Epilectic Cases Detection, a Big Data Approach

OSSON Sergio Suzerain, TANG Elody

Université de Paris Cité, France

Abstract

This paper presents three methods to detect and distinguish between healthy and epileptic cases, without the tedious procedures of manual feature extraction. We will explore three classification methods for detecting epilepsy. The first method uses the original values of the EEG signals to perform the classification. The second method uses discrete wavelet transformation to extract features from the EEG signals. Finally, the third method uses a bidirectional recurrent neural network (biLSTM) to learn the temporal features of the EEG signals and use them for classification. The performance of these methods is evaluated on the database of the University of Bonn and shows results comparable to state-of-the-art approaches.

1. Introduction

According to the World Health Organisation [8], epilepsy is a chronic, non-communicable brain disease that affects approximately 50 million people worldwide. It is characterised by recurrent epileptic seizures, which can manifest themselves in different ways depending on the area of the brain affected.

It is a very distressing disease for those affected, as well as for their families and loved ones, as the seizures can be very violent. They can lead to convulsions and loss of consciousness, and cause physical injury. In addition, people with epilepsy may experience fatigue, anxiety and depression, as well as difficulty concentrating and remembering information. These symptoms can have a significant impact on patients' daily lives, such as difficulty performing work or school tasks, driving or participating in social activities.

For this reason, the detection and appropriate treatment of epilepsy is essential to help patients manage their disease and improve their quality of life. So far, most of the methods proposed to detect and predict epilepsy rely on the extraction of features by hand from electroencephalographic (EEG) signals. However, these methods can be time-consuming and tedious, which limits their usefulness in a clinical setting.

To develop new methods for epilepsy detection and treatment, we have investigated newer approaches, exploring the use of machine learning and deep neural networks. These methods show great potential for improving the detection and treatment of epilepsy. In this report, we present three different methods to distinguish between healthy and epilectic cases.

Having introduced epilepsy and the challenges it presents to people with the condition, their families and healthcare professionals, it is essential to review existing research in this area. This work provides us with a better understanding of the issues and objectives of epilepsy detection and treatment, and the methods used to date. This gives us the keys to developing new, innovative and effective approaches to the detection and treatment of this disease.

2. Related Work

A number of research studies on the detection and treatment of epilepsy have already made significant progress in this area. This work has mainly focused on the detection of epilepsy from electroencephalographic (EEG) signals, using signal processing techniques and machine learning. In this section, we will present a review of existing work in this area, in order to better understand the methods already used and the results obtained.

- The paper "Big Data in Cognitive Neuroscience: Opportunities and Challenges", written by Kamalaker Dadi and Bapi Raju Surampudi (2023) [5] introduces the field of cognitive neuroscience, which studies the neural basis of mental abilities such as perception, learning, vision and language. The challenge is to construct the cognitive functional architecture of the brain by observing brain regions. To this end, he discusses the identification of brain regions with specific functions and mentions the morphometric characteristics of brain structures and its behaviour on the human body. It is important to note that some neurological diseases, such as epilepsy, have a significant impact on brain function, which is why cognitive neuroscience has a key role to play in understanding the brain mechanisms involved in epilepsy. This article also introduces functional magnetic resonance imaging (IRMf) [9] which is a dominant tool in cognitive neuroscience. fMRI is used to map signal changes in brain areas. Maps are then obtained to analyse brain activity, which can be very useful in determining the areas that trigger an epileptic seizure. Massive data in cognitive neuroscience offer exciting opportunities to understand brain function on a large scale and to develop new theories of brain function. However, there are still challenges to be overcome, such as properly removing noise from fMRI signals.

April 27, 2023

- The paper "Multimodal detection of epilepsy", written by Loukas Ilias, Dimitris Askounis and John Psarras (2022) [6] describes two new methods for distinguishing between different cases in epilepsy. The first method involves using the Short Term Fourier Transform (STFT) to generate a three-channel image from a single-channel electroencephalogram (EEG) signal, which is then fed to pre-trained deep learning models. The second method involves running each EEG signal through two convolutional neural networks (CNNs) to extract the low and high frequency features, as well as using STFT to generate an image. The results show that the proposed methods perform well. Furthermore, these experiments showed that each component of the models was effective for epilepsy classification. The authors plan to extend their method to multimodal models for epilepsy detection using magnetic resonance imaging data, and to explore multi-channel EEG data and data imbalances in future work.

- The article "Machine Learning Algorithms for Epilepsy Detection Based on Published EEG Databases: A Systematic Review", written by Andreas Miltiadous , Katerina Tzimourta, Nikolaos Giannakeas, Markos G. Tsipoura, Euripis Glavas, Konstantinos Kalafatakis and Alexandros T. Tzallas (2023) [1] presents a review of methods for automatic detection of epilepsy from electroencephalogram (EEG) over the last five years. Researchers have developed signal processing and machine learning techniques based on the observation of epileptic seizures and the analysis of the results of tests such as EEG, brain imaging, blood and urine tests, and neurological examination. In the text, several techniques for detecting epilepsy are mentioned:

1. **EEG (electroencephalography) data analysis:** this technique involves recording the electrical activity of the brain using electrodes placed on the scalp. EEG data analysis is used to detect abnormalities that characterise epilepsy, such as epileptiform discharges.

2. **ECG (electrocardiography) data analysis:** this technique involves recording the electrical activity of the heart using electrodes placed on the skin. Analysis of ECG data can reveal cardiac abnormalities that are often associated with epilepsy.

3. **EMG (electromyography) data analysis:** this technique involves recording the electrical activity of muscles using electrodes placed on the skin. Analysis of EMG data can reveal involuntary movements that are often associated with epilepsy.

