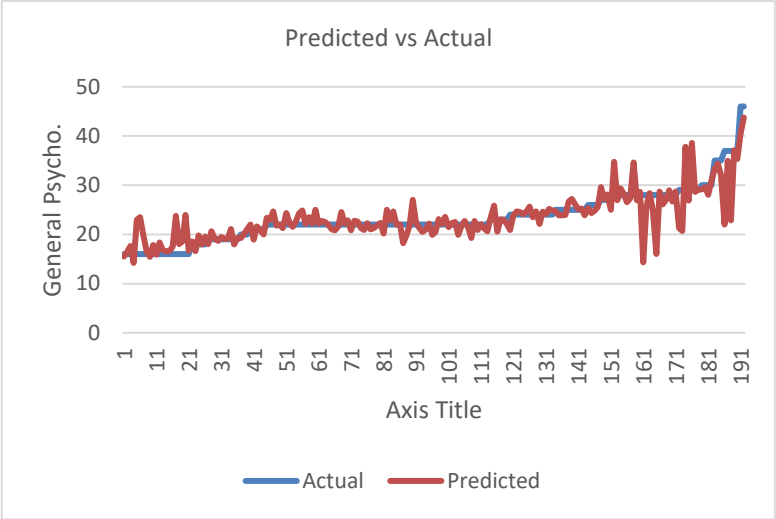

Figure 17. General Psychological Scores vs the Actual as rated by the Psychiatrist

4. DISCUSSIONS

The extended k-means algorithm, which functions as a thresholding algorithm in discriminating between voice and silence frames (voice activity detection) in speech recording, proves to be effective in segregating silent segments from voice segments of the recorded speeches.

RFE and XGBoost feature selectors identified nine features that optimize discrimination between schizophrenia and healthy groups, and they are kurtosis, active-ratio, localShimmer, 3rd and 4th IC (Independent Component) of mfcc, ddpJitter, skewness, mean-power, and f1mean. The features common to both RFE and XGBoost feature selectors are active-ratio and skewness.

An XGBoost whose hyper-parameters were tuned by GridSearchCV was trained with the selected features and achieved precision, recall, F1-score, and accuracy of 97.89%, 99.29%, 98.58%, and 98.6%, respectively on the test dataset.

Table 6 shows the current result and that of previous studies that achieved classification accuracy that is above 90%.

Table 6 Comparison with Related Studies with Classification Accuracy Above 80%

Authors	Data source Type	No speech transcription	Classifier	Performance Result
[13]. He et al. 2021	Voice Recording	No Voice Transcription	Decision Tree	Accuracy=98.2%
[33] Espinola et al. 2021	Voice Recording	No speech transcription	SVM with the Pearson VII universal kernel	Accuracy=91.76%
[34] Huang et al. 2022	Voice Recording	No Voice Transcription	Linear discriminant analysis (LDA)	Accuracy=89%
[17]. Stassen et al. 1995.	Voice Recording	Transcribed voice recordings	Linear discriminant analysis (LDA)	Accuracy=95.2%
[35] Zhang et al. 2022	Video and speech	No Voice Transcription	RandomForest	Accuracy=96.60%
[16]. Mota et al. 2017.	Voice Recording	Transcribed voice recordings	Naïve Bayes	Accuracy=97%
Current study	Voice Recording	No Transcription	XGBoost	Accuracy=98.6%

An existing gap was addressed by this study with the development of regression models that estimate schizophrenia symptom severity from acoustic features which will serve as a framework to guide related studies in the future. All the acoustic features were used to train three MLP regression models for estimating schizophrenia's positive, negative, and cognitive symptom severity scores. The models achieved correlation coefficients of 0.888, 0.806, and 0.786 and Mean Absolute Errors of 1.975, 2.856, and 1.555 respectively for positive, negative, and cognitive or general psychopathology severity regressors.

Closely related works that implemented schizophrenia symptom severity as regression tasks using EEG and SMRI are presented in Table 7 with their respective results for comparison with this study.

