Comparative Analysis of ML and CNN Modes for Predicting Schizophrenia Disease using EEG data

Suraj Jaiswal

Indian Institute of Technology, Gandhinagar, Gujarat 382355
jaiswalsuraj@iitgn.ac.in

1 Abstract

One of the main causes of impairment worldwide is mental disorders. Accurate diagnosis is the first step in treating these disorders, but because there aren't any well-established clinical diagnostics, doing so is difficult. As we discuss in this paper, machine learning algorithms could be able to offer a potential answer to this issue. We describe and compare some techniques for the detection of mental disease diagnosis automatically based on Machine Learning algorithms and deep learning on EEG time series data. This approach can classify patients with Alzheimer's disease. We show that this same approach can classify patients with Schizophrenia with high accuracy.

Keywords: automated detection system; schizophrenia; deep learning; deep learning algorithm

2 Introduction

2.1 Big picture

The **big picture** is to improve the accuracy of diagnosis for mental disorders, which is currently difficult due to the lack of well-established clinical diagnostics. The research discusses ML and DL techniques applied to EEG time series data to detect mental diseases automatically. The paper specifically focuses on Alzheimer's disease using EEG signal data.

2.2 relevant literature review

The **relevant literature review** includes previous studies on the use of EEG data for diagnosis and machine learning algorithms for classification. Alzheimer's disease causes brain tissue to deteriorate, which impairs cognitive and social skills. Another common neurological disease is schizophrenia. It causes hallucinations, incoherent thinking, delusions, decreased intellectual functioning, difficulty in expressing emotions, and agitation. Currently, 25 million people suffer from Alzheimer's, while 26 million people with Schizophrenia.

2.3 Gaps

The **gap** in the current research is the lack of accurate diagnostic tools for mental disorders, which can be addressed by using machine learning algorithms. To provide a quantitative evaluation of mental disorders, methods based on Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), and Positron Emission Tomography (PET) have been used to aid professionals in the diagnostic process, but these are expensive.

2.4 Current Hypothesis

The **current hypothesis** is that machine learning algorithms can accurately classify patients with Alzheimer's disease and Schizophrenia based on EEG time series data. The application of machine learning and deep learning algorithms is a non-invasive and cost-efficient alternative for making an accurate diagnosis of these diseases, which is essential for effective therapy.

3 Schizophrenia Dataset

EEG in Schizophrenia: The dataset comprised 14 patients with paranoid schizophrenia and 14 healthy controls. Data were acquired with the sampling frequency of 250 Hz using the standard 10-20 EEG montage with 19 EEG channels: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, O2. The reference electrode was placed between electrodes Fz and Cz.

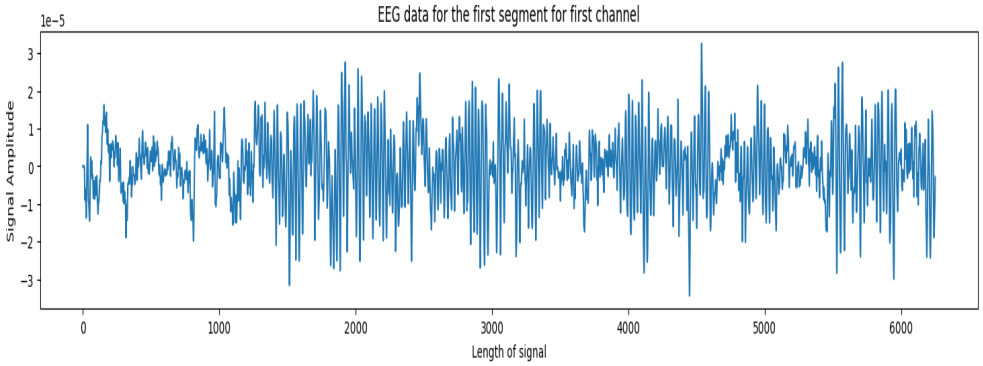

Fig. 1. EEG data for the first segment, the first channel

4 Algorithm Used

ML models and Deep Neural Networks Convolutional Neural Networks.