4. **Use of machine learning:** This technique involves training algorithms to recognise patterns that characterise epilepsy from EEG, ECG or EMG data. These algorithms can then be used to automatically detect epilepsy in patients.

In summary, the epilepsy detection techniques mentioned in the text are based on the analysis of electrical data from the brain, heart and muscles, and the use of machine learning to detect abnormalities associated with epilepsy.

Common to all three texts is the use of data analysis techniques for the detection of epilepsy. In particular, the analysis of electrophysiological data, such as EEG, is mentioned in all three texts. Furthermore, the second and third texts also mention the use of deep learning and machine learning techniques for the detection of epilepsy from electrophysiological data. Finally, the second and third texts also mention the possibility of using multimodal data, such as brain imaging data, to improve epilepsy detection.

- The paper "Accelerating Data Analysis in Simulation Neuroscience with Big Data Technologies" [4] discusses the challenges of analyzing large-scale neural simulations and proposes a framework that combines big data technologies and neuroscience-specific tools to accelerate the data analysis process.

They begin by highlighting the increasing size and complexity of neural simulations, which can generate massive amounts of data that are difficult to manage and analyze using traditional methods. They discuss the importance of efficient data analysis in order to gain insights into the functioning of neural circuits and to advance our understanding of brain function and disease.

The paper then introduces a framework called the Neurolytics framework , alongside a new data layout, which leverages big data technologies such as Apache Spark and Hadoop to enable fast, scalable, and distributed data processing. The authors describe how this framework can be used to process large-scale neural simulation data, including spike trains and time-series data, in an efficient and automated way.

In addition to big data technologies, the framework also includes neuroscience-specific tools such as the NEURON simulation environment. These tools enable the integration of simulation data with other types of neuroscience data, such as neuroimaging and electrophysiological recordings, and facilitate the exploration of complex neural circuit dynamics.

The paper concludes by demonstrating the effectiveness of the framework in a case study involving the simulation of a large-scale cortical model. The authors show that the framework can significantly accelerate the data analysis process, allowing for faster discovery of relevant features and more efficient exploration of complex neural dynamics.

Overall, the paper provides a valuable contribution to the field of simulation neuroscience by proposing a novel framework that addresses the challenges of analyzing large-scale neural simulations and enables more efficient and comprehensive exploration of neural circuit dynamics.

- The paper "Epileptic Seizure Prediction Based on Hybrid Seek Optimization Tuned Ensemble Classifier Using EEG Signals" [2] is about predicting epileptic seizures using electroencephalogram (EEG) signals, and specifically proposes a novel approach called the "Hybrid Seek Optimization Tuned Ensemble Classifier" (HSOTEC) for this task.

The HSOTEC approach proposed in the paper combines three different types of classifiers, AdaBoost, random forest (RF), and the decision tree (DT), using an optimization algorithm to create an ensemble model that is tuned specifically for predicting epileptic seizures. The paper reports on experiments using a publicly available EEG dataset, showing that the HSOTEC approach achieves higher accuracy and lower false positive rates than other commonly used methods for seizure prediction.

Overall, the paper provides a detailed technical description of the HSOTEC approach and its implementation, as well as experimental results demonstrating its effectiveness for predicting epileptic seizures using EEG signals.

- The paper "An Interpretable Deep Learning Classifier for Epileptic Seizure Prediction Using EEG Data" [3] proposes a deep learning model for predicting epileptic seizures using EEG (Electroencephalogram) data. The authors of the paper aim to develop a model that not only accurately predicts seizures but also provides interpretability to the predictions, allowing clinicians to better understand the reasoning behind the model's predictions.

The authors use a convolutional neural network (CNN) to extract features from the EEG data and then use a long short-term memory (LSTM) network to make the final prediction. They use a dataset of EEG recordings from patients with epilepsy and evaluate their model's performance using several metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

The authors also propose a visualization technique to help clinicians interpret the model's predictions. They use attention maps to highlight the regions of the EEG recordings that are most important for the model's prediction.

Overall, the paper demonstrates that the proposed model can achieve high accuracy in predicting epileptic seizures and provides interpretability to the predictions, which can help clinicians make more informed decisions about patient care.

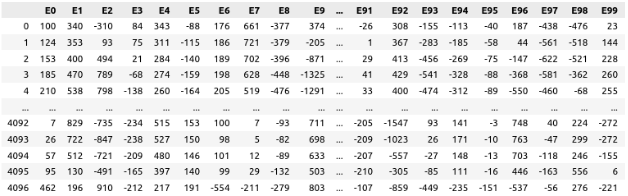
3. Experimentations

3.1. The database

In this experimentation, we used The Bonn University EEG Database, which is one of the largest and most comprehensive collections of EEG recordings in the world, with over 1,500 recordings from more than 500 subjects. The recordings were obtained using a variety of EEG systems, including the 10-20 system, the 10-10 system, and the 64-channel system, and cover a wide range of cognitive and motor tasks.

The database is maintained by the Department of Epileptology at the University of Bonn [7], and it is freely available to researchers around the world. It is an essential resource for researchers studying brain function and behavior, and has been used in countless studies on topics such as attention, perception, memory, language, and emotion.

In addition to the EEG recordings themselves, the database also includes detailed metadata, including subject demographics, experimental conditions, and task parameters. This metadata is essential for ensuring that the recordings are used appropriately and that the results of any research conducted using the database are accurate and reliable.