Table 7 Closely Related Studies on Schizophrenia Symptom Severity

SN	Authors	Unstructured Source Data	Algorithm	Result
1	Kim et at [20]	Electroencephalography (EEG)	General linear model (GLM)	The correlation coefficient ranges from -0.6 to -0.702. Average Mean Absolute Error=2.71
2	Alimi et al [9]	Structural Magnetic Resonance Imaging (SMRI)	MLP	Correlation coefficient=0.811 Mean Absolute Error=1.44
3	Current study	Speech	MLPs	The average correlation coefficient=0.827 The average Mean Absolute Error=2.129

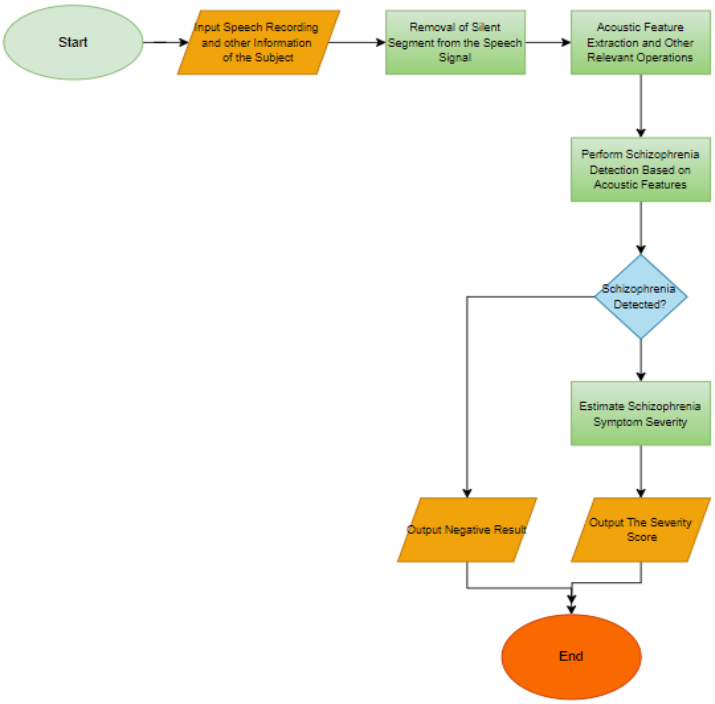

Figure 18. Solution Architecture for Practical Implementation

With the capability to detect and estimate schizophrenia severity, medical practitioners are armed with both qualitative and quantitative information that will make them more effective in schizophrenia treatment. With symptom severity estimation, the level of attention will be determined, as well as help with decision-making on the concentration of the anti-psychotic medication or its frequency of administration. The estimation capability will also help to monitor which serves as a feedback mechanism to determine the effectiveness of administered treatment. Above all, with voice recording as input to the detection and severity estimation assessment tool, the cost of data acquisition will drop significantly compared with MRI and EEG; it is also non-invasive and has a short cycle time for data acquisition. Schizophrenia diagnostic tests may be performed as often as needed with ease, and the cycle time is anticipated to be just a few minutes.

For practical utilization of the research outcome for clinical benefits, the solution architecture in Figure 18 was suggested. The subject's voice file, age, and gender information will be submitted to the proposed

diagnosis system. The necessary preprocessing and feature extraction operations will be executed. The classifier will then determine if the subject is schizophrenic or not based on the acoustic features extracted. In a situation where schizophrenia is not detected, the process ends, if schizophrenia is detected, a set of features is passed to the regression models to determine positive, negative, general psychopathology, and total PANSS severity scores.

5. CONCLUSION

In contrast to other research that classified severity into bands, such as low and high severity levels, based on voice or speech recordings, this study used acoustic features to estimate the severity of schizophrenia symptoms as continuous numerical scores. With Bayesian-optimized deep neural network architectures, state-of-the-art regression performance indicators were achieved. A high-performance schizophrenia detector classifier was also built using a subset of the acoustic features that maximize segregation between healthy and schizophrenia groups.

These cutting-edge models for detecting schizophrenia and estimating its severity will function as an expert system that will provide physicians with qualitative and quantitative data on the disorder's current status. These insights will significantly enhance physicians' decision-making and result in more economical and successful treatment of schizophrenia.

The number of participants in this study is small: fifty-nine (59) healthy controls and sixty (60) schizophrenia subjects. Though oversampling and data augmentation were applied, it is encouraged that similar research should be conducted with a reasonable number of participants to improve the generalization of the models.

REFERENCES

[1] A. Barbato, "Schizophrenia and public health," WHO Nations Ment. Heal. Initiat. World Heal. Organ. Div. Ment. Heal. Prev. Subst. Abus., 1998, doi: 10.1016/s0924-9338(99)80120-5.

