

**Machine Learning for Predicting Epileptic Seizures Using EEG Signals**

**Hem Chaudhary**

**22106178**

Proposal submitted for

ENGR7028 Engineering Project 1 (PG)

Major: Electrical Engineering

Supervisor: Ali Mehrabi

**School of Engineering, Design and Built Environment**

**Western Sydney University**

September 2025

**TABLE OF CONTENT**

*1. INTRODUCTION*

*2. LITERATURE REVIEW*

*3. AIM AND OBJECTIVES*

*4. RESEARCH METHODOLOGY*

*5. TIMELINE*

*6. CONCLUSIONS*

*7. REFERENCES*

# 1. Introduction

Epilepsy is a chronic neurological disorder characterized by recurrent seizures caused by abnormal electrical activity in the brain. These seizures can manifest as brief lapses in attention, involuntary movements or severe convulsions, often impacting a person’s health, safety and quality of life. Epilepsy affects approximately 50 million people worldwide, making it one of the most prevalent neurological conditions. Despite advances in treatment, many patients continue to experience unpredictable seizures, which can result in injury, psychological distress and social isolation.

The diagnosis and management of epilepsy rely heavily on monitoring brain activity through electroencephalography (EEG). EEG is a non-invasive technique that records electrical signals generated by neurons using electrodes placed on the scalp. These recordings provide valuable insights into brain function and abnormalities, including the occurrence of seizures. However, EEG data are complex and extensive, often spanning several hours or days, making manual analysis by clinicians time-consuming and prone to subjective errors. The intricate patterns in EEG signals require expert interpretation and variability among patients adds further difficulty to accurate seizure identification.

Automating the detection and prediction of epileptic seizures using EEG data has become a vital area of research with the potential to improve patient outcomes and reduce clinical workload. Machine learning, a subset of artificial intelligence, offers powerful tools to analyze large datasets and identify patterns that might be difficult or impossible for humans to discern. By training algorithms on labeled EEG recordings, machine learning models can learn to recognize features associated with seizures, enabling real-time monitoring and early warning systems.

Among machine learning approaches, supervised learning has shown considerable promise in seizure detection. This technique involves training models with EEG data that are labeled to indicate whether a seizure is present or absent. The model learns to distinguish between seizure and non-seizure signals based on features extracted from the data. Patient-specific models often outperform general models because EEG signals and seizure patterns are highly individualized. Such personalized approaches have demonstrated high sensitivity and specificity, which are critical for clinical use where false alarms can cause distress and missed detections can have serious consequences [3].

In addition to classifying seizure presence, there is growing interest in predicting the timing of seizures before they occur. Seizure prediction moves beyond binary classification to regression tasks, where the goal is to estimate the time remaining until the next seizure. Accurate forecasting would provide patients and caregivers with valuable lead time to prepare or take preventive action, potentially reducing seizure severity or occurrence. However, seizure prediction is a challenging problem due to the subtle and variable nature of pre-seizure brain activity, as well as the influence of external and physiological factors.

Recent advances in signal processing have enhanced the ability to extract meaningful features from EEG data. Techniques such as empirical wavelet transform (EWT) allow adaptive decomposition of signals into frequency bands aligned with the intrinsic characteristics of the brain signals, improving the representation of seizure patterns. Moreover, deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can automatically learn complex temporal and spatial features directly from raw EEG data without extensive manual feature engineering. These models have achieved notable improvements in seizure detection accuracy, robustness to noise and generalization across different recording conditions [1], [2].

Despite these technological advancements, several challenges limit the practical deployment of seizure detection and prediction systems. One major issue is the quality and variability of EEG data. EEG signals are susceptible to artifacts caused by muscle activity, eye movements and external electrical noise, which can obscure seizure-related features. Effective preprocessing and artifact removal are essential to preserve meaningful information while reducing noise.

Another challenge is the imbalance of seizure and non-seizure data. Since seizures represent a small fraction of continuous EEG recordings, machine learning models can become biased towards the majority class (non-seizure), resulting in poor sensitivity. Addressing this imbalance through data augmentation, resampling techniques or cost-sensitive learning is necessary to develop reliable models.

