Tampere University

SeyedHasan MirHosseini

SEIZURE DETECTION FROM EEG SIGNALS USING A MULTIMODAL DEEP LEARNING MODEL

Master of Science Thesis

Faculty of Information Technology and Communication Sciences

Examiners: Prof. Jaakko Peltonen, Saana Seppälä,

July 2025

ABSTRACT

SeyedHasan MirHosseini: Seizure Detection from EEG Signals Using a Multimodal Deep Learning Model
Master of Science Thesis
Tampere University
Statistical Data Analytics
July 2025

Seizure is a neurobiological disorder affecting millions of individuals across the globe, making accurate detection from EEG signals crucial for diagnosis and improving patient outcomes. Compared to traditional machine learning approaches typically requiring manual feature extraction, deep learning has advanced seizure detection by automating feature extraction. However, most studies rely on unimodal EEG data, failing to capture the full complexity of brain dynamics.

To address this limitation, this thesis proposes a multimodal deep learning model that integrates both raw EEG signals, processed using LSTM networks, and spectrograms analyzed by Convolutional Neural Networks CNNs. The outputs from these two modalities are then fused and processed through a Transformer architecture to model cross-channel dependencies and classify signals. To evaluate the effectiveness of the proposed model, two unimodal models: an LSTM-Transformer model using raw EEG signals and a CNN-Transformer model using spectrograms, were also developed as baselines.

The results showed that while the unimodal CNN-Transformer model achieved the highest overall accuracy with 90%, highlighting the effectiveness of spectral features, the proposed model exhibited superior class-specific performance, the highest recall for LPD (93%) and the precision for seizure (91%). The findings indicate that combining temporal and spectral modalities can enhance sensitivity and specificity in seizure detection. Future research could explore alternative fusion strategies along with more diverse classes to improve the model's generalizability.

Keywords: Seizure, LPD, Multimodal deep learning, LSTM, CNN, Transformer.

The originality of this thesis has been checked using the Turnitin Originality Check service.

USE OF AI IN THIS THESIS

I have utilized AI tools in my thesis:

- No
- ✓ Yes

The AI tools utilized in my thesis and their purposes are described below:

- ChatGPT (versions 3.5 and 4o) by OpenAI
- Claude Haiku (version 3.5) by Anthropic

These tools provided support in the following ways:

- Improving writing quality: ChatGPT helped rephrase sentences, enhance cohesion and coherence, and correct grammatical errors.
- Technical assistance: Claude Haiku aided in debugging Python code and offered programming-related suggestions during model development.

Both tools were used throughout the research process. While ChatGPT was mainly used for refining the thesis text, Claude assisted with coding issues and corrections during the development of the model.

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications

PREFACE

This thesis is submitted in fulfillment of the requirements for Master's degree in Statistical Data Analytics at Tampere University. The interdisciplinary nature of this work reflects both my academic interests and my aspiration to contribute to meaningful advancements in the field of computational neuroscience.

The process of conducting this thesis has been both challenging and transformative. It not only significantly enhanced my understanding of neural signals and advanced deep learning architectures but also ignited a passion for pursuing research in this field with a broader scope and more sophisticated models.

I would like to express my sincere gratitude to my supervisors, Professor Jaakko Peltonen and Saana Seppälä, for their support, guidance, encouragement, and insightful feedback throughout this process. I also extend my appreciation to the faculty members and colleagues at the Department of Computing Sciences and Electrical Engineering for creating a supportive research environment.

Beyond its academic contributions, this work marks a significant personal milestone. It is the culmination of years of learning, resilience, and unwavering dedication, and it would not have been possible without the incredible support of my family. Their belief in me has been a constant source of strength and inspiration.

Ultimately, I hope that the contributions of this thesis, however modest, will support ongoing efforts to improve neurological care and inspire further exploration at the intersection of AI and healthcare.

SeyedHasan MirHosseini

Tampere, July 2025

Contents

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

%	Percentage
→	Indicates model flow or sequence (e.g., LSTM → Transformer)
AUC	Area Under the Curve
CHB-MIT	Children’s Hospital Boston - Massachusetts Institute of Technology EEG dataset
CNN	Convolutional Neural Network
DL	Deep Learning
EEG	Electroencephalography
f1-score	Harmonic mean of precision and recall
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GPD	Generalized Periodic Discharges
GRDA	Generalized Rhythmic Delta Activity
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
KL Divergence	Kullback–Leibler Divergence
LPD	Lateralized Periodic Discharges
LRDA	Lateralized Rhythmic Delta Activity
LSTM	Long Short-Term Memory
ML	Machine Learning
PCA	Principal Component Analysis
ReLU	Rectified Linear Unit
SHAP	SHapley Additive exPlanations
STFT	Short-Time Fourier Transform
TCN	Temporal Convolutional Network
TN	True Negative
TP	True Positive
ViT	Vision Transformer
ADC	Analog-to-Digital Converter

1. Introduction

According to the World Health Organization (2024), Millions of people worldwide are suffering from epilepsy, a neurological condition resulted from spontaneous and recurring seizures. Seizures arise when a group of neurons undergo excessive, synchronized electrical activity, potentially affecting one or multiple brain regions (Dutta et al, 2024). These discharges can lead to a wide spectrum of symptoms, ranging from subtle lapses in attention or brief muscle spasms to severe, prolonged convulsions (Fisher et al., 2014; Sirven & Shafer, 2020). To record, diagnose, and monitor such brain activities, electroencephalography (EEG) is widely used due to its ability to capture real-time electrical signals from the brain (Halford et al., 2018). However, manual interpretation of EEG recordings is inherently challenging. Neurologists must identify subtle and often transient patterns that signal abnormal activity, which introduces a high degree of subjectivity and is susceptible to human error. Moreover, inter-rater variability among clinicians can lead to inconsistencies in diagnosis and delays in treatment initiation (Acharya et al., 2013; Tatum et al., 2011).

To address these limitations, the integration of computational tools particularly machine learning has emerged as a promising approach. Over the past decade, significant efforts have been made to develop automatic seizure detection systems using conventional machine learning algorithms, including decision trees, support vector machines (SVMs), and ensemble models (Benjamens et al., 2020; Shoeb, 2009; Han, Liu, & Friedman, 2024). While these methods have shown potential, their reliance on manually crafted features and prior domain knowledge imposes several constraints. The variability in EEG features across individuals and over time, coupled with the complexity of feature engineering, often results in limited generalizability (Dutta et al., 2024; Sanei & Chambers, 2013). Furthermore, EEG signals pose significant challenges for analysis due to their nonlinear, highly dynamic, and high-dimensional characteristics. The nonlinear nature stems from the intricate interactions among neuronal circuits, which often render traditional statistical models insufficient for accurate representation (Stam, 2005). Another analytical challenge arises from the combination of EEG's high temporal resolution and its multichannel nature. While the millisecond-scale sampling captures rapid fluctuations in brain activity (Buzsáki, 2006), the simultaneous recordings from numerous scalp electrodes result in a high-dimensional dataset that reflects complex spatial patterns of neural function (Makeig et al., 1995).

To overcome the analytical challenges, deep learning offers a robust and scalable solution. Unlike traditional methods that rely on hand-crafted features, deep learning models can automatically extract meaningful representations directly from raw data. Specifically, Long Short-Term Memory (LSTM) networks and Convolutional Neural

Networks (CNNs) have demonstrated strong performance in EEG-based seizure detection (Fraiwan & Alkhodari, 2020; Tian et al., 2019). CNNs are adept at capturing spatial features in frequency-transformed representations like spectrograms, whereas LSTMs excel at modelling the sequential dependencies inherent in raw EEG signals.

While LSTM and CNN architectures handle specific aspects of EEG complexity, these models often focus on isolated modalities. The full characterization of EEG signals requires capturing both short- and long-term temporal patterns, nonlinear spectral features, and inter-channel dependencies in a unified manner. To this end, multimodal deep learning has emerged as a promising approach by integrating complementary information from multiple data representations, thereby offering a more holistic understanding of brain activity. Despite these advances, the full potential of multimodal approaches remains underexplored, leaving significant room for further investigation into their effectiveness and broader applicability. Building on this idea, the current thesis proposes a novel multimodal deep learning model to detect seizures from EEG signals, tackling the limitations of single-modality approaches by leveraging both temporal and spectral representations to capture a broader range of signal characteristics, improving robustness, sensitivity to diverse seizure patterns, and generalizability across patients. The key innovation lies in the integration of temporal and spectral representations, captured respectively by LSTM and CNN modules, and Transformer to address cross-channel dependencies. Unlike conventional models that treat these modalities independently, the proposed method employs a fusion strategy in which the outputs of the LSTM and CNN are concatenated and passed through a Transformer architecture, enabling the model to capture complex dependencies across both modalities and channels. This multimodal design complements the individual strengths of temporal and spectral modelling by capturing a broader range of EEG characteristics.