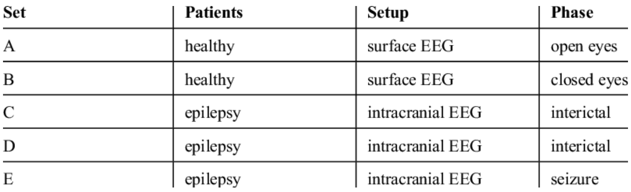


	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	...	E91	E92	E93	E94	E95	E96	E97	E98	E99
0	100	340	-310	84	343	-88	176	661	-377	374	...	-26	308	-155	-113	-40	187	-438	-476	23
1	124	353	93	75	311	-115	186	721	-379	-205	...	1	367	-283	-185	-58	44	-561	-518	144
2	153	400	494	21	284	-140	189	702	-396	-871	...	29	413	-456	-269	-75	-147	-622	-521	228
3	185	470	789	-68	274	-159	198	628	-448	-1325	...	41	429	-541	-328	-88	-368	-581	-362	260
4	210	538	798	-138	260	-164	205	519	-476	-1291	...	33	400	-474	-312	-89	-550	-460	-68	255
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4092	7	829	-735	-234	515	153	100	7	-93	711	...	-205	-1547	93	141	-3	748	40	224	-272
4093	26	722	-847	-238	527	150	98	5	-82	698	...	-209	-1023	26	171	-10	763	-47	299	-272
4094	57	512	-721	-209	480	146	101	12	-89	633	...	-207	-557	-27	148	-13	703	-118	246	-155
4095	95	130	-491	-165	397	140	99	29	-132	503	...	-210	-305	-85	111	-16	446	-163	556	6
4096	462	196	910	-212	217	191	-554	-211	-279	803	...	-107	-859	-449	-235	-151	-537	-56	276	-221

Figure 1: Set E of Bonn University EEG database

The Bonn University EEG Database is a collection of EEG recordings that has several characteristics, including the number of sets, frequency, and other features. Here are some of the key characteristics of the dataset:

- **Number of Sets:** The Bonn University EEG Database consists of five sets of EEG recordings, referred to as "sets A-E". Each set contains EEG recordings from different subjects and under different experimental conditions.



Set	Patients	Setup	Phase
A	healthy	surface EEG	open eyes
B	healthy	surface EEG	closed eyes
C	epilepsy	intracranial EEG	interictal
D	epilepsy	intracranial EEG	interictal
E	epilepsy	intracranial EEG	seizure

Figure 2: Overview of Bonn University EEG database

- **Sampling Frequency:** The EEG recordings in the database were sampled at a frequency of 173.61 Hz. This means that the voltage values for each electrode were recorded 173.61 times per second.

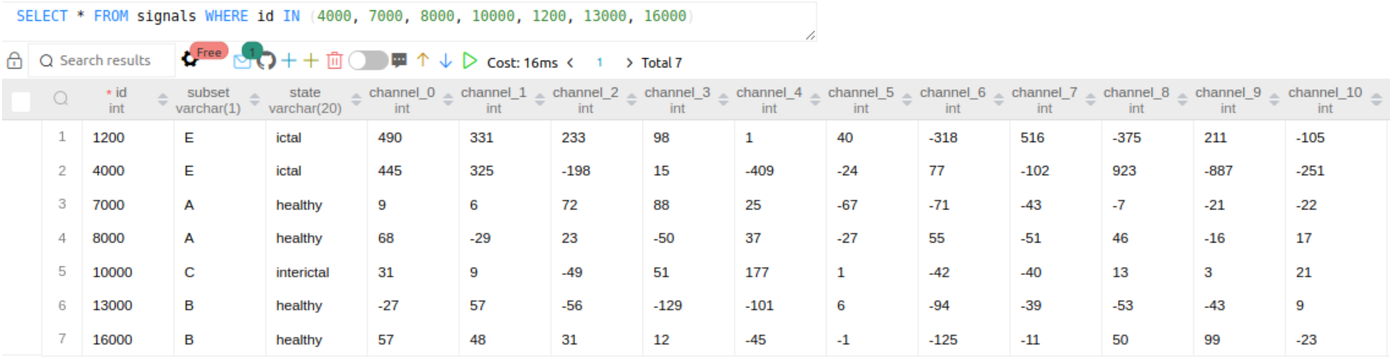
- **Electrode Configuration:** The EEG recordings in the database were obtained using a variety of electrode configurations, including the 10-20 system, the 10-10 system, and the 64-channel system. The specific electrode configuration used for each recording is included in the metadata for that recording.

- **Task and Stimulus Types:** The recordings in the database cover a wide range of cognitive and motor tasks, including auditory and visual stimuli, motor tasks, and cognitive tasks such as mental arithmetic and reading. The metadata for each recording includes information on the specific task and stimulus type used.

- **Duration:** The duration of each EEG recording varies, with some recordings lasting a few minutes and others lasting up to an hour.

- **File Format:** The EEG recordings are provided in European Data Format (EDF) files, which is a standard file format for EEG data.

These characteristics of the Bonn University EEG Database are important to consider when analyzing the data and designing experiments using this dataset 3.



```
SELECT * FROM signals WHERE id IN (4000, 7000, 8000, 10000, 1200, 13000, 16000)
```

	* id int	subset varchar(1)	state varchar(20)	channel_0 int	channel_1 int	channel_2 int	channel_3 int	channel_4 int	channel_5 int	channel_6 int	channel_7 int	channel_8 int	channel_9 int	channel_10 int
1	1200	E	ictal	490	331	233	98	1	40	-318	516	-375	211	-105
2	4000	E	ictal	445	325	-198	15	-409	-24	77	-102	923	-887	-251
3	7000	A	healthy	9	6	72	88	25	-67	-71	-43	-7	-21	-22
4	8000	A	healthy	68	-29	23	-50	37	-27	55	-51	46	-16	17
5	10000	C	interictal	31	9	-49	51	177	1	-42	-40	13	3	21
6	13000	B	healthy	-27	57	-56	-129	-101	6	-94	-39	-53	-43	9
7	16000	B	healthy	57	48	31	12	-45	-1	-125	-11	50	99	-23

Figure 3: Query of some data in the MySQL database

Overall, the Bonn University EEG Database is a valuable resource for researchers in the field of neuroscience. It provides a wealth of high-quality EEG recordings and metadata, and has played an important role in advancing our understanding of the human brain.