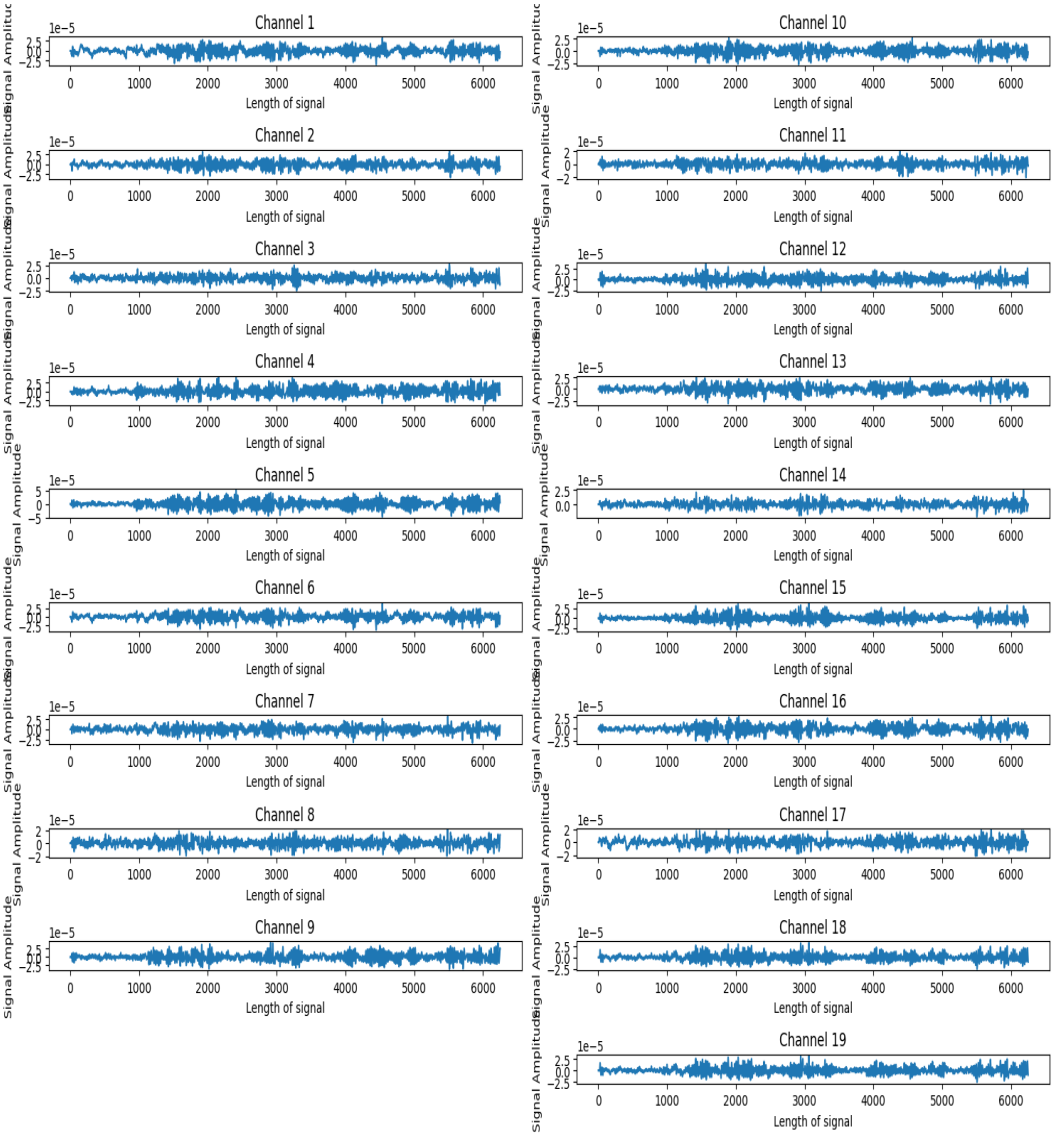

Fig. 2. EEG data for the first segment for all 19 channels

5 Methodology

5.1 Data loading and Preprocessing

In this section, we address certain technical issues, including dataset loading and feature extraction. The dataset comprised 14 patients with paranoid schizophrenia and 14 healthy controls in European Data Format (EDF): a standard file format designed for the exchange and storage of medical time series data. The dataset is time series data. We load it using the Glob module of the Python library, which imports all data from the user machine into our code. Using MNE:(an Open-source Python package for exploring, visualizing, and analyzing human neuro-physiological data), we fetch the raw EEG data in .edf format using the method $mne.io.read_raw_edf$. After loading the data, we do the following for each 28 edf file:

1)$data.set_eeg_reference()$: This line applies an average reference to the EEG data. The reference channel is calculated as the average of all channels and subtracted from each channel. This is done to reduce the noise and improve the quality of data.

2)$datax.filter(low_freq = 1, high_freq = 45)$: This line applies a bandpass filter to the data, which removes any frequencies outside the range of 1-45 Hz. This is done to reduce noise and remove any unwanted signals outside the frequency range of interest.

3)$epochs = mne.make_fixed_length_epochs(data, duration = 25, overlap = 0)$: This line segments the continuous data into epochs of fixed length. The duration argument specifies the length of each epoch in seconds, and the overlap argument specifies the amount of overlap between adjacent epochs. This will make the data in the form of (segments, channels, and length of signal) Where: $Lengthofsignal = duration * sampling_frequency$

We can see this is a two-class classification problem, so we make labels each segment as 0 for the healthy patient and 1 for those with Schizophrenia. Now this makes data ready to be passed to ML models and Convolutional Neural Networks.

5.2 Applying all major Machine Learning Models

We create features of the data which are mean, standard deviation, variance, minimum, maximum, argmin, argmax, mean square, root mean square, absolute difference in signals, skewness, and kurtosis. We split this data into train and test sets.