In addition to introducing the proposed multimodal architecture, this study also aims to evaluate the performance of two hybrid baseline models: an LSTM-Transformer, designed to capture both short-term dynamics and long-range temporal dependencies in raw EEG signals, and a CNN-Transformer, developed to extract discriminative spectral features from time–frequency representations (spectrograms). The research further investigates whether integrating these modalities within a unified framework enhances seizure detection performance compared to unimodal approaches. Accordingly, the central research question guiding this work is: Does the combination of temporal and spectral features, along with modelling cross-channel dependencies, improve seizure classification accuracy?

To guide the reader through this investigation, the thesis is structured into seven chapters. Chapter 1 introduces the research problem, objectives, and central question. Chapter 2 provides a theoretical overview of seizures and abnormal brain activity, including key clinical concepts such as Lateralized Periodic Discharges. Chapter 3 introduces deep learning models relevant to this work, including LSTM, CNN, and Transformer architectures, and outlines the foundations of multimodal learning. Chapter 4 reviews prior work on EEG-based seizure detection and highlights the existing research gap. Chapter 5 describes the research methodology, covering data collection, preprocessing, model design, and evaluation criteria. Chapter 6 presents the experimental results, comparing the performance of unimodal and multimodal models across seizure types. Finally, Chapter 7 discusses the findings in depth, reflects on their implications, addresses limitations, and outlines directions for future research to improve seizure detection and model generalization.

2. Abnormal Brain Activity in EEG

2.1. Electroencephalography (EEG)

Brainwaves are patterns of electrical activity produced by neuronal firing, indicating the brain's functional state (Niedermeyer & da Silva, 2005). These electrical signals, generated by ionic currents within neurons, can be detected non-invasively at the scalp surface through voltage fluctuations (Nunez & Srinivasan, 2006).

EEG offers a non-invasive, real-time means of monitoring brainwave patterns (Sanei & Chambers, 2007). It involves placing electrodes on the scalp to capture voltage fluctuations resulting from ionic currents within neurons so that it can detect the brain's electrical activity (Fisch, 2009). Figure 2.1 illustrates the international 10–20 system, the most widely used in both clinical and research settings, providing consistent positioning of electrodes across individuals and facilitates the interpretation and comparison of EEG recordings. According to the figure, each electrode label corresponds to a specific brain region: F (frontal), T (temporal), C (central), P (parietal), O (occipital), and Fp (frontopolar). The accompanying numbers indicate lateral positioning, with odd numbers (e.g., F3) referring to the left hemisphere, even numbers (e.g., F4) to the right hemisphere, and 'z' (e.g., Cz) designating midline placements. This spatial nomenclature facilitates consistent recording and interpretation across studies and clinical applications. In addition to EEG electrodes, an EKG (electrocardiogram) channel is often included to monitor cardiac activity and assist in distinguishing neurological signals from physiological artifacts.

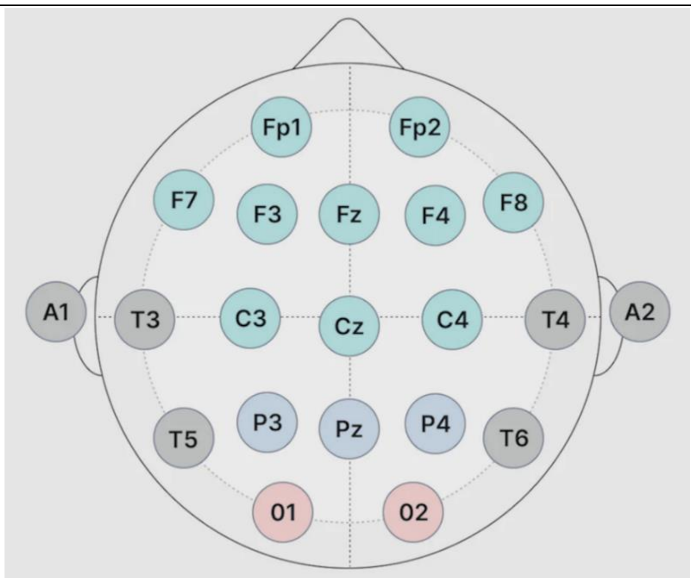

Figure 2.1: International 10-20 EEG Electrode Placement System (Emotiv (n.d.))

The history of EEG backs to the 19th century, with early experiments by Carlo Matteucci, Emil Du Bois-Reymond, and Richard Caton. Caton first recorded brain activity in 1875. Following this, Pioneers such as Hans Berger further advanced the field, recording the first human EEG in the 1920s. Throughout the 1930s-1950s, multichannel recordings, biological amplifiers, and studies of epileptic seizures revolutionized EEG research. Modern EEG technology now includes long-term recordings and integration with other neuroimaging techniques, like fMRI and MEG, for comprehensive brain activity analysis. Modern EEGs is non-invasive long-term recordings to study brain activity in greater depth (Priyanka A., et al, 2016).

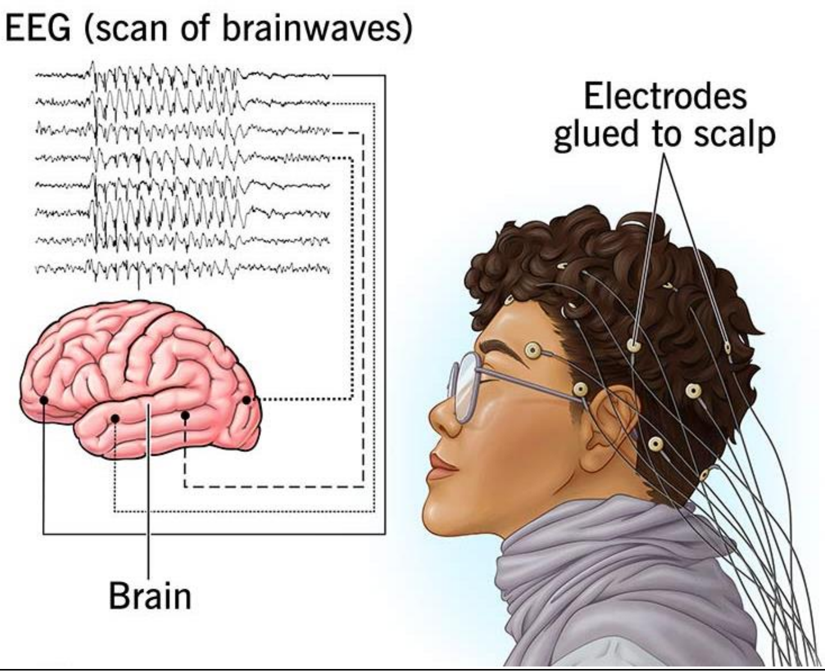

Figure 2.2: How EEG works (my.clevelandclinic.org).

Figure 2.2 illustrates the overall process of how EEG works, from signal generation in the brain to its digital analysis. EEG captures electrical activity produced by synchronous firing of cortical neurons, particularly from pyramidal cells. Electrodes placed on the scalp detect these microvolt-level signals, which are then amplified and filtered to remove noise and retain relevant frequency components. The analog signals are converted into digital form using an analog-to-digital converter (ADC), enabling computational analysis.

Moreover, Brainwaves are rhythmic or repetitive patterns of neural activity in the central nervous system, typically measured in microvolts (μV), and traditionally categorized into five principal frequency bands: delta, theta, alpha, beta, and gamma, each associated with specific mental states and neurophysiological functions. These bands are illustrated in Table 2.1. Hans Berger, the pioneer of EEG, first identified alpha and beta waves in 1929, associating alpha with a relaxed wakeful state and beta with active thinking (Berger, 1929). Later, Walter (1936) characterized delta activity, typically observed during deep sleep. Wolter and Dovey (1944) introduced the concept of theta waves, linked to drowsiness and meditation (McLeod et al., 1983). Gamma waves were identified by Jasper and Andrews in 1938 as oscillations exceeding 30 Hz, reflecting cognitive integration and high-level information processing (Jasper & Andrews, 1938). Table 2.1 and Figure 2.3 represents wave samples with dominant frequencies belonging to beta, alpha, theta, and delta bands and gamma waves.

