Abstract—Over the recent years, Schizophrenia has become a serious mental disorder that is affecting almost 21 million people globally. There are different symptoms to recognize schizophrenia from healthy people. It can affect the thinking pattern of the brain. Delusions, hallucinations, and disorganized speech are the common symptoms of Schizophrenia. In this study, we have used electroencephalography (EEG) signals to analyze and diagnose Schizophrenia using machine learning algorithms and found that temporal features performed well as compared to statistical features. EEG signals are the best way to analyze this disorder as they are intimately linked with human thinking patterns and provide information about brain activities. The present work proposes a Machine Learning (ML) model based on Logistic Regression (LR) along with two feature extraction libraries Time Series Feature Extraction Library (TSFEL) and MNE Python toolkit to diagnose Schizophrenia from EEG signals. The results are analyzed based on 5 different sampling techniques. The dataset was cross-validated using leave one subject out cross-validation (LOSOCV) using Scikit learn and achieve greater accuracy, sensitivity, specificity, macro average recall, and macro f1 score on temporal features respectively.

Keywords—Artificial intelligence (AI); Logistic Regression (LR); smote-class weight (S-CW); borderline smote-class weight (BS-CW); electroencephalography; Time Series Feature Extraction Library (TSFEL)

I. INTRODUCTION

Schizophrenia is one of the most common and severe mental disorders which is affecting more than 21 million people around the globe[1] and almost 50% of the total population of men [2] are suffering from this mental disease than women. This mental disorder directly affects the thinking pattern of human beings if it is not treated properly and can cause discrimination, stigmatization, and disobedience of human rights [3], [4].

Delusions, hallucinations, and disorganized speech [5] are the most common examples of this severe psychotic disorder. Hallucinations are sensory illusions that appear to be real [6] but are generated by your mind. Delusions are false beliefs that contradict reality and are not true. One cannot distinguish between what is real and what is imagined [7]. The patients having this psychotic disorder have a relative life expectancy is between 10 to 15 years and it also increases the risk of suicide to 10% which is not a great sign for human beings [8].

The diagnosis and analysis of schizophrenic patients can be done through the use of electroencephalogram (EEG) signals [9], [10]. The EEG signals have unique characteristics, variability, and dimensionality [11] as they can provide information about the electrical activities of a human brain [12], [13], and also they have the great potential to predict whether a person is a healthy control or schizophrenic [14]–[16]. In the medical field, EEG signals have vast applications like it can be used to detect epilepsy, coma, clinical death, and schizophrenia [17]. The scalp-based activity of EEG signals exhibits oscillations at different frequencies [18]. The main advantage of using EEG signal is that it is non-invasive, cheap, and possesses a high temporal resolution which gives a clear advantage over other techniques being used to diagnose schizophrenia [19].

According to researchers, there are five types of frequencies emitted by the human brain. Based on their frequency bands and locations they are categorized into delta (δ), theta (θ), alpha (α), beta (β), and gamma (γ) respectively which is shown in Table I [20].

Schizophrenia is classified into five categories according to the Diagnostic and Statistical Manual of Mental Disorders, 4th edition DSM-IV[21], [22]. The well-known five categories are further classified into positive and negative symptoms. Delusions and hallucinations are positive symptoms while avolition, alogia, and anhedonia are negative symptoms [23], [24].