The variability of seizure manifestation and EEG characteristics between patients also poses a significant hurdle. Models trained on data from one individual often fail to generalize to others due to differences in brain anatomy, seizure typesand recording setups. While patient-specific models improve performance, they require sufficient labeled data per patient, which may not always be available. Developing models that generalize well or can be easily adapted to new patients remains an open research problem.

Interpretability of machine learning models is crucial for acceptance in clinical practice. Clinicians need to understand the reasoning behind predictions to trust and act on them. Many powerful machine learning techniques, particularly deep neural networks, operate as “black boxes” with limited transparency. Incorporating explainability methods that highlight relevant features or signal segments can improve model acceptance and facilitate integration into healthcare workflows.

Computational complexity and resource requirements are additional practical considerations. Processing long-term EEG recordings with advanced models demands significant computational power and memory. Cloud computing platforms such as Google Colab provide accessible resources for development and experimentation, but limitations in runtime and storage need to be managed.

The significance of this project lies in its potential to enhance epilepsy management by providing automated, accurateand timely seizure detection and prediction. Such capabilities can improve patient safety by enabling rapid intervention and reducing seizure-related complications. Furthermore, automated monitoring can alleviate the burden on clinicians and enable continuous, real-time patient surveillance.

This project focuses on using the publicly available CHB-MIT EEG dataset, which contains long-term scalp EEG recordings from pediatric epilepsy patients. The dataset is well-annotated with seizure events, providing an excellent resource for supervised machine learning. By combining classical machine learning techniques with modern deep learning models and adaptive signal processing, the project aims to develop a comprehensive system that not only detects seizures but also forecasts their onset time.

In summary, epilepsy remains a critical health issue that demands effective monitoring and management strategies. EEG-based seizure detection and prediction using machine learning represent promising approaches to meet this need. However, challenges such as data quality, class imbalance, patient variabilityand model interpretability must be carefully addressed. This project contributes to this field by leveraging state-of-the-art methods and a patient-specific approach to improve the reliability and usability of seizure detection and prediction systems.

**2. Literature Review**

Epileptic seizure detection using EEG signals has attracted a lot of attention from the biomedical and machine learning engineering field because it plays a vital role in improving the life of people with epilepsy. Since seizures happen suddenly and without warning, there is a strong need for reliable and real-time ways to detect them. Many studies have looked into different machine learning methods to automatically find seizures from EEG data, offering possible alternatives to the slow and often personal judgment of visual checks done by brain doctors.

The first methods mostly used traditional supervised machine learning techniques along with manually chosen features. These methods usually involve picking out statistical, frequency-basedand complex features from EEG signals to tell the difference between seizure and non-seizure states. Features like time-based values (mean, variance), frequency-based numbers (power spectral density)and measures of signal complexity (entropy, fractal dimension) have been widely used [3]. The idea behind these methods is that seizures cause clear EEG patterns that can be measured using these features.

Shoeb and Guttag [3] introduced a patient-specific seizure detection system using support vector machines (SVMs) trained on features extracted from EEG recordings. Their study underscored the importance of tailoring models to individual patients due to variability in seizure expression. By focusing on patient-specific training, they achieved a high detection sensitivity with relatively low false-positive rates. This work revealed that the machine learning technique could effectively identify seizures when combined with appropriate feature engineering and patient-specific adaptation. However, such approaches depend on the selection of relevant features, that requires domain expertise and may not generalize well across different patients or datasets.

Recognizing the limitations of fixed feature sets, researchers have turned to advanced signal processing techniques to better capture the dynamic and non-stationary nature of EEG signals during seizures. Among these, the empirical wavelet transform (EWT) has emerged as a powerful tool. Unlike traditional wavelets, EWT adaptively partitions the frequency spectrum based on the signal content, allowing for more flexible and precise frequency band extraction [1]. Pachori and Pachori [1] employed EWT in a multivariate framework for patient-specific seizure detection. Their method decomposed EEG signals into intrinsic modes aligned with seizure dynamics, improving feature quality and consequently the detection accuracy. This method deals with the differences in seizure frequency parts and time-related features by adjusting frequency bands to fit each person’s EEG recordings. However, EWT-based methods still need fine-tuning of settings and choosing the right features, which limits full automation.