Table 2.1: Characteristics of the five basic brain waves (Abhang et al, 2016)

Frequency band	Frequency	Brain states
Gamma	> 35 Hz	Concentration, solving problems
Beta	12–35 Hz	Anxiety dominant, active, external attention, relaxed
Alpha	8–12 Hz	Very relaxed, passive attention
Theta	4–8 Hz	Deeply relaxed, inward focused
Delta	0.5–4 Hz	Sleep

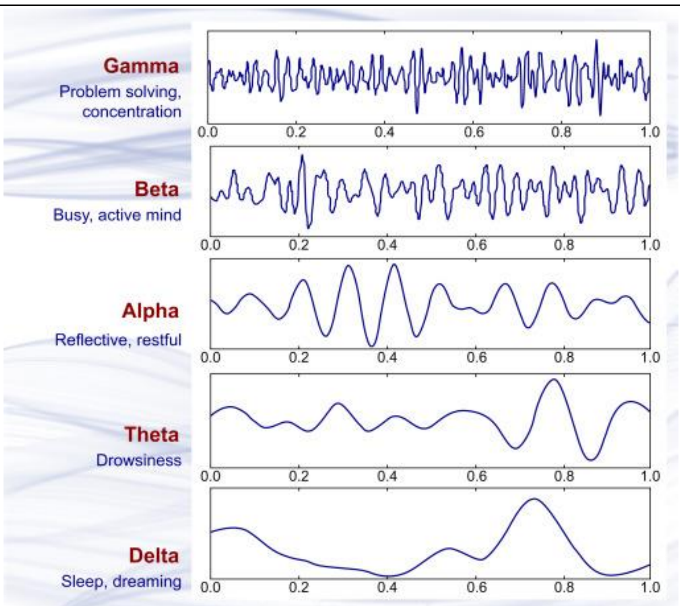

Figure 2.3: Brain wave samples by frequency band (Priyanka A. et al, 2016).

In a healthy brain, these frequency bands dynamically change based on cognitive state, behavioural condition, and age. A relaxed adult with closed eyes will typically exhibit dominant alpha rhythms over posterior regions, while beta waves become more prominent during active mental tasks. Similarly, delta and theta bands are prominent during different stages of sleep and are more prevalent in younger individuals. These rhythmic variations are not only markers of neural state transitions but also reflect underlying functional connectivity within cortical networks (Buzsáki & Draguhn, 2004; Başar et al., 2001).

Crucially, trained neurologists can recognize these patterns to assess normal developmental trajectories and identify pathological states. Altered frequency distributions, abnormal wave morphologies, or sudden bursts of rhythmic activity can indicate neurological disorders such as epilepsy, encephalopathies, or degenerative conditions. The temporal and spatial organization of these waveforms enables both qualitative and quantitative analyses of brain health.

In the context of seizure detection, which is central to this study, understanding normal brain rhythms is essential for distinguishing between physiological patterns and pathological discharges. Seizures are often characterized by distinct changes in amplitude, frequency, synchronization, and wave morphology (Fisher et al., 2005; Niedermeyer & da Silva, 2005). These deviations from baseline rhythms, including spikes, sharp waves, and rhythmic bursts, are typically observed during epileptiform events and form the core basis for clinical diagnosis and algorithmic detection in automated systems (Smith, 2005; Shoeb, 2009).

Therefore, a thorough understanding of typical EEG frequency bands and their normal variations across different physiological states lays the foundation for identifying abnormal EEG patterns, a topic further discussed in the following section.

2.2. Seizure

Under normal physiological conditions, brain waves exhibit structured and predictable patterns that reflect the functional and cognitive state of the brain (Niedermeyer & da Silva, 2005; Sanei & Chambers, 2013). Changes in brain waves patterns in EEG recordings often indicate abnormal brain activity. Such abnormalities may manifest in various forms, including the distortion or loss of normal patterns, the emergence or intensification of abnormal waveforms, or in some cases, the complete absence of detectable EEG signals. These deviations suggest that the brain is not functioning within its typical physiological range. Sharbrough et al. (1999) made an effort to categorize abnormalities based on their waveform characteristics and clinical conditions. The first proposed category was termed widespread intermittent slow-wave activity, typically occuring within the delta frequency range and is often linked to generalized cerebral dysfunction. It was followed by the second type referred to as bilateral persistent EEG patterns. This type is linked to impaired conscious processing and reduced cortical responsiveness. Finally, The third category was named as focal persistent EEG activity, which is indicative of localized disturbances, often reflecting abnormalities restricted to specific brain regions (Sharbrough, F. W., et al, 1999).

Among the most prominent manifestations of abnormal electrical brain activity is a seizure, defined as a sudden, uncontrolled discharge of neurons that disrupts normal brain function. This disruption can result in unpredictable changes in behavior, movement, sensation, awareness, emotions, or consciousness (Chang & Lowenstein, 2003; Fisher et al., 2014). When such seizures occur recurrently without an identifiable acute cause, the condition is classified as epilepsy (ILAE, 2014). Understanding the EEG signatures of seizure activity is therefore central to accurate diagnosis, monitoring, and treatment planning.

Abnormal EEG patterns are commonly linked to seizure-prone states and other neurological dysfunctions. One such pattern is burst-suppression activity, often observed in encephalopathies, anesthesia, or during postictal states. This pattern is characterized by high-amplitude bursts of activity alternating with periods of suppression and may become more pronounced during hyperventilation, drowsiness, or eye closure, conditions known to enhance cortical excitability and reveal latent epileptiform activity (Husain, 2011; Sutter et al., 2016). Although not specific to epilepsy, the presence of burst-suppression patterns can indicate a brain highly vulnerable to seizures.

Another important EEG abnormality is bilateral persistent rhythmic activity, such as generalized/lateralized rhythmic delta activity or periodic discharges. These patterns are frequently observed in patients with altered levels of consciousness and are especially relevant in cases of non-convulsive status epilepticus (NCSE), toxic-metabolic encephalopathies, and diffuse brain injury (Trinka et al., 2015; Hirsch et al., 2021). While not always epileptic in origin, these patterns can mask subclinical seizures and require careful interpretation in context.

Finally, focal persistent EEG abnormalities play a key role in the localization of seizure onset zones, particularly in focal epilepsy. These abnormalities may involve sustained focal slowing, asymmetry in frequency or amplitude, or the presence of interictal epileptiform discharges (IEDs) such as spikes or sharp waves (Kane et al., 2017; Smith, 2020).

Events recorded in EEG data are typically categorized into three temporal phases. Ictal events occur during the seizure itself and are often marked by distinct rhythmic discharges. Preictal events precede the seizure and may span several minutes to several hours, offering a potential window for seizure prediction. Interictal events refer to all other abnormal discharges occurring between seizures and often include spikes or sharp waves not associated with overt clinical symptoms (Zurdo-Tabernero et al., 2023).

In a large-scale study involving nearly 5,000 critically ill patients, Rodriguez Ruiz et al. (2017) investigated the relationship between EEG patterns and seizure risk. The study found that certain rhythmic and periodic discharges were significantly associated with increased seizure likelihood. Among these, lateralized periodic discharges (LPDs) were most strongly linked to seizures, followed by lateralized rhythmic delta activity (LRDA) and generalized periodic discharges (GPDs). Generalized rhythmic delta activity (GRDA) was associated with the lowest seizure risk among the patterns studied.

Building upon this evidence, the American Clinical Neurophysiology Society (ACNS) updated its Standardized Critical Care EEG Terminology in 2021 to include classification of seizures based on rhythmic and periodic EEG features. According to this framework, generalized periodic discharges (GPDs) and generalized rhythmic delta activity (GRDA) are grouped under generalized seizure activity due to their bilateral, synchronous distribution. In contrast, lateralized periodic discharges (LPDs) and lateralized rhythmic delta activity (LRDA) are categorized as focal (or lateralized) seizures, as their activity is confined to one cerebral hemisphere. This classification has become a valuable tool in both clinical neurophysiology and automated seizure detection systems.

2.3. Lateralized Periodic Discharges

According to ACNS in 2021, Periodic Discharges (PDs) refer to the repetitive occurrence of waveforms that exhibit relatively uniform morphology and duration, with clearly identifiable intervals separating each event (Hirsch et al, 2021). These waveforms, illustrated in Figure 2.4, occur at approximately regular intervals, and the inter-discharge interval (i.e., the time between successive discharges) is a defining feature. The term "nearly regular intervals" is operationally defined as having a cycle length that varies by less than 50% between successive cycles in more than 50% of the observed pairs. As can be seen in Figure 2.5, A discharge is characterized as a waveform lasting less than 0.5 seconds, regardless of the number of phases, or a waveform equal to or exceeding 0.5 seconds in duration but containing no more than three phases. In contrast, a burst is defined as a waveform lasting 0.5 seconds or longer and comprising at least four phases. Both discharges and bursts are required to stand out clearly from the background EEG activity to be classified as such.