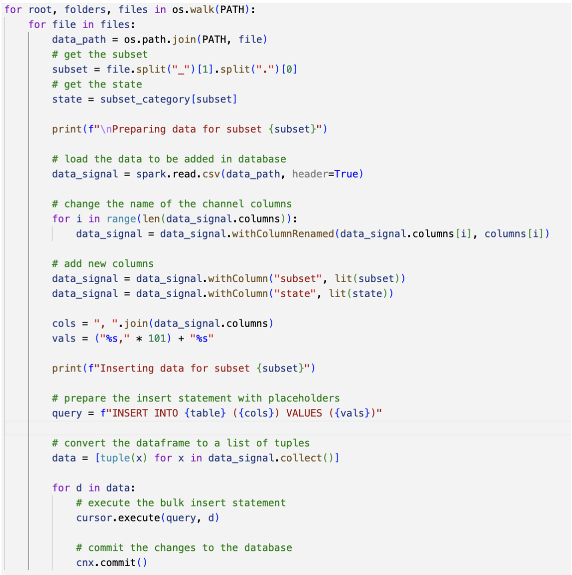
3.2. *The Database Management System*

The database management system (DBMS) used in the experimentation for storing the EEG signals of the Bonn University database is MySQL. MySQL is an open-source, relational DBMS that is widely used for web applications and is known for its ease of use, high performance, and scalability. It was chosen for this project due to its reliability, stability, and compatibility with a wide range of programming languages.

The process used for inserting the data into the MySQL database involved first establishing a connection to the MySQL server, creating a table with the appropriate columns and data types, and then iterating over the files in the data directory. For each file, the data was loaded into a Spark dataframe, the column names were changed to match the column names of the MySQL table, and new columns were added to specify the subset and state of the data. The dataframe was then converted to a list of tuples, and the data was inserted into the MySQL database using a prepared statement with placeholders. Finally, the changes were committed to the database and the connection was closed. This process was repeated for each file in the directory until all data had been successfully inserted into the database.

Our code an implementation of inserting EEG signal data from the Bonn University dataset into a MySQL database using Python and Spark. MySQL is a Relational Database Management System (RDBMS) that is widely used for storing and managing structured data. In this case, MySQL is being used to store EEG signal data in a structured format so that it can be easily queried and analyzed. The process of inserting the data involves several steps:

1. Establish a connection to the MySQL server using the ‘mysql.connector‘ library.
2. Define the table name and column names for the EEG signals.
3. Construct a ‘CREATE TABLE‘ statement using the table and column definitions.
4. Execute the ‘CREATE TABLE‘ statement to create the table in the MySQL database.
5. Load the EEG signal data using the Spark framework and rename the channel columns.
6. Add new columns to the Spark dataframe for the subset and state of the EEG signals.
7. Prepare an ‘INSERT INTO‘ statement with placeholders for the values to be inserted.
8. Convert the Spark dataframe to a list of tuples.
9. Execute the ‘INSERT INTO‘ statement using the ‘execute()‘ method of the MySQL cursor object.
10. Commit the changes to the MySQL database using the ‘commit()‘ method of the MySQL connection object.
11. Repeat steps 5-10 for all EEG signal files in the dataset.



```python
for root, folders, files in os.walk(PATH):
    for file in files:
        data_path = os.path.join(PATH, file)
        # get the subset
        subset = file.split("_")[1].split(".")[0]
        # get the state
        state = subset_category[subset]

        print(f"\nPreparing data for subset {subset}")

        # load the data to be added in database
        data_signal = spark.read.csv(data_path, header=True)

        # change the name of the channel columns
        for i in range(len(data_signal.columns)):
            data_signal = data_signal.withColumnRenamed(data_signal.columns[i], columns[i])

        # add new columns
        data_signal = data_signal.withColumn("subset", lit(subset))
        data_signal = data_signal.withColumn("state", lit(state))

        cols = ", ".join(data_signal.columns)
        vals = ("%s," * 101) + "%s"

        print(f"Inserting data for subset {subset}")

        # prepare the insert statement with placeholders
        query = f"INSERT INTO {table} ({cols}) VALUES ({vals})"

        # convert the dataframe to a list of tuples
        data = [tuple(x) for x in data_signal.collect()]

        for d in data:
            # execute the bulk insert statement
            cursor.execute(query, d)

            # commit the changes to the database
            cnx.commit()
```

Figure 4: Data insertion in MySQL Database

This implementation uses Spark to load and manipulate the EEG signal data because Spark is well-suited for processing large datasets in a distributed and parallelized way. The code also includes a dictionary that maps the different subsets of EEG signals to their corresponding categories (healthy, interictal, or ictal). This information is added as columns to the MySQL table so that it can be easily queried and analyzed later.

3.3. The Methods

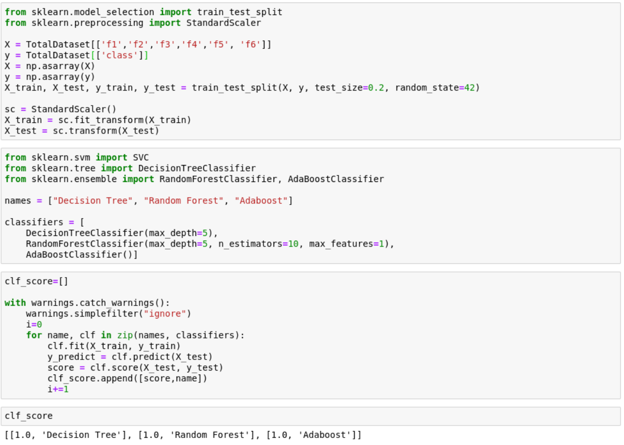
Now that we have seen how to set up the database, we will explore three classification methods for detecting epilepsy. The first method uses the original values of the EEG signals to perform the classification. The second method uses discrete wavelet transformation to extract features from the EEG signals. Finally, the third method uses a bidirectional recurrent neural network (biLSTM) to learn the temporal features of the EEG signals and use them for classification.