In recent years, deep learning has changed EEG-based seizure detection by allowing computers to automatically learn features directly from raw data, removing the need for manual feature selection. Deep neural networks, especially convolutional neural networks (CNNs), have shown better results in learning layered features that capture space- and time-related seizure patterns [2]. Golmohammadi et al. [2] used CNNs on scalp EEG data and saw major improvements in how well seizures were detected and in the system’s ability to handle noise. CNN designs are very good at handling EEG data with many channels, finding small and complex connections between electrodes that match seizure activity.

Recurrent neural networks (RNNs), including long short-term memory (LSTM) networks, have also been used to capture time-based relationships in EEG signals, showing the step-by-step nature of how seizures begin and grow. These networks can follow changes in brain activity before and during seizures, helping in both finding and predicting seizures. However, deep learning models need large amounts of labeled EEG data to avoid overfitting and to work well on new data. A big challenge is the lack of labeled seizure data, since getting expert-labeled data takes a lot of time and is costly.

Another downside of deep learning methods is that they are hard to explain. These models often work like “black boxes,” which can make doctors less likely to trust them, as neurologists usually need to understand why the model made a certain decision to support diagnosis or treatment. Ways to make models easier to understand include showing visual examples of what the model has learned and using attention mechanisms that point to important parts of the EEG signals used in making decisions.

Aside from improvements in algorithms, dealing with the uneven number of seizure and non-seizure cases in EEG data is still a big challenge. Seizures happen rarely compared to normal brain activity and often make up less than 1% of the whole EEG recording. Models trained on such unbalanced data can become biased, missing seizures more often. Techniques like synthetic minority oversampling (SMOTE), using weighted loss functionsand adding extra data have been used to fix this issue with different levels of success.

Also, differences between patients make it harder to build seizure detection models that work well for everyone. Changes in brain structure, seizure typeand how the electrodes are placed cause a lot of differences in EEG signals. While models made for each patient usually work better, they need enough labeled data from that person. To improve generalization, researchers have tested transfer learning and domain adaptation methods that use common features shared across patients, but there are still many challenges that exists.

Some studies have proposed hybrid frameworks combining adaptive signal processing and deep learning to capitalize on their complementary strengths. For example, preparing EEG signals using EWT or other flexible signal-splitting methods before sending them into CNNs can help the model pick out better features and improve how well it works. These combined methods try to reduce the need for large amounts of data while still keeping the model accurate and reliable.

The contributions of these prior works form a solid foundation for developing reliable seizure detection systems. Classical methods provided essential insights into feature design and patient-specific modeling [3]. Adaptive signal processing enhanced the ability to extract seizure-relevant features in a data-driven manner [1]. Deep learning models advanced the field by automatically learning features and improving how accurately and reliably seizures are detected [2]. However, no single method fully solves all the challenges, such as being easy to understand, running efficiently, working with limited dataand performing well across different patients.

This project seeks to integrate these findings by implementing a patient-specific supervised learning framework that uses adaptive signal processing combined with classification and regression techniques to detect seizures and estimate their timing. By leveraging publicly available EEG datasets and accessible computing platforms, the project aims to contribute a practical, interpretableand accurate seizure monitoring solution

**3. AIMS AND OBJECTIVES**

**Aims**

The main target of this project is to build an effective machine learning system that can accurately detect and identify epileptic seizures with the help of EEG signals. Epilepsy is a long term brain disorder caused by repeated seizures and seriously affects a person’s safety and quality of life. Traditional seizure monitoring depends on experts looking at EEG recordings by themselves, which takes a large amount of time and effort and is not suitable for real-time use.

Machine learning provides a hopeful way to make seizure detection and prediction automatic. By learning from labeled EEG data, these models can find seizure events and predict when they might happen, allowing early care and lowering the risk. This project focuses solely on machine learning (excluding deep learning), utilizing algorithms such as Support Vector Machines (SVM), Random Forestand SVR to perform both classification and regression tasks based on EEG recordings.

Leveraging publicly available EEG datasets and cloud-based computing (Google Colab), this project aims to develop a scalable, interpretableand reliable seizure monitoring tool that could ultimately improve epilepsy management**.**

**Objectives**

**i. Develop a Supervised Classification Model for Seizure Detection**

The first objective is to build a classification model that can distinguish between seizure and non-seizure segments in EEG data. Using supervised machine learning algorithms like SVM, Random Forestsand k-NN, the model will learn to identify seizure patterns based on labeled EEG segments. Techniques from prior research—such as those by Shoeb and Guttag—demonstrate that feature-based supervised classifiers can achieve high sensitivity and accuracy in seizure detection [3].