| Activity | Frequencies (Hz) | Signal Characteristics | Behavior |
|----------|-----------------|------------------------|----------|
| Delta    | 0.005-0.004     | Lowest Frequency       | Sleeping |
| Theta    | 0.004-0.008     | Low Frequency          | Drowsy   |
| Alpha    | 0.008-0.013     | High Frequency         | Relaxing |
| Beta     | 0.013-0.03      | High Frequency         | Busy     |
| Gamma    | 0.03-0.04       | Highest Frequency      | Concentration |
Several techniques and methods have been employed to diagnose schizophrenia with the help of electroencephalography. In [25] the authors have used a thirteen (13) layers Convolutional Neural Network model to diagnose practical, normal, and seizure classes. In [26] they have used a deep learning algorithm such as CNN with random forest. They have implemented a voting layer to differentiate between those individuals who are at high-risk with schizophrenia and healthy controls. In [27] authors have used a CNN model to diagnose and evaluate different partitions of EEG to visualize unusual brain activities. In [28] authors have used EEG signals to compare real and imaginary music and classified them using CNN. In [29] they have used a cross trail encoding technique with the aid of convolutional autoencoders and used EEG signals. The dataset used in the training was very small. In [30] they have utilized a single electrode approach to classify and diagnose schizophrenia from EEG recordings. They have used the time-frequency technique to differentiate between schizophrenia and healthy controls. The authors have proposed a state-of-the-art model to recognize Alzheimer's disease with the help of logistic regression and achieved higher accuracy as compared to the domain knowledge-based handcrafted features [31].

The motivation behind this research is to evaluate the performance of different sampling techniques with the help of EEG signals for the diagnosis of schizophrenia and also to assist and verify the psychiatrist's decision because the diagnosis takes around 6-12 months as it is based on the questionnaire surveys.

The objective of this research is to develop a Machine Learning (ML) model and to compare the effect of different sampling techniques on the mentioned dataset that can validate the doctor’s decision and quickly diagnose this severe mental disorder.

The following is the structure of this research paper. The introduction is included in Section I. Section II explains the methodology and different python toolkits used in the experiment. Section III includes the results on three different techniques with filtered, no Z score, and no filter and unfiltered datasets. Section IV concludes the conclusion and how this study can be useful to diagnose Schizophrenia with the help of machine learning algorithms. Section V explains the future work.

II. METHODOLOGY

A. EEG Dataset and Preprocessing

The raw EEG data of fourteen (14) patients having schizophrenia, comprised of seven (7) males and seven (7) females with their average ages of 27.9 ± 3.3 and 28.3 ± 4.1 years, respectively. The experiment was carried out at the Institute of Psychiatry and Neurology in Warsaw, Poland [32]. Similarly, fourteen (14) healthy patients having no major disease were recruited of the same gender and same age group with seven (7) males and (7) females, respectively. The raw data were collected with the consent of all the participants.

Data were collected at a sample rate of 250 Hz using the International 10-20 system [33] as depicted in Fig. 1. Raw data were obtained when the patients were in a relaxed state with their eyes closed. The channels utilized to collect data were Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, and O2 respectively. The acquired EEG data was partitioned into different partitions and are considered stationary signals. Each segment had a window duration of 25 seconds (6250 samples). There were 1142 EEG segments in total, with each segment containing 6250 x 19 sample points and were normalized with Z-score before being sent to the logistic regression (LR).

B. Research Toolkits

1) Time Series Feature Extraction Library (TSFEL): TSFEL is one of the most effective available libraries of python to compute the extracted features of EEG signals. It helps the data scientist to evaluate a variety of domain knowledge features as well as handcrafted features. It can compute 60 distinct features which are extracted from statistical, temporal, and spectral domains [34].

2) MNE tool python: MNE Python toolkit [35] is an open-source python package used to evaluate and analyze human neurophysiological data such as EEG, MEG, sEEG, ECoG, NIRS, and many more. This toolkit is very helpful to visualize EEG signals. It can compute 28 univariate features and 6 bivariate features.

C. Proposed Methodology

The raw EEG data of schizophrenia was downloaded from the Repository of Open Data (RepOD), Department of Methods of Brain Imaging and Functional Research of Nervous System[36]. A bandpass filter of 0.1 Hz to 45 Hz was applied to remove the unwanted frequencies and data.
segmentation was done in the first phase. After data segmentation different features were extracted with the help of feature extraction libraries such as Time Series Feature Extraction Library (TSFEL) and MNE Python toolkit. Logistic Regression (LR) was utilized for classification and to differentiate the schizophrenic and healthy control patients depicted in Fig. 2.