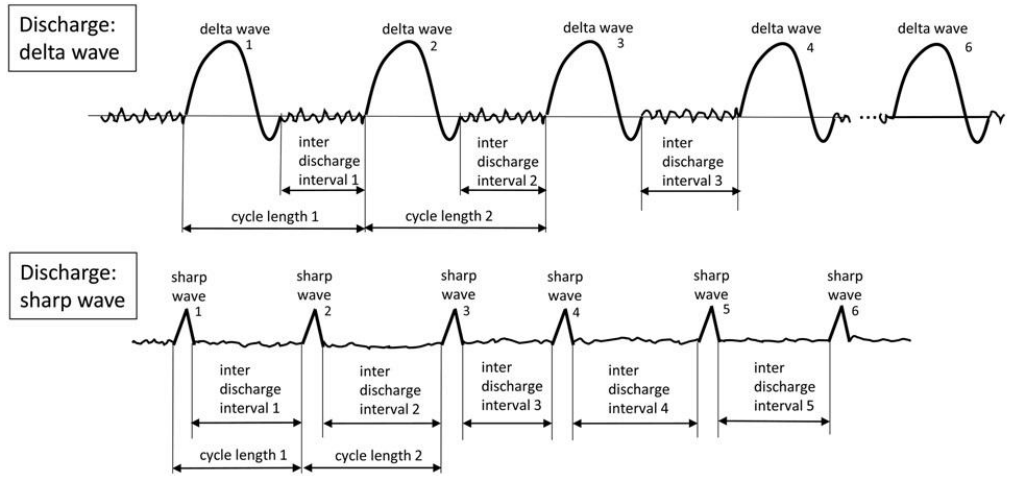

Figure 2.4: Periodic Discharges (PDs) (Hirsch et al, 2021)

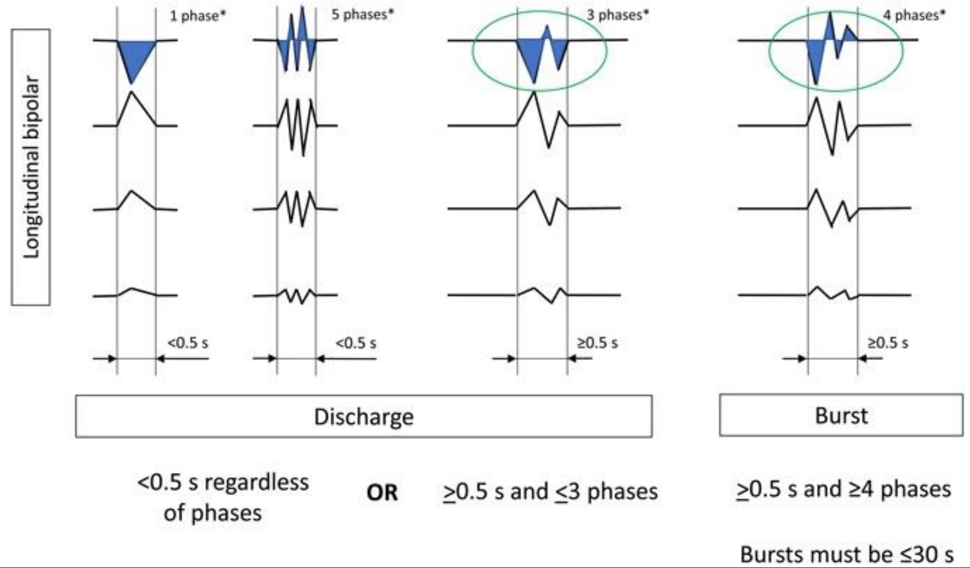

Figure 2.5: Discharge vs. Burst (Hirsch et al, 2021).

LPDs are unilateral periodic discharges consisting of sharp waves or spikes occurring at 0.5–4 Hz, with moderate to high amplitude (50–300 μV). These discharges are asymmetric and appear over one hemisphere, often indicating localized pathology like stroke, infection, or tumors. LPDs tend to evolve over time, both in morphology and frequency, and are associated with increased risk of seizures. They represent ongoing focal cortical irritation and may correlate with clinical symptoms like altered consciousness. Such patterns can be seen in Figure 2.6. When periodic discharges appear on both hemispheres but with consistently higher amplitude on one side, they are termed bilateral asymmetric. These patterns are classified as lateralized, bilateral asymmetric discharges. For example, PDs seen bilaterally but stronger on the left are referred to as "Left LPDs, bilateral asymmetric." Figure 2.7 illustrates this pattern. If discharges are observed on both sides but consistently start earlier on one hemisphere, they are known as bilateral asynchronous. This pattern is described as lateralized, bilateral asynchronous. For instance, bifrontal LPDs leading on the left are termed "Left LPDs, bilateral asynchronous." Figure 2.8 illustrates this activity with a typical interhemispheric time lag.

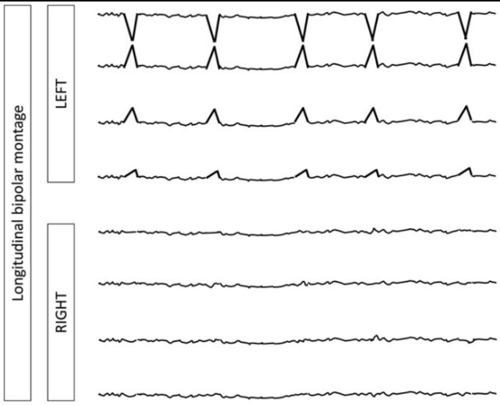

Figure 2.6: LPDs, unilateral (Hirsch et al, 2021).

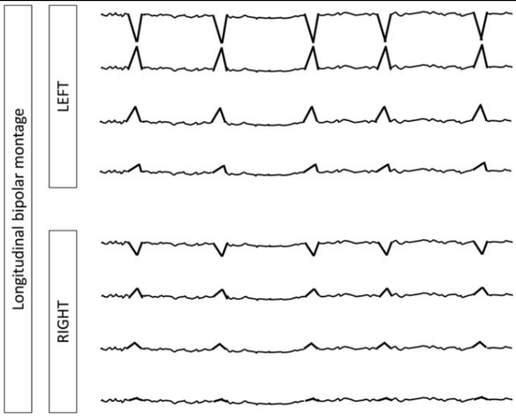

Figure 2.7: LPDs, bilateral asymmetric (Hirsch et al, 2021).

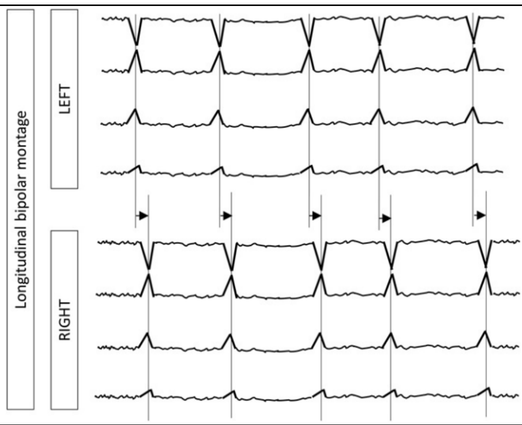

Figure 2.8: LPDs, bilateral asynchronous (Hirsch et al, 2021).

3. Multimodal Deep Learning Models

3.1. Modalities

A modality refers to a way of perceiving or representing natural phenomena, such as speech, images, or vibrations. Modalities range from raw, like audio signals and EEG signals, to abstract such as detected emotions or objects (Baltrušaitis et al., 2019). Multimodal learning involves the integration and interpretation of multiple, diverse data modalities that are interconnected and interact with one another to enhance understanding and prediction (Ramachandram & Taylor, 2017).

According to Baltrušaitis et al. (2019) and Ramachandram & Taylor, (2017), three foundational principles guide multimodal research:

- **Heterogeneity**: Modalities differ in structure, format, and representation. For instance, speech and images vary significantly in how information is encoded and processed.
- **Connections**: Despite their differences, modalities often share complementary information, identifiable through statistical patterns or semantic relationships.
- **Interactions**: New meaning emerges when modalities are integrated, enabling richer inference than any single modality alone.

Building upon these principles, this study proposes a multimodal approach that combines two complementary EEG modalities: temporal features derived from raw EEG time-series data, and spectral features extracted from their time–frequency representations (spectrograms). This combination leverages both short- and long-term temporal patterns and frequency-domain dynamics, which have been shown to be useful in EEG-based classification tasks (Roy et al., 2019). To effectively process and fuse these diverse features, this study employs a multimodal deep learning framework. Accordingly, the following sections delve into the deep learning models used in this study, including LSTM, CNN, and Transformer architectures.

3.2. Deep Learning Architectures

Deep learning is one of the most transformative developments in computer science, impacting nearly all scientific domains. This section briefly traces its evolution from foundational ideas to modern applications.

In 1957, Frank Rosenblatt introduced the perceptron, a two-layer model capable of recognizing simple patterns (F. Rosenblatt, 1958). However, enthusiasm faded when Minsky and Papert showed that it could not learn the XOR function (M. Minsky and S. A. Papert, 1969). Further skepticism arose in the 1970s and 1980s with the universal approximation theorem, which proved that a single hidden layer could approximate any continuous function (G. Cybenko, 1989; K. Hornik, 1991). While theoretically powerful, such shallow models lacked efficiency and scalability.