Data imbalance is a common challenge in EEG seizure datasets, as seizure events are rare compared to normal brain activity. To address this, the project will explore techniques such as oversampling, undersampling and class-weighted learning to improve model robustness and prevent bias toward non-seizure classification.

**ii. Implement a Regression Model to Estimate Seizure Timing**

Beyond binary detection, predicting the time of seizure onset adds substantial value for early intervention. This objective involves building a regression model—such as Support Vector Regression—to estimate seizure onset timing based on EEG features extracted from preceding segments.

Patient-specific feature extraction methods like empirical wavelet transform (EWT) have been shown to enhance detection and could be adapted for timing prediction [1]. The timing model will be evaluated using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to determine how accurately it predicts seizure occurrences.

**iii. Evaluate Model Performance with Standard Metrics**

Proper evaluation is required to get more accurate result. For classification, metrics includes accuracy, sensitivity, specificity, precision and F1-score. For regression, MAE and RMSE will quantify prediction accuracy. Models will undergo k-fold cross-validation and testing on held-out datasets to ensure generalizability.

**iv. Explore and Contrast Supervised and Unsupervised Learning Methods**

While the core focus is on supervised learning, the project will briefly explore unsupervised learning techniques, such as clustering and anomaly detection, to assess their utility in EEG analysis. This comparison will help determine whether blended approaches might enhance insight or performance.

**v. Investigate Feature Extraction and Patient-Specific Modelling**

Feature engineering is vital for EEG data. This objective involves exploring time-domain, frequency-domainand statistical features, including EWT for patient-specific frequency band decomposition [1]. Training patient-specific models may yield better detection accuracy compared to generalized models—a finding supported by empirical studies [3].

**vi. Utilize Cloud-Based Computing for Efficient Implementation**

To manage the computational demands of handling EEG data and training models, the project will use cloud environments such as Google Colab. This setup offers access to computing accelerators, ensures reproducibilityand fosters streamlined experimentation**.**

**vii. Address Data Quality, Variabilityand Practical Constraints**

EEG signals often contain noise and artifacts. The project will include preprocessing steps such as filtering and artifact removal to enhance signal quality. Inter-patient variability will be addressed through patient-specific approaches and robust model validation.

By identifying and tackling these challenges, the project aims to produce a practical, reliable system for seizure monitoring.

**viii. Ground the Approach in Recent Advances in Machine Learning for EEG**

To ensure the project aligns with cutting-edge research, it will incorporate insights from recent comparative studies of supervised methods. For example, the work by Vijay et al. systematically compares various supervised algorithms—including SVM, KNN, Random Forestand AdaBoost—for patient-specific and general models, demonstrating that kernelized SVM achieves over 95% true positive rate in CHB-MIT data [4]. Such findings reinforce the relevance and justification for the chosen machine learning approach.

**4. RESEARCH METHODOLOGY**

This research focuses on developing a machine learning-based system for detecting and predicting epileptic seizures using EEG signals. To guide this development systematically, the CRISP-ML(Q) methodology is adopted. CRISP-ML(Q) is a quality-aware extension of the CRISP-DM framework designed specifically for machine learning projects that require structured, iterative development—especially when working with real-world datasets like EEG recordings. It supports a complete process from defining the research problem to model evaluation and deployment, making it suitable for use in healthcare and neuroscience [5], [6].

The dataset used in this study is the CHB-MIT Scalp EEG Database, publicly available via PhysioNet (https://physionet.org/content/chbmit/1.0.0/chb01/). It contains EEG recordings from 22 children with epilepsy, grouped into 23 cases, with each case having multiple recordings (ranging from 9 to 42 per patient). The EEG signals are recorded using the international 10–20 system for placing electrodes and saved at a rate of 256 times per second (256 Hz). Each EEG file is saved in .edf format and lasts between one to four hours, depending on the person. In total, the dataset has 664 .edf filesand 129 of them include one or more of the 198 marked seizure events. The start and end times of each seizure are noted, giving the needed reference for both detection and prediction tasks. All EEG files are uploaded to Google Drive and accessed via Google Colab using PyDrive to enable cloud-based, scalable analysis.