![Proposed Model for Automated Recognition of Schizophrenia.](image)

**Fig. 2.** Proposed Model for Automated Recognition of Schizophrenia.

### III. EXPERIMENTAL RESULT

In the experimental phase, two feature extraction libraries are used such as TSFEL and MNE Python toolkit. The experimental results are divided into three phases. First, we have analyzed results by applying Z score normalization (Filtered), No Z score normalization and no filter in the second phase, and unfiltered data in the third phase. For the classification purpose, Logistic Regression (LR) was employed as a machine learning algorithm.

1) **Logistic regression on TSFEL filtered data:** LR was implemented on the TSFEL library with Z score normalization and applied five different sampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) with Class Weight abbreviated as S-CW, Borderline SMOTE Class Weight (BS-CW), Random Oversampling Class Weight (ROS-CW), None-Class Weight (N-CW) and None-None (N-N). It is found that S-CW and ROS-CW achieved an accuracy of 77.90% which is quite good when we have a smaller data size as shown in Table II and Fig. 3.

**TABLE II. SUMMARY OF LR ON TSFEL FILTERED DATA**

| Channels | Sampling Technique | Macro-recall | Macro F1 score | Sens. | Spec. | Acc.  |
|----------|--------------------|--------------|----------------|-------|-------|-------|
| All      | S-CW               | 78.07        | 77.89          | 75.34 | 80.81 | 77.90 |
| All      | BS-CW              | 77.77        | 77.62          | 75.51 | 80.03 | 77.63 |
| All      | ROS-CW             | 78.10        | 77.90          | 75    | 81.20 | 77.90 |
| All      | N-CW               | 77.37        | 77.37          | 78.59 | 76.16 | 77.45 |
| All      | N-N                | 77.70        | 77.53          | 75.17 | 80.23 | 77.54 |

2) **Logistic regression on TSFEL with No Z score & no filtered data:** In Table III, S-CW AND BS-CW achieved the higher accuracy of 82.27% and 82.72%, respectively when filtering and normalization was not implemented shown in Fig. 4.

**TABLE III. SUMMARY OF LR ON TSFEL NO Z SCORE AND NO FILTERED DATA**

| Channels | Sampling Technique | Macro-recall | Macro F1 score | Sens. | Spec. | Acc.  |
|----------|--------------------|--------------|----------------|-------|-------|-------|
| All      | S-CW               | 82.36        | 82.25          | 80.82 | 83.91 | 82.27 |
| All      | BS-CW              | 82.88        | 82.71          | 80.30 | 85.46 | 82.72 |
| All      | ROS-CW             | 82.27        | 81.80          | 74.82 | 89.72 | 81.81 |
| All      | N-CW               | 82.19        | 81.71          | 74.65 | 89.72 | 81.72 |
| All      | N-N                | 81.58        | 81.50          | 80.99 | 82.17 | 81.54 |

![Summary of LR on TSFEL No Z Score and No Filtered Data.](image)

**Fig. 3.** Summary of LR on TSFEL Filtered Data.

**Fig. 4.** Summary of LR on TSFEL No Z Score and No Filtered Data.
3) Logistic regression on TSFEL unfiltered data: In Table IV, LR was implemented on unfiltered data and interestingly the accuracy of N-CW was 79.36 which was very good as compared to other sampling techniques shown in Fig. 5.