This is fed to LazyClassifier class of the LazyPredict library. This applies to all major ML models on the data. Following are all ML models LGBMClassifier, XGBClassifier, ExtraTreesClassifier, Random Forest Classifier, AdaBoost Classifier, Bagging Classifier, Linear Discriminant Analysis, Linear Suppor Vector Classifier, Calibrated Classifier CV, Logistic Regression, Stochastic Gradient Descent Classifier, Ridge Classifier, Decision Tree Classifier, perceptron, Passive

Aggressive Classifier, Support Vector Classifier, Bernoulli Naive Bayes, K Neighbors Classifier, Nearest Centroid, Gaussian Naive Bayes, etc. We took accuracy and time as metrics to select the best Machine learning model.

5.3 Convolutional Neural Network

CNN. No need to create features as it automatically finds patterns using filters in a series of Convolutional layers and MaxPool layers. This CNN architecture consists of 32,098 parameters which are all trainable parameters. We do cross-validation with kfold = 5 and take an average accuracy obtained for each fold for 100 epochs. We can see the architecture of the CNN model in the figure.

6 Results

The present study investigated the performance of various machine learning algorithms and a convolutional neural network (CNN) for the classification of patients suffering from Schizophrenia Disease using EEG signals. The study used Google Colab with 12.7 GB as max system RAM and 12 GB GPU RAM for implementation.

The CNN architecture comprised two convolutional layers, two max-pooling layers, and two dropout layers, followed by a global average pooling layer and two fully connected layers. The CNN model achieved an accuracy of 0.92 and took 596 seconds for training. The machine learning models employed in the study included LGBMClassifier, XGBClassifier, ExtraTreesClassifier, RandomForestClassifier, AdaBoostClassifier, BaggingClassifier, LinearDiscriminantAnalysis, LinearSVC, CalibratedClassifierCV, LogisticRegression, SGDClassifier, RidgeClassifier, DecisionTreeClassifier, RidgeClassifierCV, Perceptron, PassiveAggressiveClassifier, QuadraticDiscriminantAnalysis, NuSVC, SVC, and ExtraTreeClassifier. The accuracy of these models ranged from 0.72 to 0.95. Among these models, LGBMClassifier achieved the highest accuracy of 0.95, followed by XGBClassifier (0.93) and ExtraTreesClassifier (0.92).