Progress stalled for several years due to difficulties in training deep networks, which contributed to a period known as the “AI winter,” characterized by reduced funding and waning interest in artificial intelligence research (Werbos, 1975). The turning point arrived with the development and widespread adoption of the backpropagation algorithm in the 1980s. This method enabled efficient computation of gradients, allowing deep neural networks to be trained more effectively and overcoming many previous obstacles. Building on this breakthrough, Deep Neural Networks (DNNs), typically organized as multilayer perceptrons, transformed machine learning by automating feature extraction (LeCun, Bengio, & Hinton, 2015). Unlike their shallow predecessors, DNNs capture complex, nonlinear patterns through hierarchical layers, providing more powerful and flexible representations of data (Jordan & Mitchell, 2015).

Figure 3.1 provides an overview of a typical deep learning architecture. The model begins with an input layer, which receives raw data such as EEG signals or their transformed representations. This data is then passed through multiple hidden layers, each composed of numerous interconnected neurons. These hidden layers apply nonlinear activation functions to the incoming signals, enabling the network to learn complex, hierarchical features that range from simple patterns in early layers to more abstract representations in deeper layers. The depth and number of hidden layers allow the model to progressively capture increasingly sophisticated relationships within the data. The connections between neurons are weighted, and these weights are iteratively adjusted during training using algorithms like backpropagation combined with an optimization method such as gradient descent. This process minimizes the error between the predicted and actual outputs. Finally, the transformed information reaches the output layer, which produces predictions tailored to the task, such as classification probabilities in seizure detection. Depending on the task, the output layer may use activation functions like softmax for multiclass classification or sigmoid for binary outcomes.

This layered structure allows deep learning models to automatically extract relevant features and model complex dependencies without manual feature engineering, making them particularly effective for challenging tasks involving high-dimensional and nonlinear data like EEG signals.

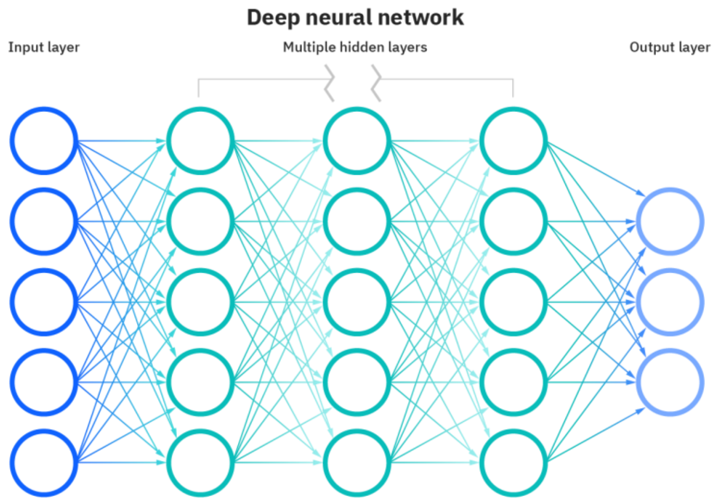

Figure 3.1: An overview of a typical deep learning architecture (Image via IBM).

3.3. LSTM

The LSTM architecture was introduced in 1997 by Sepp Hochreiter and Jürgen Schmidhuber in the paper "Long Short-Term Memory." It was designed to tackle the vanishing gradient problem, which had limited the effectiveness of traditional Recurrent Neural Networks (RNNs) in modeling long-term dependencies in sequential data. The key innovation in LSTM was its memory cell, which allowed the network to retain important information over long time steps without suffering from gradient decay. The addition of gates—Input, Forget, and Output gates—enabled the selective updating, erasing, and reading of information from the memory cell (Hochreiter, S., & Schmidhuber, J. 1997).

Following their introduction, LSTMs underwent several enhancements to boost performance. A key improvement was the addition of the forget gate by Felix Gers in 2000, which enabled better control over memory retention and made LSTMs more robust to noise (Gers, F. A., et al., 2000). Early 2000s applications in speech recognition and handwriting generation demonstrated LSTMs' ability to model long-term dependencies.

LSTMs gained momentum in the 2010s, particularly after a 2013 Google study led by Alex Graves showed they outperformed traditional models in speech recognition (Graves, A., et al., 2013). This success spurred broader adoption in machine translation and text generation.

A breakthrough came with Sequence-to-Sequence (Seq2Seq) models in 2014, where LSTMs were used for both encoding and decoding sequences, achieving strong results in neural machine translation (Sutskever, I., et al., 2014). Bidirectional LSTMs (BiLSTMs), which process data in both time directions, further improved performance in tasks like POS tagging, named entity recognition, and speech synthesis (Graves, A. & Schmidhuber, J., 2005).

Beyond language, LSTMs were used in handwriting generation (Graves, A., 2013) and in models like *Show and Tell* for image captioning, combining CNNs with LSTM decoders (Vinyals, O., Toshev, A., 2015). In healthcare, LSTMs have been widely used for time-series analysis of ECGs and patient records, offering powerful tools for predicting clinical outcomes (Lipton, Z. C., et al., 2016). LSTMs addressed key limitations in RNNs and enabled effective modeling of long-term dependencies across domains from language to healthcare.

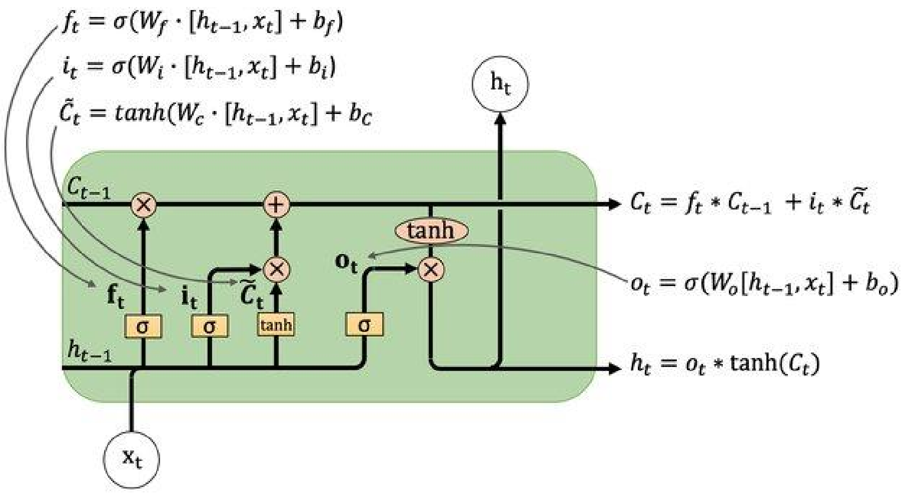

Figure 3.2: Architecture of a LSTM unit (d2l.ai/)

***f_t**: Forget gate*

***i_t**: Input gate*

***C_t**: Cell state*

***o_t**: Output gate*

***h_t**: Hidden state*

***$\bar{c}_t$**: Candidate cell*

Figure 3.2 illustrates the architecture of a LSTM unit. At the core of an LSTM unit is a cell state, which acts as the memory of the network and flows through the unit with minor linear interactions, preserving information across time steps. The flow of information into and out of this cell state is controlled by three key gates:

- **Forget Gate**: this gate plays a crucial role in regulating memory flow within the LSTM cell. At each time step, it receives the current input vector and the previous hidden state, which are concatenated and passed through a sigmoid activation function. The resulting vector contains values between 0 and 1, representing the degree to which each element of the previous cell state should be retained or discarded. A value close to 0 indicates that the corresponding information is no longer relevant and should be forgotten, while a value near 1 signifies that the information remains important for ongoing temporal reasoning. This mechanism enables the LSTM to dynamically ignored outdated or irrelevant features, addressing the issue of long-term dependencies in sequential data.

- **Input Gate**: it is responsible for incorporating new information into the memory cell. Similar to the forget gate, it processes the concatenation of the current input and the previous hidden state. A sigmoid layer first determines which dimensions of the input should be updated, acting as a gatekeeper. In parallel, a tanh activation generates a vector of candidate values representing potential additions to the cell state. These two outputs are then combined through element-wise multiplication, resulting in a filtered update vector that selectively injects relevant information into the cell's long-term memory. This dual-step process ensures that only significant features are integrated, improving the model's ability to learn and retain meaningful temporal patterns.
- **Output Gate**: this gate defines what part of the internal memory (cell state) should influence the output at the current time step. It operates in two stages. First, the current cell state is passed through a tanh activation to squash its values into a manageable range, effectively normalizing the signal. Then, a sigmoid layer based on the current input and previous hidden state produces a gating vector that determines which dimensions of the transformed cell state should be exposed to the outside. The element-wise product of these two vectors forms the new hidden state, which serves both as output for the current time step and as an input to the next LSTM cell. This controlled exposure allows the model to regulate information flow across time, balancing memory retention with adaptability.

The combination of these gates enables the LSTM to selectively remember or forget information, making it particularly effective for modeling complex temporal patterns, such as those found in raw EEG recordings.