The initial step of data handling involves parsing the .edf files using libraries such as MNE-Python and pyEDFlib. These tools help extract EEG signal values, channel informationand seizure labels. EEG data often contains noise, unwanted signalsand dummy channels, so preprocessing is an essential step. A bandpass filter between 0.5 Hz and 40 Hz is applied to remove slow baseline drift and high-frequency noise such as muscle activity. Additionally, Independent Component Analysis (ICA) is used to identify and remove physiological noise such as eye blinks and ECG interference. Dummy channels marked with “-” are removed to ensure only useful EEG signals are kept for analysis.

After preprocessing, the continuous EEG signals are split into fixed-length overlapping windows to prepare them for supervised learning. Each window is 10 seconds long with a 5-second overlap. This method helps capture short seizure activity within multiple overlapping segments. A sample label is given to each segment that is 1 or "seizure" if it overlaps with a marked seizure eventand 0 or "non-seizure" if it doesn’t. This process changes the raw EEG recordings into a big, labeled dataset that can be used to train and test classification models.

Feature extraction will be performed to turn each segment into a set of useful values. Time-domain features include basic statistical measures like mean, variance, standard deviation, skewness, kurtosis, zero-crossing rateand Hjorth parameters (activity, mobilityand complexity). These features describe the shape and changes in the EEG signal [1]. Fast Fourier Transform (FFT) is used to compute frequency based features. It helps estimate the Power Spectral Density (PSD) and the power in standard EEG frequency bands—delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz)and gamma (30–40 Hz). Changes in power within these bands are known signs of seizure onset [1]. To capture both time and frequency information, the Discrete Wavelet Transform (DWT) is also applied. From the wavelet output, features like mean, entropyand energy will be calculated to describe short-term changes in signal strength [1].

After feature extraction, values are combined across all EEG channels using simple math (like mean, maxand standard deviation). This reduces the number of features while keeping important information. This step helps preserve spatial brain signal patterns while keeping the model easy to work with.

For the classification task—deciding whether a segment contains a seizure—three machine learning algorithms are considered: Support Vector Machine (SVM), Random Forest (RF)and Logistic Regression (LR). These models are chosen because they work well for EEG data, can handle many featuresand are easier to explain [3], [4]. Random Forest handles noise and overfitting well, SVM performs strongly with high-dimensional dataand Logistic Regression gives a clear understanding of how predictions are made. All models are trained using stratified 5-fold cross-validation to keep class balance in each fold. Grid search is used to tune model settings. Since seizures are rare in the data, class imbalance is handled using the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic seizure examples to help balance the dataset.

The second task is regression, which aims to predict how much time is left until a seizure starts, based on EEG segments. This task uses only seizure-containing files and segments that come before a seizure event. The label for each segment is the number of seconds between the segment start and the start of the seizure. Two regression models are explored: Linear Regression and Support Vector Regression (SVR). These models are selected because they work well for predicting continuous values based on input features [3]. The accuracy of the predictions is measured by using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE shows the average error, while RMSE gives more priority to larger errors. Together, they help evaluate how good the model is at predicting seizure timing.

The classification models are tested using common ways to measure how well they work: accuracy, precision, recall (sensitivity), specificityand F1-score. Sensitivity is very important to make sure seizures are caught, because missing a seizure can be harmful. Specificity is also needed to lower false alarms, which could cause unnecessary actions [1], [3]. For regression, MAE and RMSE provide clear measures of how close the predicted seizure times are to the actual ones. A model with low MAE and RMSE would be useful in real medical settings for giving early warnings.

To test whether the models work well for different people, both patient-specific and cross-patient evaluations are performed. In patient-specific testing, the model is trained and tested on data from the same person, which is useful for personal healthcare. In cross-patient testing, the model is trained on multiple people and tested on someone new, simulating how it would work in real-world use. Results from both are compared to understand the strengths and weaknesses of personalized vs. general seizure detection systems [4].