The CNN model took the longest time for training (596 seconds), followed by LGBMClassifier (5.16 seconds), XGBClassifier (4.38 seconds), and CalibratedClassifierCV (2.56 seconds). The remaining models took less than 2 seconds for training. The results suggest that the CNN model and machine learning algorithms can potentially classify patients suffering from Schizophrenia using EEG signals. The highest accuracy achieved by LGBMClassifier suggests that this algorithm is better suited for this task. However, the longer training time required for the CNN model may be a limiting factor in real-world applications.

Overall, the study provides important insights into the performance of machine learning algorithms and a CNN model for the classification of patients suffering from Schizophrenia Disease using EEG signals. The findings can be useful for developing more accurate and efficient emotion recognition systems that can be applied in various domains, including healthcare, education, and entertainment.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_4 (Conv1D)           (None, 6248, 32)          1856

 conv1d_5 (Conv1D)           (None, 6246, 32)          3104

 max_pooling1d_2 (MaxPooling  (None, 3123, 32)         0
 1D)

 dropout_3 (Dropout)         (None, 3123, 32)          0

 conv1d_6 (Conv1D)           (None, 3121, 64)          6208

 conv1d_7 (Conv1D)           (None, 3119, 64)          12352

 max_pooling1d_3 (MaxPooling  (None, 1559, 64)         0
 1D)

 dropout_4 (Dropout)         (None, 1559, 64)          0

 global_average_pooling1d_1   (None, 64)               0
 (GlobalAveragePooling1D)