**TABLE IV. SUMMARY OF LR ON TSFEL UNFILTERED DATA**

| Channels | Sampling Technique | Macro-recall | Macro F1 score | Sens. | Spec. | Acc. |
|----------|--------------------|--------------|----------------|-------|-------|------|
| All      | S-CW               | 79.17        | 78.90          | 74.82 | 83.52 | 78.90|
| All      | BS-CW              | 78.25        | 78.15          | 77.05 | 79.45 | 78.18|
| All      | ROS-CW             | 79.41        | 79.17          | 75.68 | 83.13 | 79.18|
| All      | N-CW               | 79.60        | 79.36          | 75.68 | 83.25 | 79.36|
| All      | N-N                | 78.25        | 78.15          | 77.05 | 79.45 | 78.18|

4) Logistic regression on MNE filtered data: In Table V, MNE-Python toolkit was used on Filtered data with Z score normalization and BS-CW accuracy was 77.45% as compared to other sampling techniques used in the experiment shown in Fig. 6.

5) Logistic regression on MNE with no Z score and no filter data: In Table VI, LR was used with MNE, and No Z Score, and No Filtered data, and S-CW performed better than other oversampling techniques. S-CW achieved an accuracy of 91.63% which is very good as shown in Fig. 7.

**TABLE V. SUMMARY OF LR ON MNE FILTERED DATA**

| Channels | Sampling Technique | Macro-recall | Macro F1 score | Sens. | Spec. | Acc. |
|----------|--------------------|--------------|----------------|-------|-------|------|
| All      | S-CW               | 77.13        | 76.98          | 75    | 79.26 | 77   |
| All      | BS-CW              | 77.61        | 77.44          | 75    | 80.23 | 77.45|
| All      | ROS-CW             | 76.86        | 76.71          | 74.65 | 79.06 | 76.72|
| All      | N-CW               | 77.03        | 76.89          | 75    | 79.06 | 76.90|
| All      | N-N                | 76.57        | 76.43          | 74.65 | 78.48 | 76.45|

6) Logistic regression on MNE unfiltered data: In Table VII, LR was used on the unfiltered data and S-CW and N-CW performed better and achieved the accuracies of 79.45% each shown in Fig. 8.

7) Comparison of logistic regression on TSFEL: After applying LR on three different datasets such as filtered, no z score, and no filter and unfiltered data we found that BS-CW has the highest unweighted macro recall value of 82.88 on no z score and no filter data as shown in Table VIII.
8) Comparison of logistic regression on MNE python:
Similarly, we have applied LR on the MNE Python toolkit to observe the performance of three different datasets and we have analyzed that N-CW has achieved the highest unweighted macro recall value of 80.49 as compared to others as shown in Table IX.

| Channels | Sampling Technique | Macro-recall | Macro F1 score | Sens. | Spec. | Acc. |
|----------|--------------------|--------------|---------------|-------|-------|------|
| All      | S-CW               | 79.37        | 79.37         | 80.65 | 78.10 | 79.45|
| All      | BS-CW              | 79.29        | 79.28         | 80.47 | 78.10 | 79.36|
| All      | ROS-CW             | 78.73        | 78.73         | 80.13 | 77.32 | 78.81|
| All      | N-CW               | 79.38        | 79.38         | 80.47 | 78.29 | 79.45|
| All      | N-N                | 78.68        | 78.71         | 80.82 | 76.55 | 78.81|

IV. Conclusion

In this paper, a deep learning model is proposed to diagnose schizophrenia from EEG signals as they contain information about the electrical activities of the human brain. We have proposed a machine learning model that can diagnose Schizophrenia from EEG signals. According to World Health Organization (WHO), it is affecting almost 21 million people worldwide and it is very hard to diagnose Schizophrenia as the treatment can take from 6 months to 1 year because doctors ask several questionnaires and take the survey from the patients. Different studies suggest that it can be found more in men than women. So, we need to take help from machine learning algorithms to diagnose this chronic mental disorder as quickly as possible. In our proposed model, we have used Logistic Regression (LR) as a classifier because it provides very good results when we have smaller datasets. We have evaluated the results in three different domains. First filtered data with Z score normalization, then without Z score normalization, and finally on the unfiltered data. We have used 5 different sampling techniques like SMOTE Class Weight (S-CW), Borderline SMOTE Class Weight (BS-CW), Random oversampling Class Weight (ROS-CW), None Class Weight (N-CW), and None-None (N-N), respectively. From our observation we have analyzed that the results achieved with no z score and no filter have the highest unweighted macro recall value it is due to the EEG recordings obtained from 14 SZ and 14 HC people does not have artifacts. It is also observed that when we have applied some filtering techniques so the ML model performance significantly decreased.