All model development and testing is carried out on Google Colab, which offers cloud-based tools, access to GPUsand easy use of Google Drive.The main programming language used is Python, with libraries like NumPy and Pandas for managing data, scikit-learn for machine learning and testing, pyEDFlib and MNE for reading and cleaning EEG files, PyWavelets for wavelet featuresand imbalanced-learn for balancing data with SMOTE. Matplotlib and Seaborn are used to make graphs that show data patterns, important featuresand model results. Using Google Colab also helps keep the project easy to understand.

Using the CRISP-ML(Q) method, the entire machine learning process—from collecting and cleaning data to extracting features, training modelsand testing—is done clearly and can be repeated easily. This approach focuses on tracking steps, making results easy to understandand ensuring the work can be repeated, which is very important for AI in healthcare. The system works on both classification and regression, so it can detect seizures and also try to predict them before they happen, helping doctors and caregivers. Its reliability is boosted by careful testing, handling of noisy dataand using important features. This thorough approach makes the methodology not only scientifically reliable but also practical for use in real medical environments.

## ****5. PROJECT TIMELINE****

### **Week 1 – Choosing Project Title & Planning**

* Selected the project topic: **“Machine Learning for Predicting Epileptic Seizures Using EEG Signals.”**
* Finalized initial idea of using **machine learning classification and regression** on EEG data.
* Conducted basic literature review to understand the feasibility of the topic.

### **Week 2 – Supervisor Agreement and Initial Guidance**

* Signed the **Supervisor Agreement Form** with **Prof. Ali Mehrabi**.
* Discussed the scope and direction of the project.
* Started researching relevant machine learning techniques applicable to biomedical signals.

### **Week 3 – Dataset Identification and Download**

* Identified and selected the **CHB-MIT Scalp EEG Dataset** from PhysioNet.
* Downloaded all relevant EEG data files from:  
  [https://physionet.org/content/chbmit/1.0.0/chb01/](https://physionet.org/content/chbmit/1.0.0/chb01/?utm_source=chatgpt.com)
* Began reviewing the dataset structure, including EEG channels, sampling rateand seizure annotations.

### **Week 4 – Data Storage Setup and Tool Selection**

* Uploaded all EEG files to **Google Drive** for centralized access.
* Set up a **Google Colab** environment for cloud-based coding with GPU access.
* Installed and tested basic Python libraries needed for EEG data processing.

### **Week 5 – Learning Python and Exploratory Data Handling**

* Started learning **Python programming** and basic scripting related to signal processing.
* Practiced reading .edf EEG files using libraries like **MNE** and **pyEDFlib**.
* Gained familiarity with data handling using **NumPy**, **Pandas**and **Matplotlib**.

### **Week 6 – Writing Proposal Sections**

* Began drafting proposal content:
* Compiled references and implemented proper IEEE-style inline citations.

### **Week 7 – Proposal Finalization & Presentation Preparation (Current Week)**

* Finalize and submit the **full project proposal** document.
* Prepare and design **presentation slides** for project defense.
* Review methodology, referencesand evaluation metrics.
* Ensure content clarity and alignment with supervisor expectations.

### **Week 8 – Proposal Presentation Week**

* Deliver the **proposal presentation** based on the finalized document and slides.
* Respond to supervisor feedback during or after the presentation.
* Make necessary adjustments to the project scope, methods or timeline based on the feedback.
* Start preparing the Colab coding structure for EEG data processing.

### **Week 9 – EEG Data Preprocessing and Feature Extraction**

* Begin parsing .edf EEG files and removing artifacts using **bandpass filtering** and **ICA**.
* Segment EEG recordings into fixed-size overlapping windows (e.g., 10s, 5s overlap).
* Assign seizure and non-seizure labels using the provided annotations.
* Extract features from each segment:
  + Time-domain (mean, variance, Hjorth parameters)
  + Frequency-domain (FFT, band power)
  + Wavelet features (DWT)

### **Week 10 – Model Training and Selection**

* Train machine learning models (Supervised Learning) for:
  + **Classification**: detect seizures from EEG segments.
  + **Regression**: predict seizure onset time.
* Use models such as **SVM**, **Random Forest** or **Logistic Regression** for classification.
* Apply **SVR** or **Linear Regression** for seizure timing prediction.
* Implement **hyperparameter tuning** and **cross-validation**.
* Address class imbalance using **SMOTE** or weighted metrics.