In this study, LSTMs are utilized to model the temporal dependencies within each EEG channel, allowing the network to learn patterns associated with different types of brain activity over time.

3.4. CNN

The evolution of CNN began with Yann LeCun's 1989 work on handwritten digit recognition using backpropagation (LeCun, Y., et al., 1989), leading to the development of LeNet-5 in 1998, a foundational architecture for document recognition (LeCun, Y., et al., 1998).

A major leap came with AlexNet in 2012, which revolutionized image classification by leveraging ReLU activations and GPU training, winning the ImageNet challenge and popularizing deep learning (Krizhevsky, A., et al., 2012). This was followed by VGGNet in 2014, which emphasized depth using simple 3×3 filters, achieving high accuracy with a deeper but straightforward architecture (Simonyan, K. & Zisserman, A., 2014).

In the same year, GoogLeNet (Inception) introduced a novel module combining multiple filter sizes, reducing parameters while maintaining performance (Szegedy, C., et al., 2015). Then in 2015, ResNet addressed vanishing gradients via residual connections, enabling networks with over 100 layers and setting new records in image classification (He, K., et al., 2016).

Later advancements include EfficientNet (2019), which optimized accuracy and efficiency through compound scaling (Tan, M. & Le, Q. V., 2019), and Vision Transformers (ViT), which in 2020 applied self-attention mechanisms from NLP to vision tasks, sometimes outperforming CNNs (Dosovitskiy, A., et al., 2021).

Figure 3.3 illustrates a schematic diagram of a basic CNN architecture, which is particularly well-suited for processing structured grid-like data such as images or time–frequency representations (e.g., spectrograms). The architecture is composed of several key layers that sequentially transform the input into a set of increasingly abstract and informative features for classification.

The process begins with the input layer, which typically receives a 2D input such as an image or spectrogram, represented as a matrix of pixel or intensity values. This input is passed to one or more convolutional layers, where small filters (or kernels) slide over the input matrix to extract local features like edges, patterns, and textures. These filters learn to activate on specific patterns during training, enabling the network to detect meaningful spatial features.

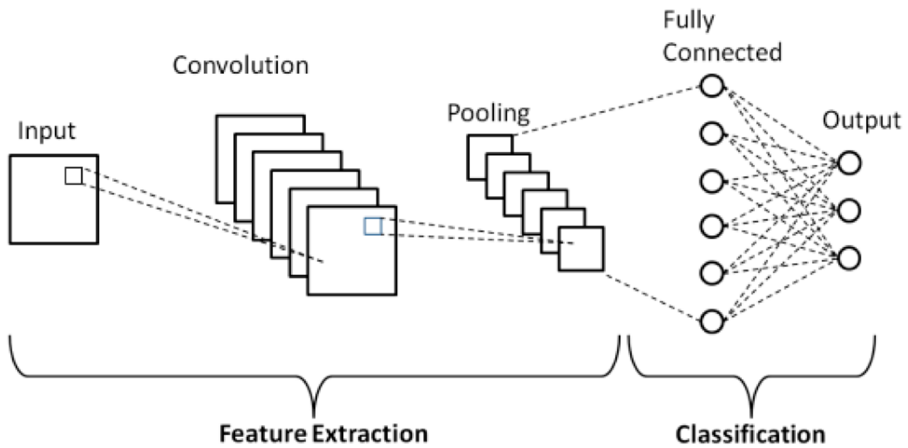

Figure 3.3: Schematic diagram of a basic CNN architecture (Rhee, 2018).

Each convolutional layer is typically followed by a nonlinear activation function, such as the Rectified Linear Unit (ReLU), which introduces non-linearity into the model and allows it to learn complex mappings from input to output. After the activation, a pooling layer (commonly max pooling) reduces the spatial dimensions of the data, which decreases computational load and provides a degree of spatial invariance.

As the data passes through multiple convolutional and pooling layers, the network captures increasingly abstract and high-level features. These feature maps are then flattened and passed to one or more fully connected (dense) layers, which interpret the learned features and produce the final output. In classification tasks, the last layer typically uses a softmax or sigmoid activation function to output class probabilities.

This hierarchical feature learning makes CNNs highly effective in tasks involving structured patterns and spatial correlations, such as classifying EEG spectrograms based on seizure activity.

3.5. Transformers

The Transformer architecture, introduced by Vaswani et al. (2017) in the paper "Attention Is All You Need", marked a paradigm shift in deep learning models for sequential data. Prior to this, RNNs and LSTMs dominated tasks such as language modeling and machine translation. However, these models suffered from inherent limitations, particularly their sequential nature, which constrained parallelization and often struggled with long-range dependencies due to vanishing gradients. The Transformer overcame these issues by introducing a self-attention mechanism, which allows the model to weigh the importance of different elements in a sequence simultaneously. Each input token is projected into query, key, and value vectors, and the attention weights are computed based on the dot product of queries and keys, scaled and passed through a SoftMax function. Multi-head attention, another core component, enables the model to capture multiple types of relationships in parallel. Since Transformers do not process tokens sequentially, positional encodings are added to the input embeddings to retain information about the order of the sequence (Vaswani et al., 2017). This architecture significantly improved both accuracy and training efficiency across various natural language processing (NLP) tasks. Subsequent models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) extended the Transformer's encoder to a bidirectional setting, improving contextual understanding. On the generative side, models like GPT (Generative Pre-trained Transformer) and its successors leveraged Transformer decoders for large-scale language modeling and demonstrated strong generalization with minimal task-specific tuning (Radford et al., 2018; Brown et al., 2020).

Beyond NLP, the Transformer was successfully adapted to computer vision through the Vision Transformer (ViT), which treats image patches as tokens and applies the same self-attention mechanisms used in text processing (Dosovitskiy et al., 2020). ViT's performance on image classification benchmarks illustrated the architecture's flexibility and domain-general capacity. A defining feature of the Transformer is its scalability. Unlike RNNs, it allows for parallel computation over entire input sequences, drastically accelerating training on large datasets. This property has underpinned the success of large-scale pre-trained models like GPT-3, which consists of 175 billion parameters and demonstrates strong performance on a variety of language tasks with little or no fine-tuning (Brown et al., 2020). The Transformer has thus become the foundation of many state-of-the-art models across domains. Its design—centered on attention, parallelization, and long-range dependency modeling—has not only enhanced NLP tasks but also influenced developments in vision, audio, and multimodal learning.

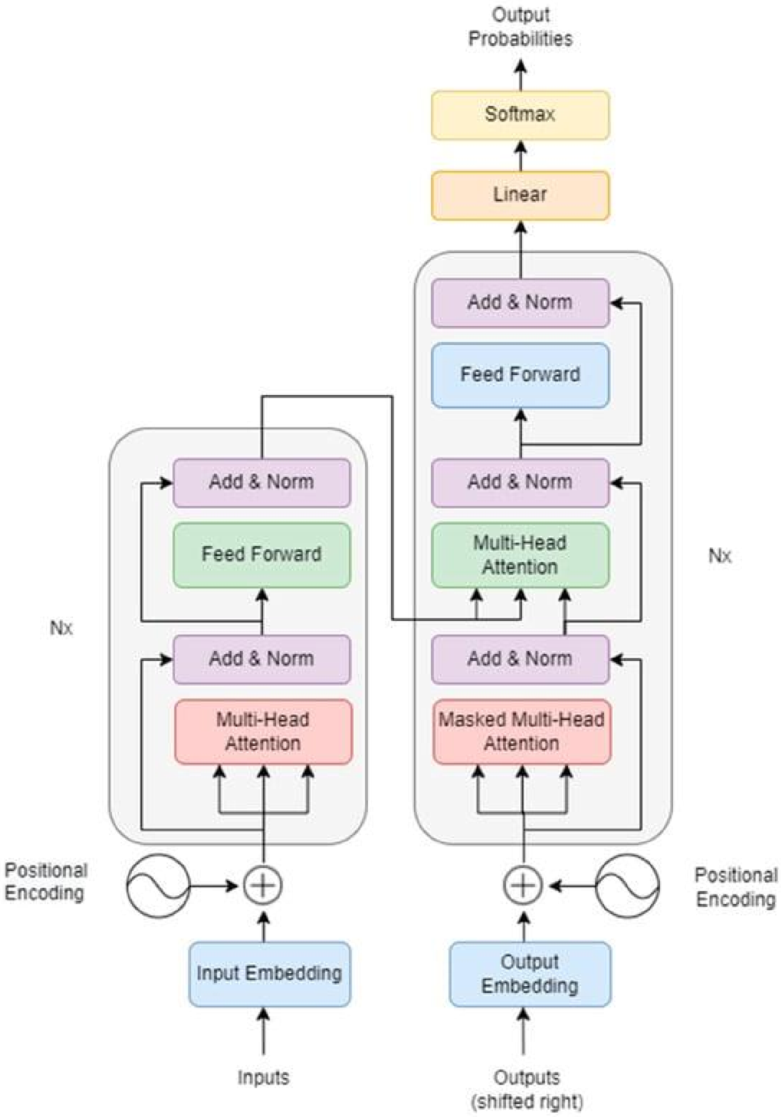

Figure 3.4: The encoder-decoder structure of the Transformer architecture (Vaswani and et al, 2017)