 dense_2 (Dense)             (None, 128)               8320

 dropout_5 (Dropout)         (None, 128)               0

 dense_3 (Dense)             (None, 2)                 258

=================================================================
Total params: 32,098
Trainable params: 32,098
Non-trainable params: 0
```

Fig. 3. architecture of CNN model

Model	Accuracy	Time Taken
Convolutional Neural Networks	0.92	596
LGBMClassifier	0.95	5.16
XGBClassifier	0.93	4.38
ExtraTreesClassifier	0.92	0.79
RandomForestClassifier	0.92	2.62
AdaBoostClassifier	0.92	2.56
BaggingClassifier	0.91	1.73
LinearDiscriminantAnalysis	0.90	0.36
LinearSVC	0.89	0.74
CalibratedClassifierCV	0.88	2.56
LogisticRegression	0.87	0.23
SGDClassifier	0.87	0.15
RidgeClassifier	0.86	0.13
DecisionTreeClassifier	0.86	0.41
RidgeClassifierCV	0.85	0.29
Perceptron	0.84	0.19
PassiveAggressiveClassifier	0.84	0.22
QuadraticDiscriminantAnalysis	0.83	0.26
NuSVC	0.81	0.44
SVC	0.73	0.49
ExtraTreeClassifier	0.72	0.03

Fig. 4. Accuracy and time(in seconds) for all ML models and CNN model

7 Discussion

7.1 Interpretation of results

We explored the potential of machine learning algorithms and a convolutional neural network (CNN) for classifying patients suffering from schizophrenia using EEG time series data. The CNN model achieved an accuracy of **0.92**, and the machine learning models achieved accuracies ranging from 0.72 to 0.95, with LGBMClassifier achieving the highest accuracy of **0.95**. However, the CNN model required significantly longer training time compared to the other models.

7.2 Comparison to other works

The proposed model in the schizophrenia research paper(link in reference) generated classification accuracies of 81.26 percent for subject-based testing using CNN; My model produces 92 percent accuracy for CNN and 95 percent accuracy using the Machine learning algorithm- LGBM Classifier.

The study's results suggest that both CNN and machine learning algorithms have the potential for disease classification using EEG signals. The findings can aid in developing more accurate and efficient disease detection systems for various diseases. The model developed in our study and described herein could potentially also be used to diagnose other neurological disorders such as Alzheimers,

Parkinson's disease, and epilepsy. Apart from the CNN model, other deep learning methods, such as long short-term memory (LSTM) and autoencoders, could also be explored in the diagnosis of SZ.

7.3 Limitations

CNN method Limitation 1)Small data pool: The CNN model used in the proposed system was developed using a small data pool of only 14 healthy subjects and 14 SZ patients. This may limit the generalizability of the model to larger and more diverse datasets, as the model may not capture the full range of variability in brain activity among individuals with SZ.

2)Costly computation: Compared to traditional machine learning techniques, CNN is computationally expensive and may require specialized hardware to run efficiently. This can limit the scalability of the proposed system, particularly for applications that require real-time analysis of brain activity.

3)Lack of interpretability: Although the proposed system achieves high classification accuracy, the underlying features that the CNN model uses to make its predictions are not easily interpretable. This limits our understanding of the biological mechanisms underlying SZ and may hinder the development of targeted interventions for the disorder.

ML methods Limitations 1) Need for feature engineering: Most traditional machine learning methods require feature engineering, which involves manually selecting and extracting relevant features from the raw data. This can be time-consuming and requires domain expertise.

2) Sensitivity to noise and outliers: Traditional machine learning methods can be sensitive to noisy or outlier data points, which can lead to overfitting or underfitting of the model.

3) Limited capacity for complex patterns: Traditional machine learning methods may struggle to identify complex patterns in high-dimensional data, such as images or audio. This can limit their ability to perform well on certain types of classification tasks.

4) Limited scalability: Traditional machine learning methods may struggle to scale up to handle very large datasets or high-throughput applications.

5)Limited ability to learn from new data: Once a traditional machine learning model has been trained, it may be difficult to update or adapt it to new data without retraining the entire model from scratch. This can limit its ability to learn from new data in real-time applications.

7.4 Future Work

Future work can include exploring larger datasets to evaluate the models' performance further. Additionally, it would be interesting to investigate the transferability of the trained models to other datasets and populations. Further research can also focus on optimizing the hyperparameters of the models to improve their

performance. Finally, the inclusion of other types of data, such as genetic and behavioral data, can be explored to enhance the accuracy of the classification models.

8 Conclusion

In conclusion, the study demonstrates the potential of both machine learning algorithms and a convolutional neural network for accurately classifying patients suffering from Schizophrenia Disease using EEG signals. The results show that **LGBMClassifier** achieved the highest accuracy among the machine learning models, which is **0.95**, while the **CNN model** achieved an accuracy of **0.92**. Although the CNN model took the longest time to train, it still has the potential for practical applications.

In the near future, we intend to use a larger data set to test our model, further optimize the models, and combine the web-based cloud method to identify the early stages of Schizophrenia.

References

1. Detection of Early Stage Alzheimer's Disease using EEG Relative Power with Deep Neural Network by Donghyeon Kim; Kiseon Kim https://ieeexplore.ieee.org/abstract/document/8512231
2. Deep Convolutional Neural Network Model for Automated Diagnosis of Schizophrenia Using EEG Signals by Shu Lih Oh, Jahmunah Vicnesh, Edward J Ciaccio, Rajamanickam Yuvaraj, and U Rajendra Acharya. https://www.mdpi.com/2076-3417/9/14/2870
3. Schizophrenia disease EEG data-set : https://repod.icm.edu.pl/dataset.xhtml?persistentId=doi:10.18150/repod.0107441
4. Github link for my code for the above implementation : https://github.com/jaiswalsuraj487/EEG$_{T}imeSeriesData_{C}lassification_{t}ask$

overleaf link: https://www.overleaf.com/read/grppzwcpwggw