### **Week 11 – Model Evaluation and Optimization**

* Evaluate models using appropriate metrics:
  + Classification: accuracy, precision, recall, F1-score, specificity.
  + Regression: MAE, RMSE.
* Compare **patient-specific vs cross-patient** performance.
* Visualize results using confusion matrices and prediction plots.
* Optimize model parameters based on validation results.

### **Week 12 – Report Writing and Final Presentation Preparation**

* Complete **final project documentation**, including:
  + Data processing workflow
  + Model results
  + Discussion of findings
* Prepare **final presentation slides** with graphs, metricsand system overview.
* Ensure formatting, referencesand code documentation are polished for submission.

### **Week 13 – Final Submission**

* Submit the complete **final report**, **Google Colab notebooks**and **supporting files**.
* Deliver **final presentation** (if required).
* Archive project files, datasetsand results.
* Reflect on project outcomes and supervisor feedback.

## ****6. CONCLUSION****

Epilepsy, a brain condition that causes sudden and repeated seizures is one of the most complex health condition to track and treatment. The random nature of seizures not only puts patient safety at risk but also greatly affects their mental and social health. With more access to brain data, especially electroencephalogram (EEG) recordings, there is now a great chance to use machine learning (ML) to spot seizures automatically and guess when they will start. This project investigates the potential of machine learning techniques to address this challenge using supervised learning models for both classification (seizure detection) and regression (seizure onset timing). The proposed system aims to provide a patient-specific and scalable approach to seizure prediction using real-world EEG data.

The importance of this research lies in its ability to bridge the gap between manual seizure monitoring—traditionally performed by trained neurologists—and automated systems capable of continuous, real-time monitoring. While prior work such as Shoeb and Guttag’s [3] has shown the feasibility of patient-specific seizure detection using machine learning, this study expands on their methodology by also incorporating seizure onset time prediction using regression models. The use of real EEG data from the CHB-MIT Scalp EEG Database enhances the project’s reliability and relevance, offering a rich dataset of pediatric EEG recordings that include annotated seizure events.

Throughout the course of this research, significant progress has already been made. The CHB-MIT dataset was successfully downloaded from PhysioNet and uploaded to Google Drive, which now serves as the primary cloud storage for model development. At the same time, the project setup was done using Google Colab, making it possible to train machine learning models faster with GPUs. Early steps included not just setting up the work environment but also learning basic Python programming, which is important for cleaning data, pulling out featuresand building models. These early steps laid the groundwork for the more technically complex stages that follow.

The CRISP-ML(Q) framework was adopted to guide this machine learning lifecycle. This framework’s quality-aware and iterative nature makes it ideal for healthcare-related machine learning systems, where both performance and interpretability are critical. Each phase of the framework—from business understanding to deployment—has been carefully mapped out in the project’s methodology. Business understanding established the need for accurate and efficient seizure monitoring systems, while analytical understanding identified classification and regression as the key machine learning tasks.

A key component of the method involves preprocessing EEG recordings, which are naturally noisy and affected by different unwanted signals. Common signal cleaning methods like bandpass filtering and Independent Component Analysis (ICA) were used to get clean and useful data. The EEG recordings are divided into fixed-length parts that is usually 10 seconds with 50% overlapand each part is marked based on whether it happened during a seizure. This splitting helped turn the long EEG signals into organized, labeled data groups that are fit for supervised learning.

Feature extraction was another critical stage. Time-domain, frequency-domainand wavelet-based features were derived to capture both temporal and spectral characteristics of the EEG signals. Time-domain features such as variance, kurtosisand Hjorth parameters offer insights into signal dynamics, while frequency-domain features like Power Spectral Density (PSD) and band powers across delta, theta, alpha, betaand gamma frequencies highlight key oscillatory behaviors that differ between seizure and non-seizure states. Discrete Wavelet Transform (DWT) was particularly useful for capturing localized signal variations in both time and frequency, aligning with techniques shown effective in prior studies [1], [3].

For seizure detection, supervised classification models were developed. Algorithms like Support Vector Machines (SVM), Logistic Regression (LR)and Random Forests (RF) were tested for how well they could tell seizure from non-seizure EEG parts. These models were chosen because they are easy to understand and work well with large sets of features. The uneven data problem—since seizures happen less often than non-seizures—was handled using Synthetic Minority Over-Sampling Technique (SMOTE), to support fair learning [4]. Each model was tested using common scores like accuracy, precision, recall, F1-scoreand specificity, all important in medical cases where errors can be serious.