Figure 3.4 presents the architecture of a Transformer. The Transformer's core mechanism is self-attention, which allows each input element (e.g., a time step or feature vector) to attend to all other elements in the sequence, dynamically weighting their importance. This mechanism enables the model to focus on the most relevant parts of the input regardless of their position in the sequence. The architecture typically consists of the following key components:

- **Input Embedding and Positional Encoding**: Since Transformers lack recurrence or convolution, they require positional encoding to inject information about the order of elements in a sequence. These encodings are added to the input embeddings, allowing the model to differentiate between the positions of inputs. A commonly used positional encoding is:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{3.1}$$

  Where pos is the position, i is the dimension index, d_{model} is the model dimensionality

- **Self-Attention Mechanism**: Each token (or EEG feature vector) is transformed into a query (Q), key (K), and value (V) using learned linear projections. The scaled dot-product attention is then computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.2}$$

  Where $Q \in R^{n\times d_k}$ (queries), $K \in R^{n\times d_k}$ (keys), $V \in R^{n\times d_k}$ (values), d_K is the dimensionality of the keys, n is the sequence length.
  The dot product of Q and K measures similarity, and the result is scaled by $\frac{1}{\sqrt{d_k}}$ to prevent. The softmax function normalizes the attention scores, ensuring they form a probability distribution.

- **Multi-Head Self-Attention**: This module applies multiple attention operations in parallel (heads), each capturing different types of dependencies between elements.

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \tag{3.3}$$

  Where each head_i is computed as:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$
$$W_i^Q \in R^{d_{model}\times d_k}$$
$$W_i^k \in R^{d_{model}\times d_k}$$
$$W_i^v \in R^{d_{model}\times d_k}$$
$$W^o \in R^{hd_v\times d_{model}}$$

  Here, the inputs are projected h times, and attention is computed in parallel for each head. The results are concatenated and linearly transformed.

- **Feedforward Layers**: After the attention layers, a position-wise fully connected feedforward network is applied independently to each position in the sequence with a ReLU activation in between. This allows the model to introduce non-linearity and transform representations.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{3.4}$$
$$W_1 \in R^{d_{model} \times d_{ff}}$$
$$W_2 \in R^{d_{ff} \times d_{kmodel}}$$

- **Layer Normalization and Residual Connections**: Each sub-layer (attention or feedforward) is followed by a residual connection and layer normalization, which help with training stability and convergence.

$$\begin{gathered} Encoder\ Layer(x) = LayerNorm\big(x + MultiHead(x, x, x) + FFN(x)\big) \\ Decoder(x) = LayerNorm\big(x + MaskedMultiHead(x, x, x) \\ + MultiHead(x, Encoder\ Output, Encoder\ Output) + FFN(x)\big) \end{gathered} \tag{3.5}$$

- **Stacked Encoder or Decoder Blocks**: In encoder-only models like EEG analysis, several layers of attention and feedforward blocks are stacked. In full encoder-decoder structures such as translation tasks, the decoder is added with masked attention to prevent future information leakage.

4. Related works

4.1. Overview

This section critically reviews recent studies to identify current limitations and potential opportunities in the field. Researchers have explored various models to extract meaningful patterns from the complex temporal and spectral characteristics of EEG data. These approaches include both traditional machine learning algorithms and deep learning architectures, each offering different strengths in handling signal variability, noise, and interpretability. While notable progress has been made in improving classification accuracy and early seizure detection, key challenges remain, particularly in terms of model generalizability, multimodal integration, and clinical applicability. By analyzing the methodological trends, model performances, and evaluation strategies across the literature, this review aims to establish a foundation for identifying unresolved challenges. Table 4.1 provides a summary of the most relevant and up-to-date studies, outlining their objectives, employed models, and reported results. Based on this review, the section concludes by articulating the key research gaps that motivate the present study.

4.2. Prior studies

Dutta et al. (2023) proposed a multi-head self-attention model that captures both short-term temporal dynamics and long-distance dependencies in EEG data. The model also uses a bidirectional LSTM (BiLSTM) to extract features and log SoftMax for classification. Tested on the CHB-MIT dataset, it achieved an accuracy of 96.2%, sensitivity of 96.5%, specificity of 97.04%, and F1-score of 96.6%. This approach enhances efficiency and generalization in detecting seizures across multi-channel EEG recordings, offering valuable support for epilepsy diagnosis (Dutta et. al, 2024).

Xia et al. (2023) introduced a hybrid LSTM-Transformer model to predict epileptic seizures using scalp EEG data. By applying the short-time Fourier transform (STFT) to extract time–frequency features, the model distinguishes between the interictal and preictal states. Combining the LSTM's strength in processing variable-length information with the Transformer's ability to capture long-distance dependencies, the model offers robust feature extraction. Tested on the CHB-MIT dataset, it achieved a sensitivity of 99.75%, an AUC of 99.39%, and a 0/h false prediction rate, demonstrating its effectiveness for accurate seizure prediction (L. Xia et al., 2024).

Tsiouris et al. (2018) introduced a Long Short-Term Memory (LSTM) network for epileptic seizure prediction using EEG signals. A thorough pre-analysis was conducted to optimize the LSTM architecture, and a two-layer LSTM network was selected to evaluate performance across four preictal windows (15 minutes to 2 hours). The model utilized various features, including time-frequency domain features, cross-correlation between EEG channels, and graph-theoretic features. Tested on the CHB-MIT dataset, the model predicted 185 seizures with a high sensitivity and a false prediction rate (FPR) of 0.11–0.02 false alarms per hour (Tsiouris, K. M., 2018).

Hussein et al. (2022) proposed a novel approach for forecasting epileptic seizures using EEG data by designing a model specifically capable of analyzing data from multiple channels simultaneously. Instead of working with raw signals directly, they converted EEG recordings into visual patterns known as scalograms, using the continuous wavelet transform to capture both frequency and time information. These scalograms were broken into smaller segments, allowing a specialized transformer model called the Multi-Channel Vision Transformer (MViT) to extract complex features that span spatial, temporal, and spectral domains. By applying this method to three widely used EEG datasets, the researchers demonstrated that their model could accurately detect patterns related to seizures, with particularly high sensitivity for both surface-level with a sensitivity of 99.80% and invasive EEG recordings between 90.28% and 91.15%. These results showed a noticeable improvement over previous deep learning and signal-based techniques in the same domain (Hussein et al, 2022).

Li et al. (2022) propose an EEG-based seizure prediction method using a Transformer-guided CNN (TGCNN), combining the strengths of both CNN and Transformer models. The method utilizes STFT to extract time-frequency features from EEG signals. These features are processed by an alternating structure that captures both local and global dependencies, addressing CNN's limitation in long-distance dependence and Transformer's lack of local feature extraction. The model's final prediction is generated through a global average pooling and fully connected layer. The TGCNN achieved 91.5% sensitivity, a false prediction rate (FPR) of 0.145/h, and an AUC of 93.5% on the CHB-MIT database, and 82.2% sensitivity, 0.06/h FPR, and 83.5% AUC on the Kaggle dataset (Li et al, 2022).

In their study, Kode et al. (2024) proposed a method for detecting epileptic seizures by utilizing both Machine Learning (ML) and Deep Learning (DL) techniques on EEG signals. The research focuses on analyzing time-series data, contrasting with other studies that primarily transform EEG signals into images or time-frequency domain features. Several classifiers were employed, including Extreme Gradient Boosting (XGBoost), TabNet, Random Forest (RF), and a One-Dimensional Convolutional Neural Network (1D-CNN). Among these, the 1D-CNN outperformed the others, achieving an accuracy of 99% on the UCI Epileptic Seizure Recognition dataset. This approach demonstrated a high level of accuracy, sensitivity, and precision, showcasing the efficacy of deep learning for seizure detection in EEG signals (Kode et al. 2024).

Aluvalu et al. (2024) proposed a cognitive smart healthcare framework for seizure prediction using a Multimodal Convolutional Neural Network (MMFCNN) integrated with IoT-cloud technologies. The framework uses smart EEG sensors to record patient data and employs a deep learning model for seizure detection. The MMFCNN consists of three blocks: feature extraction, multimodal fusion, and seizure classification. A unique cross-attention mechanism enhances feature extraction by highlighting key information. The model achieves a seizure detection accuracy of 99.2%, providing a robust solution for real-time healthcare monitoring and emergency response (Aluvalu et al, 2024).