For cross-validation of ML model the leave-one-subject-out cross-validation technique using Scikit Learn has been utilized to validate the results in the form of evaluations parameters macro recall, macro f1 score, sensitivity, specificity, and accuracy, respectively.

V. Future Work

This research has still some limitations as the proposed model can predict and diagnose schizophrenic patients (SP) and the healthy control (HC) from the EEG signals. It cannot predict the disease severity. Also, this experiment has been done on the smaller dataset, but it can be carried out on the larger datasets as well to verify the model’s accuracy.

ACKNOWLEDGMENT

We would like to express our gratitude to Syed Zafi Sherhan Shah for his helpful, constructive talks and support to carry out the related experiments.

REFERENCES

[1] 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.

[2] S. L. James et al., “Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017,” Lancet, vol. 392, no. 10159, pp. 1789–1858, 2018.
Are early measured resting state EEG parameters predictive for upper limb motor impairment six months poststroke?,” Clin. Neurophysiol., vol. 132, no. 1, pp. 56–67, 2021.

B. J. Vorderwülbecke, A. G. Baroumand, L. M. N. F. Rodríguez Vidal, “Deep learning for emotion recognition using brain connectivity and its Spatial Information Seong-eun Moon Soo Seok Jung Lee Republic of Korea,” 2018 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 2556–2560, 2018.

R. Akhter, K. Lawal, and S. A. Mazumder, “Classification of Common and Uncommon Tones by P300 Feature Extraction and Identification of Accurate P300 Wave by Machine Learning Algorithm,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 10, pp. 646–652, 2020.

B. J. Vorderwülbecke, A. G. Baroumand, L. Spinelli, M. Seeck, P. van Mierlo, and S. Vulliémoz, “Automated interictal source localisation based on high-density EEG,” Seizure, vol. 92, no. October, pp. 244–251, 2021.

M. Saes, C. G. M. Meskers, A. Daffertshofer, E. E. H. van Wegen, and G. Kwakkel, “Are early measured resting-state EEG parameters predictive for upper limb motor impairment six months poststroke?,” Clin. Neurophysiol., vol. 132, no. 1, pp. 56–62, 2021.

A. Craik, Y. He, and J. L. Contreras-Vidal, “Deep learning for electroencephalogram (EEG) classification tasks: A review,” J. Neural Eng., vol. 16, no. 3, 2019.

H. U. Amin, W. Mumtaz, A. R. Subhani, M. N. M. Saad, and A. S. Malik, “Classification of EEG signals based on pattern recognition approach,” Front. Comput. Neurosci., vol. 11, no. November, pp. 1–12, 2017.

R. T. Schirrmeister et al., “Deep learning with convolutional neural networks for EEG decoding and visualization,” Hum. Brain Mapp., vol. 38, no. 11, pp. 5391–5420, 2017.

T. Bose, S. D. Sivakumar, and B. Kesavamurthy, “Identification of Schizophrenia Using EEG Alpha Band Power During Hyperventilation and Post-hyperventilation,” J. Med. Biol. Eng., vol. 36, no. 6, pp. 901–911, 2016.

B. Kaliappan, K. Rajamanickam, and S. Jayapal, “A Novel Classifier Algorithm for EEG Signal Based Person Authentication from Cz Channel with 2D-Wavelet Compression for the Online Voting System Using Touch Panel,” Aust. J. Basic Appl. Sci., vol. 8, no. 5, pp. 399–409, 2014.