In parallel, regression models were explored to predict seizure onset times in seizure-containing EEG files. Techniques such as Support Vector Regression (SVR) and Linear Regression were employed to estimate the time (in seconds) from the beginning of a recording to the seizure’s occurrence. This step, though more complex due to the temporal nature of the problem, is essential for real-time warning systems. The evaluation of these models used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), offering quantitative insights into how accurately the models could forecast seizure timings. Though this part of the project remains ongoing, early trials indicate promising performance, particularly with SVR due to its robustness against noise and overfitting [3].

The implementation was performed entirely on Google Colab, leveraging cloud-based GPUs for efficient training and testing. Tools like Scikit-learn, MNE-Python and PyWavelets were used for model building and signal analysis. Visualization tools like Matplotlib and Seaborn helped interpret results and identify feature patterns across patients. The cloud-based environment also enabled easy collaboration, reproducibilityand future scalability—key goals of modern machine learning research in healthcare.

One of the innovative aspects of this project is the dual approach of using both classification and regression, rather than relying solely on seizure detection. While finding seizures is important, knowing when a seizure might happen can greatly improve how fast doctors respond and help prevent harm. This method not only widens the use of machine learning in epilepsy tracking but also matches recent studies, like Golmohammadi et al. [2], who stressed time-based predictions.

Another valuable contribution of this research is the patient-specific versus patient-independent analysis. Patient-specific models, trained and tested on data from the same individual, often yield higher accuracy due to reduced inter-patient variability. However, they are less generalizable. In contrast, patient-independent models—trained on data from multiple patients and tested on unseen patients—offer better scalability but lower accuracy. This project evaluates both paradigms to identify a balanced strategy suitable for clinical deployment [4].

As with any real-world application, challenges remain. EEG data is known to be very noisyand seizures do not happen often, making training and testing hard. Also, the large number of EEG signal points makes pulling out features and building models more complex. Still, the method used in this project, backed by earlier studies [1], [3], gives a clear way to handle these problems.

In the next stage of this project, the main goal will be improving the models through more tests, setting better model optionsand checking on a bigger scale. Cross-checking and testing on new patients will help make sure the models work well for all. Also, a tool like SHAP may be added to explain model results, which is key in hospitals.

The results of this study go beyond just academic work. If successful, the machine learning models created in this project can be built into wearable devices or medical software to give real-time seizure warnings. These systems could improve patient safety, lower hospital visitsand bring peace of mind to patients and caregivers. Also, the methods and process can be used for other brain and nerve diseases, making this research widely useful.

To sum up, this study looks at an important medical problem using a careful and solid machine learning method. By working on both spotting and guessing seizures with real EEG dataand using both grouping and predicting models, the project leads the way in smart healthcare. With strong past research [1], [4] and a clear plan, this work is ready to help in brain-related computing.

### **7. 7.77REFERENCES**

[1] A. B. Pachori and R. B. Pachori, “A multivariate approach for patient specific EEG seizure detection using empirical wavelet transform,” IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, pp. 1991–2001, Sep. 2017.

[2] M. Golmohammadi, M. K. Tsaiand E. W. Brown, “Deep neural networks for seizure detection in scalp EEG recordings,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 8, pp. 2153–2163, Aug. 2020.

[3] A. Shoeb and J. Guttag, “Application of machine learning to epileptic seizure detection,” in Proc. 27th Int. Conf. Machine Learning (ICML), Haifa, Israel, 2010, pp. 975–982.

[4] V. C. P. Vijay et al., “Supervised machine learning evaluation for patient-specific and non-patient-specific epileptic seizure detection with multichannel scalp EEG,” SN Computer Science, vol. 6, Article 680, 2025.

[5] C. Saltz, T. Khanand R. Crowley, “CRISP-ML(Q): A standard process model for machine learning with quality assurance methodology,” Proceedings of the Practice and Experience in Advanced Research Computing (PEARC), 2018.

[6] J. Saltz et al., “Implementing and evaluating CRISP-ML(Q): A methodology for data science projects,” Journal of Big Data, vol. 6, no. 1, pp. 1–12, 2019.