3.3.1. First Method : Classification with the original values

The first method we implemented is very simple: there is no pre-processing on the data. We applied machine learning models provided by the Scikitlearn library, in particular the decision tree, the random forest and Adaboost. We extracted six features: DFA (Detrended Fluctuation Analysis), HFD (Hjorth Fractal Dimension), SVD Entropy, Spectral Entropy, Fisher Information, et PFD (Petrosian Fractal Dimension), which are included in the pyeeg library. The six features extracted are measures of electrical activity in the brain that can be used to characterize epilepsy. Here is a brief explanation of each feature:

- DFA : Detrended fluctuation analysis is a measure of how the variance of the time series changes over time. This feature can be used to quantify the regularity of brain electrical activity.
- HFD : The Hjorth Fractal Dimension is a measure of the complexity of the time series. This feature can be used to quantify the complexity of brain electrical activity.
- SVD Entropy: The SVD (Singular Value Decomposition) entropy is a measure of the complexity of the time series. This feature can also be used to quantify the complexity of brain electrical activity.
- Spectral Entropy: Spectral entropy is a measure of the frequency diversity in a time series. This characteristic can be used to quantify the frequency diversity in brain electrical activity.
- Fisher Information: Fisher information is a measure of the sensitivity of a time series to changes in its parameters. This characteristic can be used to quantify the ability of brain electrical activity to respond to stimuli.
- PFD (Petrosian Fractal Dimension): The Petrosian Fractal Dimension is a measure of the roughness of the time series. This feature can be used to quantify the roughness of brain electrical activity.

Using these features, it is possible to characterize brain electrical activity and apply the above models to the detection of epilepsy. Using these six features, machine learning models' accuracy are as follows:



```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

X = TotalDataset[['f1','f2','f3','f4','f5', 'f6']]
y = TotalDataset[['class']]
X = np.asarray(X)
y = np.asarray(y)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

```
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier

names = ["Decision Tree", "Random Forest", "Adaboost"]

classifiers = [
    DecisionTreeClassifier(max_depth=5),
    RandomForestClassifier(max_depth=5, n_estimators=10, max_features=1),
    AdaBoostClassifier()]
```

```
clf_score=[]

with warnings.catch_warnings():
    warnings.simplefilter("ignore")
    i=0
    for name, clf in zip(names, classifiers):
        clf.fit(X_train, y_train)
        y_predict = clf.predict(X_test)
        score = clf.score(X_test, y_test)
        clf_score.append([score,name])
        i+=1
```

```
clf_score
```

```
[[1.0, 'Decision Tree'], [1.0, 'Random Forest'], [1.0, 'Adaboost']]
```

Figure 5: Code and Result of the first method

Furthermore, the method based on features extracted from EEG signals can be time-consuming and requires considerable expertise in selecting the most relevant features. For this reason, alternative methods have been proposed to simplify this step. One such method is based on the discrete wavelet transform (DWT).

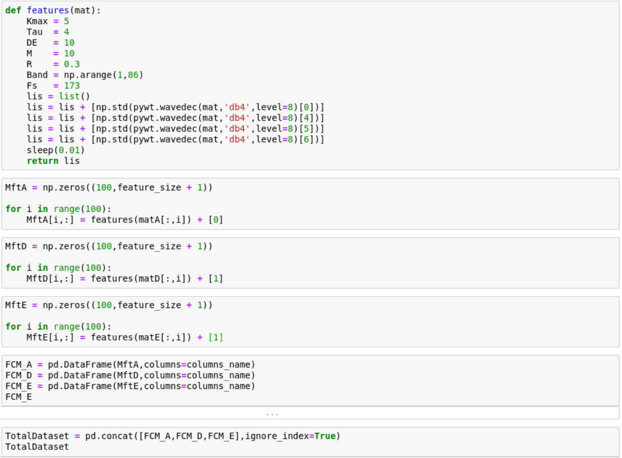
3.3.2. Second method: Classification with the discrete wavelet transform

The discrete wavelet transform is a multi-resolution signal decomposition method that allows the analysis of the different frequency components of the signal. The advantage of this method is that it provides a dense and sparse representation of the signal, which allows the size of the signal to be reduced considerably. Furthermore, DWT is an effective method for extracting relevant features from EEG signals for epilepsy detection. The discrete wavelet coefficients can be used as features to train machine learning models and improve the accuracy of epilepsy detection. By using DWT, epilepsy detection can be performed faster and with increased accuracy, making it a method of choice for the analysis of EEG signals in epilepsy detection.

We decided to decompose the EEG signals using the discrete wavelet transform before classifying them. The motivation behind this decomposition is to compress the time-varying signal, which comprises many data points, into a few parameters that represent the signal.

In the pywt library, several families of wavelets are available, such as Daubechies (db) and Biorthogonal (bior). In addition, when decomposing the signal using each of these wavelets, we need to define the level of decomposition. Daubechies (db) and Biorthogonal (bior) are families of wavelets commonly used in signal processing and image compression applications. Daubechies wavelets are a family of orthogonal wavelets that have a finite support and high regularity. They are well suited for the analysis of signals with smooth variations, such as EEG signals. Biorthogonal wavelets, on the other hand, are not orthogonal but biorthogonal, which means that they have two separate sets of wavelets that form a pair. One set is used for the decomposition of the signal, and the other set is used for the reconstruction. Biorthogonal wavelets are particularly useful for the analysis of signals with sharp changes, such as speech signals.

Both Daubechies and Biorthogonal wavelets have different orders, which correspond to the number of vanishing moments. The higher the order of the wavelet, the more vanishing moments it has, and the better it can capture the smoothness of the signal.