[2] S. Grot et al., "Converting scores between the PANSS and SAPS/SANS beyond the positive/negative dichotomy," Psychiatry Res., vol. 305, 2021, doi: 10.1016/j.psychres.2021.114199.

[3] C. A. T. Naira and C. J. L. Del Alamo, "Classification of people who suffer schizophrenia and healthy people by EEG signals using deep learning," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, pp. 511–516, 2019, doi: 10.14569/ijacsa.2019.0101067.

[4] Y. Yang et al., "Automatic classification of first-episode, drug-naive schizophrenia with multi-modal magnetic resonance imaging," Sheng Wu Yi Xue Gong Cheng Xue Za Zhi, vol. 34, no. 5, 2017, doi: 10.7507/1001-5515.201607084.

[5] Z. Aslan and M. Akin, "Automatic detection of schizophrenia by applying deep learning over spectrogram images of EEG signals," Trait. du Signal, vol. 37, no. 2, pp. 235–244, 2020, doi: 10.18280/ts.370209.

[6] J. Fu et al., "Sch-net: a deep learning architecture for automatic detection of schizophrenia," Biomed. Eng. Online, vol. 20, no. 1, 2021, doi: 10.1186/s12938-021-00915-2.

[7] L. Shen, Q. Wang, and J. Shi, "Single-modal neuroimaging computer aided diagnosis for schizophrenia based on ensemble learning using privileged information," Sheng Wu Yi Xue Gong Cheng Xue Za Zhi, vol. 37, no. 3, 2020, doi: 10.7507/1001-5515.201905029.

[8] J. Oh, B. L. Oh, K. U. Lee, J. H. Chae, and K. Yun, "Identifying Schizophrenia Using Structural MRI With a Deep Learning Algorithm," Front. Psychiatry, vol. 11, Feb. 2020, doi: 10.3389/FPSYT.2020.00016.

[9] S. Alimi, A. O. Kuyoro, M. O. Eze, and O. Akande, "Utilizing Deep Learning and SVM Models for Schizophrenia Detection and Symptom Severity Estimation Through Structural MRI," Ingénierie des systèmes d Inf., vol. 28, no. 4, pp. 993–1002, Aug. 2023, doi: 10.18280/isi.280419.

[10] J. Liu, X. Wang, X. Zhang, Y. Pan, X. Wang, and J. Wang, "MMM: classification of schizophrenia using multi-modality multi-atlas feature representation and multi-kernel learning," Multimed. Tools Appl., vol. 77, no. 22, 2018, doi: 10.1007/s11042-017-5470-7.

[11] D. W. Kim, S. H. Lee, M. Shim, and C. H. Im, "Estimation of symptom severity scores for patients with schizophrenia using ERP source activations during a facial affect discrimination task," Front. Neurosci., vol. 11, no. AUG, pp. 1–6, 2017, doi: 10.3389/fnins.2017.00436.

[12] J. N. de Boer et al., "Acoustic speech markers for schizophrenia-spectrum disorders: A diagnostic and symptom-recognition tool," Psychol. Med., 2021, doi: 10.1017/S0033291721002804.

[13] F. He, J. Fu, L. He, Y. Li, and X. Xiong, "Automatic Detection of Negative Symptoms in Schizophrenia via Acoustically Measured Features Associated with Affective Flattening," IEEE Trans. Autom. Sci. Eng., vol. 18, no. 2, pp. 586–602, 2021, doi: 10.1109/TASE.2020.3022037.

[14] J. de Boer, A. Voppel, F. Wijnen, and I. Sommer, "ACOUSTIC SPEECH MARKERS FOR SCHIZOPHRENIA...Schizophrenia International Research Society (SIRS) 2020 Congress.," Schizophr. Bull., vol. 46, 2020.

[15] D. Chakraborty et al., "Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, pp. 6024–6028, 2018, doi: 10.1109/ICASSP.2018.8462102.