R. De Filippis et al., “Machine learning techniques in a structural and functional MRI diagnostic approach in schizophrenia: A systematic review,” Neuropsychiatr. Dis. Treat., vol. 15, pp. 1605–1627, 2019, doi: 10.2147/NDT.S202418.

W. Liang, L. Cheng, and M. Tang, “Identity Recognition Using Biological Electroencephalogram Sensors,” J. Sensors, vol. 2016, 2016.

D. McLean et al., “DSM-IV ‘criterion A’ schizophrenia symptoms across ethnically different populations: evidence for differing psychotic symptom content or structural organization?,” Cult. Med. Psychiatry, vol. 38, no. 3, pp. 408–426, 2014.

R. Jardri et al., “Hallucination Research: Into the Future, and beyond,” Schizophr. Bull., vol. 45, no. 1, pp. S1–S4, 2019.

R. Tandon et al., “Definition and description of schizophrenia in the DSM-5,” Schizophr. Res., vol. 150, no. 1, pp. 3–10, 2013.

M. F. Green and P. D. Harvey, “Cognition in schizophrenia: Past, present, and future,” Schizophr. Res. Cogn., vol. 1, no. 1, pp. e1–e9, 2014.

U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, “Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals,” Comput. Biol. Med., vol. 100, pp. 270–278, 2018.

L. Chu, R. Qiu, H. Liu, Z. Zhang, and J. Wang, “Individual Recognition in Schizophrenia using Deep Learning Methods with Random Forest and Voting Classifiers: Insights from Resting State EEG Streams,” pp. 1–7, 2017, [Online]. Available: http://arxiv.org/abs/1707.03467.

S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, and U. R. Acharya, “Deep Convolutional Neural Network Model for Automated Diagnosis of Schizophrenia Using EEG Signals,” Appl. Sci., vol. 9, no. 14, p. 2870, 2019.

S. Sarkar, K. K. Reddy, A. Dorgan, C. Fidopisiastis, and M. Giering, “Wearable EEG-based activity recognition in PHM-related service environment via deep learning,” Int. J. Progn. Heal. Manag., vol. 7, no. Special Issue 6, 2016.

S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, “Deep Feature Learning for EEG Recordings,” Archiv, vol. abs/1511.0, 2015, [Online]. Available: http://arxiv.org/abs/1511.04306.

Z. Dvey-Aharon, N. Fogelson, A. Peled, and N. Intrator, “Schizophrenia detection and classification by advanced analysis of EEG recordings using a single electrode approach,” PLoS One, vol. 10, no. 4, pp. 1–12, 2015.

M. Lech and E. Pirogova, “Automated Recognition of Alzheimer’s Dementia Using Bag-of-Deep-Features and Model Ensembleing,” IEEE Access, vol. 9, pp. 88377–88390, 2021.

S. L. Oh, J. Vicnesh, E. J. Ciaccio, R. Yuvaraj, and U. R. Acharya, “Deep convolutional neural network model for automated diagnosis of Schizophrenia using EEG signals,” Appl. Sci., vol. 9, no. 14, 2019.

G. M. Rojas, C. Alvarez, C. E. Montoya, M. de la Iglesia-Vayá, J. E. Cistermans, and M. Gálvez, “Study of resting-state functional connectivity networks using EEG electrodes position as seed,” Front. Neurosci., vol. 12, no. APR, pp. 1–12, 2018.

M. Barandas et al., “TSFEL: Time Series Feature Extraction Library,” SoftwareX, vol. 11, p. 100456, 2020.

A. Gramfort et al., “MEG and EEG data analysis with MNE-Python,” Front. Neurosci., vol. 7, no. 7 DEC, pp. 1–13, 2013.

W. Olejarczyk, Elzbieta and Jernejczyk, “EEG in schizophrenia,” RepOD, 2017, [Online]. Available: https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441.