```
def features(mat):
    Kmax = 5
    Tau  = 4
    DE   = 10
    M    = 10
    R    = 0.3
    Band = np.arange(1,86)
    Fs   = 173
    lis = list()
    lis = lis + [np.std(pywt.wavedec(mat,'db4',level=8)[0])]
    lis = lis + [np.std(pywt.wavedec(mat,'db4',level=8)[4])]
    lis = lis + [np.std(pywt.wavedec(mat,'db4',level=8)[5])]
    lis = lis + [np.std(pywt.wavedec(mat,'db4',level=8)[6])]
    sleep(0.01)
    return lis
```

```
MftA = np.zeros((100,feature_size + 1))

for i in range(100):
    MftA[i,:] = features(matA[:,i]) + [0]
```

```
MftD = np.zeros((100,feature_size + 1))

for i in range(100):
    MftD[i,:] = features(matD[:,i]) + [1]
```

```
MftE = np.zeros((100,feature_size + 1))

for i in range(100):
    MftE[i,:] = features(matE[:,i]) + [1]
```

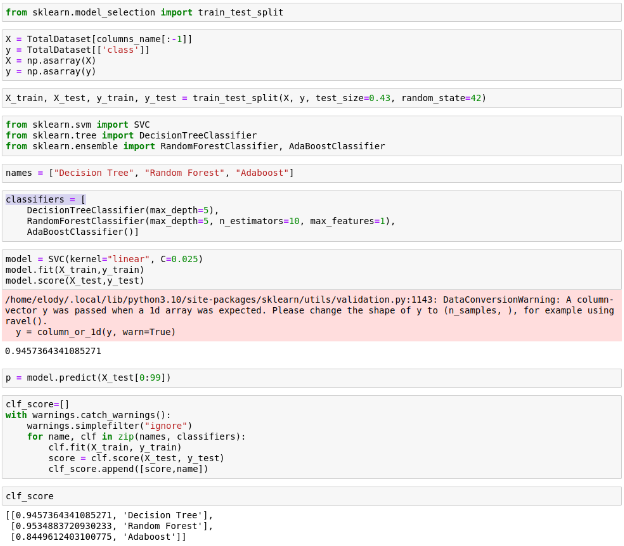
```
FCM_A = pd.DataFrame(MftA,columns=columns_name)
FCM_D = pd.DataFrame(MftD,columns=columns_name)
FCM_E = pd.DataFrame(MftE,columns=columns_name)
FCM_E
```

...

```
TotalDataset = pd.concat([FCM_A,FCM_D,FCM_E],ignore_index=True)
TotalDataset
```

Figure 6: Splitting EEG signals with the Wavelet Transform method

We tested the previous machine learning models used in the first method on wavelet decomposed data. It is interesting to note that the accuracy of the models decreases compared to the undecomposed data. The accuracy of the models after wavelet decomposition are as follows:



```
from sklearn.model_selection import train_test_split
```

```
X = TotalDataset[columns_name[:-1]]
y = TotalDataset[['class']]
X = np.asarray(X)
y = np.asarray(y)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.43, random_state=42)
```

```
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
```

```
names = ["Decision Tree", "Random Forest", "Adaboost"]
```

```
classifiers = [
    DecisionTreeClassifier(max_depth=5),
    RandomForestClassifier(max_depth=5, n_estimators=10, max_features=1),
    AdaBoostClassifier()]
```

```
model = SVC(kernel="linear", C=0.025)
model.fit(X_train,y_train)
model.score(X_test,y_test)
```

```
/home/elody/.local/lib/python3.10/site-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A column-
vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using
ravel().
  y = column_or_1d(y, warn=True)
```

```
0.9457364341085271
```

```
p = model.predict(X_test[0:99])
```

```
clf_score=[]
with warnings.catch_warnings():
    warnings.simplefilter("ignore")
    for name, clf in zip(names, classifiers):
        clf.fit(X_train, y_train)
        score = clf.score(X_test, y_test)
        clf_score.append([score,name])
```

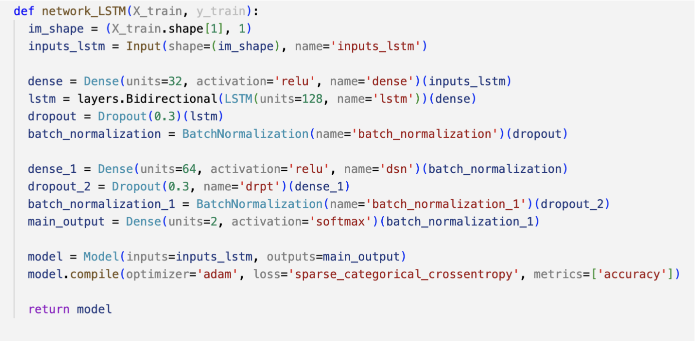
```
clf_score
```

```
[[0.9457364341085271, 'Decision Tree'],
 [0.9534883720930233, 'Random Forest'],
 [0.8449612403100775, 'Adaboost']]
```

Figure 7: Result of the second method

3.3.3. *Third method: Classification with biLSTM*

In this method, we used a bidirectional Long Short-Term Memory (BiLSTM) which is a type of recurrent neural network (RNN) that is commonly used to process sequential data such as text, speech, and time series. BiLSTM is similar to the traditional LSTM, but it processes the input sequence in both forward and backward directions, allowing the network to capture the dependencies and patterns in both directions. This helps the model to have a better understanding of the input sequence and improve its performance on various tasks, such as sentiment analysis, machine translation, and speech recognition. We didn't process the data to feed the network.