[16] N. B. Mota, M. Copelli, and S. Ribeiro, "ARTICLE Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance," vol. 3, p. 18, 2017, doi: 10.1038/s41537-017-0019-3.
[17] H. H. Stassen, M. Albers, J. Püschel, C. Scharfetter, M. Tewesmeier, and B. Woggon, "Speaking behavior and voice sound characteristics associated with negative schizophrenia," J. Psychiatr. Res., vol. 29, no. 4, pp. 277–296, 1995, doi: 10.1016/0022-3956(95)00004-O.
[18] Y. Tahir et al., "Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia," PLoS One, vol. 14, no. 4, pp. 1–17, 2019, doi: 10.1371/journal.pone.0214314.
[19] J. Piischel, H. H. Stassen, G. Bomben, C. Scharfetter, and D. Hell, "J OURNALOF PSYCHIATRIC F & EAFL ~ H Speaking behavior and speech sound characteristics in acute schizophrenia," Control, 1997.
[20] D. W. Kim, S. H. Lee, M. Shim, and C. H. Im, "Estimation of symptom severity scores for patients with schizophrenia using ERP source activations during a facial affect discrimination task," Front. Neurosci., vol. 11, no. AUG, 2017, doi: 10.3389/fnins.2017.00436.
[21] G. Fant, Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations. Berlin: Gruyter Mouton, 1971. [Online]. Available: ps://doi.org/10.1515/9783110873429
[22] K. A. Binsaeed and A. M. Hafez, "Enhancing Intrusion Detection Systems with XGBoost Feature Selection and Deep Learning Approaches," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.01405112.
[23] Z. Jiang, J. Che, M. He, and F. Yuan, "A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition," Renew. Energy, vol. 203, 2023, doi: 10.1016/j.renene.2022.12.124.
[24] S. S. Dhaliwal, A. Al Nahid, and R. Abbas, "Effective intrusion detection system using XGBoost," Inf., vol. 9, no. 7, 2018, doi: 10.3390/info9070149.
[25] B. Karan, "Speech-Based Parkinson's Disease Prediction Using XGBoost-Based Features Selection and the Stacked Ensemble of Classifiers," J. Inst. Eng. Ser. B, vol. 104, no. 2, 2023, doi: 10.1007/s40031-022-00851-2.
[26] S. Kilmen and O. Bulut, "Scale Abbreviation with Recursive Feature Elimination and Genetic Algorithms: An Illustration with the Test Emotions Questionnaire," Inf., vol. 14, no. 2, 2023, doi: 10.3390/info14020063.
[27] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image Data Augmentation for Deep Learning: A Survey," 2022, [Online]. Available: http://arxiv.org/abs/2204.08610
[28] F. Dubost et al., "Hydranet: Data augmentation for regression neural networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11767 LNCS, no. January 2020, pp. 438–446, 2019, doi: 10.1007/978-3-030-32251-9_48.
[29] N. Network, "Effects of Data Augmentation on the Nine-Axis IMU-Based," Sensors, vol. 23, no. 17, 2023, doi: doi.org/10.3390/s23177458.
[30] H. Ohno, "Auto-encoder-based generative models for data augmentation on regression problems," Soft Comput., vol. 24, no. 11, pp. 7999–8009, 2020, doi: 10.1007/s00500-019-04094-0.
[31] Y. El Khessaimi, Y. El Hafiane, A. Smith, and M. A. Barkatou, "The Effectiveness of Data Augmentation in Compressive Strength Prediction of Calcined Clay Cements Using Linear Regression Learning," NanoWorld J., vol. 9, no. September, pp. 1–6, 2023, doi: 10.17756/nwj.2023-s2-054.
[32] G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, and S. Gómez-Canaval, "Data Augmentation techniques in time series domain: a survey and taxonomy," Neural Comput. Appl., vol. 35, no. 14, pp. 10123–10145, 2023, doi: 10.1007/s00521-023-08459-3.
[33] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Vocal acoustic analysis and machine learning for the identification of schizophrenia," Res. Biomed. Eng., vol. 37, no. 1, pp. 33–46, 2021, doi: 10.1007/s42600-020-00097-1.
[34] Y. J. Huang et al., "Assessing Schizophrenia Patients Through Linguistic and Acoustic Features Using Deep Learning Techniques," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30. pp. 947–956, 2022. doi: 10.1109/TNSRE.2022.3163777.
[35] J. Zhang, H. Yang, W. Li, Y. Li, J. Qin, and L. He, "Automatic Schizophrenia Detection Using Multimodality Media via a Text Reading Task," Front. Neurosci., vol. 16, no. July, pp. 1–17, 2022, doi: 10.3389/fnins.2022.933049.