```
def network_LSTM(X_train, y_train):
    im_shape = (X_train.shape[1], 1)
    inputs_lstm = Input(shape=(im_shape), name='inputs_lstm')

    dense = Dense(units=32, activation='relu', name='dense')(inputs_lstm)
    lstm = layers.Bidirectional(LSTM(units=128, name='lstm'))(dense)
    dropout = Dropout(0.3)(lstm)
    batch_normalization = BatchNormalization(name='batch_normalization')(dropout)

    dense_1 = Dense(units=64, activation='relu', name='dsn')(batch_normalization)
    dropout_2 = Dropout(0.3, name='drpt')(dense_1)
    batch_normalization_1 = BatchNormalization(name='batch_normalization_1')(dropout_2)
    main_output = Dense(units=2, activation='softmax')(batch_normalization_1)

    model = Model(inputs=inputs_lstm, outputs=main_output)
    model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])

    return model
```

Figure 8: BiLSTM model

The BiLSTM network has several layers. Firstly, an input layer is defined, which takes the input sequences. The dense layer is a fully connected layer with 32 units and a Rectified Linear Unit (ReLU) activation function. The BiLSTM layer has 128 units and processes the input sequences in both forward and backward directions. This is followed by a dropout layer, which randomly drops 30% of the connections between the BiLSTM and the next layer to prevent overfitting. A batch normalization layer then normalizes the activations of the previous layer to speed up the training process.

Next, a second dense layer is defined with 64 units and a ReLU activation function. This is followed by a second dropout layer, which randomly drops 30% of the connections between the dense layer and the next layer. Another batch normalization layer is then added to normalize the activations of the previous layer. Finally, the output layer has two units and a softmax activation function, which is suitable for multi-class classification tasks.

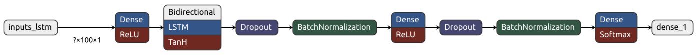


Figure 9: BiLSTM architecture

The model is compiled with the Adam optimizer and the sparse categorical cross-entropy loss function. The performance of the model is evaluated using the accuracy metric. Overall, this BiLSTM network can be used for a variety of tasks, such as text classification, sentiment analysis, and speech recognition, among others.

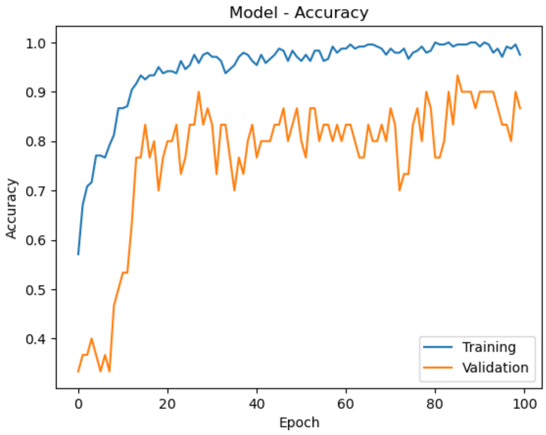


Figure 10: BiLSTM train/validation accuracy

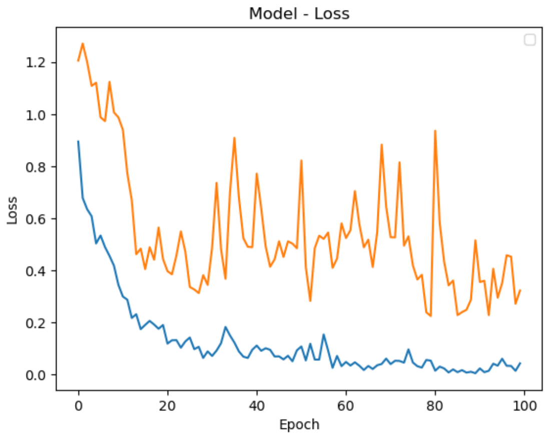


Figure 11: BiLSTM train/validation loss

3.4. Method comparison

In order to make the comparision we saved the models of the from the three methods and create a script for each of them containing the pfetching from the MySQL database, the needed preprocessing steps and the prediction using the saved models. As we have 5 subset of EEG signals, we split the data as follow:

- Set A, D, E were used to train, validate and test the models. A was the healthy patients set D and E were the epilectic cases.

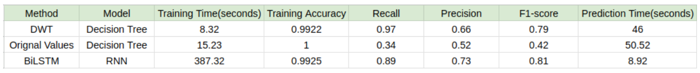
- Set B, C was used to make the analysis and compare the methods. B was the healthy patients set and C was the epilectic cases.

Based on the results of the experiment, we can see that the DWT method using a Decision Tree model had the shortest training time, achieved a high training accuracy, and had a reasonable F1-score. However, its precision was low at 0.66, indicating that it might produce many false positives. The prediction time was also relatively long at 46 seconds.

The original values method using a Decision Tree model had a perfect training accuracy, but its recall, precision, and F1-score were much lower than the DWT method. This indicates that the model is overfitting to the training data and may not generalize well to new data. The prediction time was also relatively long at 50.52 seconds.

The BiLSTM method using an RNN model had the highest training accuracy, recall, precision, and F1-score, indicating that it is likely to perform well on new data. However, it had a much longer training time, but the prediction time compared to the two others methods was way lesser.

Based on these results, the best method to use would depend on the specific requirements of the application. If speed of training is a concern, the DWT method using a Decision Tree model may be the best option, although its low precision may be a disadvantage. If accuracy is the primary concern and longer training is acceptable, the BiLSTM method using an RNN model may be the best option, moreover the prediction time is very fast.



Method	Model	Training Time(seconds)	Training Accuracy	Recall	Precision	F1-score	Prediction Time(seconds)
DWT	Decision Tree	8.32	0.9922	0.97	0.66	0.79	46
Orignal Values	Decision Tree	15.23	1	0.34	0.52	0.42	50.52
BiLSTM	RNN	387.32	0.9925	0.89	0.73	0.81	8.92

Figure 12: Metrics of the different methods

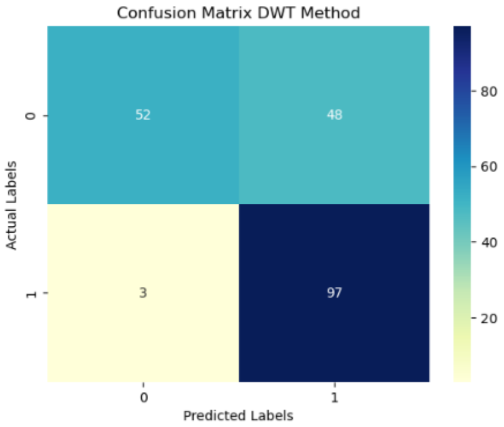


Figure 13: DWT Confusion Matrix

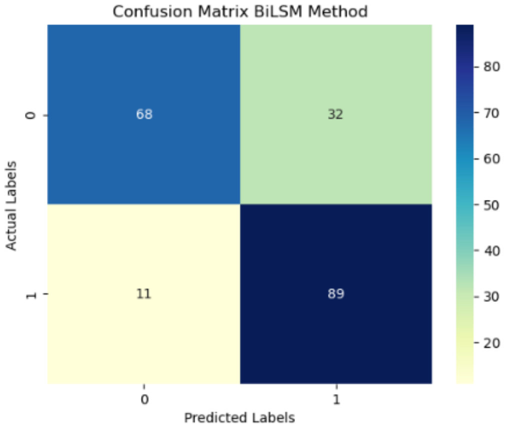


Figure 14: BiLSTM Confusion Matrix

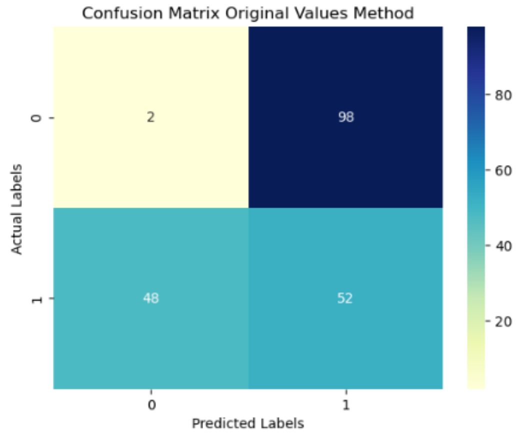


Figure 15: Original Values Confusion Matrix

4. Conclusion

In terms of predicting cases of epilepsy, there are many risk factors, such as family history, brain injury, metabolic diseases, head trauma, which can lead to the development of the disease. Therefore, it is important to stress that prediction of epilepsy should never replace clinical and medical diagnosis by a qualified health professional. Machine learning models can be used as a decision support tool to help health professionals identify individuals at high risk of developing the disease, but they should not be used as the sole method of diagnosis.

References

[1] Nikolaos Giannakeas Markos G. Tsipoura Euripis Glavas Konstantinos Kalafatakis Alexandros T. Tzallas Andreas Miltiadous, Katerina Tzimourta. *Machine Learning Algorithms for Epilepsy Detection Based on Published EEG Databases: A Systematic Review*. ., 2023.

[2] Praphula Kumar Jain Ajith Abraham Lubna Abdelkareim Gabralla Bhaskar Kapoor, Bharti Nagpal. *Epileptic Seizure Prediction Based on Hybrid Seek Optimization Tuned Ensemble Classifier Using EEG Signals*. ., 2022.

[3] Lina Abou-Abbas Imene Jemal, Neila Mezghani and Amar Mitiche. *An Interpretable Deep Learning Classifier for Epileptic Seizure Prediction Using EEG Data*. ., 2022.

[4] Felix Sch Urmann Judit Planas, Fabien Delalondre. *Accelerating Data Analysis in Simulation Neuroscience with Big Data Technologies*. ., 2022.

[5] Bapi Raju Surampudi Kamalaker Dadi. *Big Data in Cognitive Neuroscience: Opportunities and Challenges*. ., 2022.

[6] John Psarras Loukas Ilias, Dimitris Askounis. *Multimodal Detection in Epilepsie*. ., 2022.

[7] University of Boon. Database of university of boon, 2019. https://www.uni-bonn.de/en.

[8] World Health Organization. Epilepsie, 2019. https://www.who.int/fr/news-room/fact-sheets/detail/epilepsy#:~:text=%C3%80%20l\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{%\global\mathchardef\accent@spacefactor\spacefactor}\let\begingroup\let\typeout\protect\begingroup\def\MessageBreak{˙(Font)}\let\protect\immediate\write\m@ne{LaTeXFontInfo:\def{}oninputline49.}\endgroup\endgroup\relax\let\ignorespaces\relax\accent1%\egroup\spacefactor\accent@spacefactorC3%A9chelle%20mondiale%2C%20on,100%20000%20personnes%20par%20an.

[9] Wikiépdia. Irmf's definition, . https://fr.wikipedia.org/wiki/Imagerie_par_r%C3%A9sonance_magn%C3%A9tique_fonctionnelle.