Sun et al. (2024) introduced a novel approach for continuous seizure detection using transformer networks and long-term intracranial electroencephalogram (iEEG) data. The proposed model combines convolutional layers with transformers to calculate attention across EEG channels, eliminating the need for feature engineering or signal transformation. The model was evaluated on two datasets: SWEC-ETHZ and TJU-HH. Results demonstrated a high event-based sensitivity of 97.5% and 98.1%, with false detection rates (FDR) of 0.06/h and 0.22/h, respectively. The model also showed explainability by assigning higher attention to channels near the seizure onset zone, highlighting its potential for clinical application (Sun et al, 2024).

Qi et al. (2024) present a patient-specific seizure prediction method using an improved vision transformer (ViT) model for EEG channel optimization. The method focuses on extracting robust feature representations from multichannel EEG signals and selecting the optimal set of channels for prediction, aiming to reduce the number of channels without compromising performance. By applying time-frequency analysis and segmenting EEG spectrograms into patches, the model, Sel-JPM-ViT, selects critical channels for prediction. The model was tested on the Boston Children's Hospital–MIT EEG dataset and achieved an average classification accuracy of 93.65%, sensitivity of 94.70%, and specificity of 92.78%, while using only three to six channels instead of the full 22 channels, showcasing a reduction in hardware complexity for wearable devices (Qi et al, 2024).

Chiranjeevi et al. (2024) introduced a novel ensemble approach combining RNN and Time Series Transformer (TST) for automatic detection of epileptic seizures from EEG signals. The model processes raw EEG data, removing noise, and utilizes the TST model to extract relevant features. In parallel, a recurrent neural network is used to train the features and classify the EEG data. The model's performance was tested on the CHBMIT and Bonn datasets, achieving an accuracy of 94.31% and an F-score of 91.93%. The study demonstrates that the proposed method can effectively classify seizures using only 18 EEG channels, improving upon existing state-of-the-art methods (Chiranjeevi et al, 2024).

Wei et al. (2019) proposed a long-term recurrent convolutional network (LRCN) to predict epileptic seizures by converting EEG time series into 2D images for multichannel fusion. The model uses a convolutional network to extract features and an LSTM block to learn temporal sequences, identifying preictal segments. Tested on EEG data, the model achieved 93.4% accuracy, 91.88% sensitivity, and 86.13% specificity with a low false prediction rate (0.04 FP/h). Compared to other methods, the LRCN showed 5-9% improvement in sensitivity and specificity, making it a strong candidate for real-time seizure prediction systems (Wei et al, 2019).

Wu et al. (2025) introduced SeizureTransformer, a time-step seizure detection model designed for long continuous EEG recordings. Their architecture combines 1D convolutional encoders, residual CNN stacks, and Transformer encoders followed by a streamlined decoder that outputs per-time-step seizure probabilities. Evaluated on both public and proprietary datasets, the model ranked first in the 2025 International Seizure Detection Challenge, demonstrating significant improvements over prior methods in real-time detection accuracy and latency (Wu et al., 2025)

Feng et al. (2025) developed a deep learning framework called SeizureFormer to forecast long-term seizure risk through analysis of biomarkers recorded by responsive neurostimulation (RNS) devices. Instead of relying on raw EEG signals, the model utilizes structured features such as interictal epileptiform activity (IEA) and long episode events detected by RNS systems. To interpret these biomarkers, the model applies a convolutional-based patch embedding method followed by multi-head self-attention layers and squeeze-and-excitation modules, enabling it to model both short-term neural fluctuations and multi-day seizure cycles. The team validated their approach using data from five patients across prediction windows ranging from one to fourteen days. SeizureFormer achieved a mean ROC AUC of 79.44 % and PR AUC of 76.29 %, outperforming traditional statistical techniques and various deep learning baselines in terms of generalizability and resilience to class imbalance (Feng et al., 2025).

Tazaki et al. (2025) proposed a domain-adversarial CNN–BiLSTM framework tailored for inter-patient generalization in seizure detection. The CNN component is optimized via domain adversarial training to extract patient-invariant representations, while a BiLSTM captures temporal seizure evolution patterns. Tests on EEG data from 20 patients with focal epilepsy showed significant improvements in cross-patient performance compared to non-adversarial models, demonstrating strong robustness to individual variability (Tazaki et al., 2025).

Agaronyan et al. (2025) utilized a graph neural network (GNN) approach on stereo EEG (sEEG) data to predict seizure freedom outcomes in refractory epilepsy patients. Their architecture integrates multi-scale attention mechanisms to capture local and global connectivity patterns, enabling identification of seizure onset zones. Evaluated on pediatric sEEG recordings from 15 patients, the model achieved 92.4% accuracy and identified critical regions such as the anterior cingulate, with findings correlating well with clinical annotations (Agaronyan et al., 2025).

Table 4.1: Summary of the reviewed studies

Study	Model	Key Techniques	Dataset(s)	Performance
Dutta et al. (2024)	Multi-head Attention + BiLSTM	Log Softmax, Temporal + Long-range Features	CHB-MIT	Accuracy: 96.2%, Sensitivity: 96.5%, Specificity: 97.04%, F1-score: 96.6%
Xia et al. (2024)	Hybrid LSTM-Transformer	STFT, Time–Frequency Features	CHB-MIT	Sensitivity: 99.75%, AUC: 99.39%, FPR: 0/h
Tsiouris et al. (2018)	LSTM	Time–Frequency, Graph Features	CHB-MIT	Sensitivity: High, FPR: 0.11–0.02/h
Hussein et al. (2022)	Multi-channel Vision Transformer (MViT)	Scalograms, Patch Embedding	3 EEG datasets	Sensitivity: 99.80% (surface), 90.28–91.15% (invasive)
Li et al. (2022)	Transformer-Guided CNN (TGCNN)	STFT, Alternating CNN + Transformer	CHB-MIT, Kaggle	CHB-MIT: Sensitivity: 91.5%, FPR: 0.145/h, AUC: 93.5%; Kaggle: Sensitivity: 82.2%, FPR: 0.06/h, AUC: 83.5%
Kode et al. (2024)	1D-CNN + ML Ensemble	Raw EEG, Time-Series	UCI Epileptic Seizure Dataset	Accuracy: 99%
Aluvalu et al. (2024)	MMFCNN (IoT + DL)	Cross-Attention, Multimodal Fusion	Smart EEG	Accuracy: 99.2%
Sun et al. (2024)	Conv + Transformer	Raw Long-Term iEEG, Attention across Channels	SWEC-ETHZ, TJU-HH	Sensitivity: 97.5–98.1%, FDR: 0.06–0.22/h
Qi et al. (2024)	Sel-JPM-ViT	Channel Selection, Time–Frequency Analysis	BCH-MIT	Accuracy: 93.65%, Sensitivity: 94.7%, Specificity: 92.78%

Chiranjeevi et al. (2024)	Ensemble RNN + Time Series Transformer	Raw EEG, Dual Feature Extractors	CHB-MIT, Bonn	Accuracy: 94.31%, F-score: 91.93%
Wei et al. (2019)	LRCN (CNN + LSTM)	2D Image Conversion, Temporal Learning	EEG	Accuracy: 93.4%, Sensitivity: 91.88%, Specificity: 86.13%, FPR: 0.04/h
Wu et al. (2025)	SeizureTransformer	CNN + Transformer Encoder, Sequence Decoder	CHB-MIT + Others	Challenge Winner, High accuracy, Low latency
Feng et al. (2025)	SeizureFormer	CNN Embedding + Transformer + Squeeze–Excite	RNS/IEA (5 patients)	ROC AUC: 79.44%, PR AUC: 76.29%
Tazaki et al. (2025)	CNN + BiLSTM (Domain-Adversarial)	Inter-patient Generalization	20 focal epilepsy patients	High robustness to variability, Cross-patient improvement
Agaronyan et al. (2025)	Graph Neural Network (GNN)	Multi-scale Attention, Connectivity Mapping	Pediatric sEEG (15 patients)	Accuracy: 92.4%, Identified SOZs

4.3. Research Gap and Motivation

Despite significant advancements in seizure detection using deep learning models applied to EEG data, several important research gaps persist that limit both the performance and practical applicability of current approaches.

First, a considerable portion of existing studies rely on unimodal input representations. Specifically, many models focus primarily on time–frequency representations, such as spectrograms or wavelet transforms (i.e., the spectral domain), while underutilizing the raw EEG signals in the time domain. This separation of modalities means that most models fail to benefit from the unique and complementary information that exists in both domains. Seizure-related brain activity is known to exhibit both temporal characteristics such as sharp spikes or rhythmic discharges over time and spectral patterns such as changes in specific frequency bands. Ignoring either domain limits the model’s ability to detect seizures accurately and reliably across different contexts and patient conditions.

Second, although Transformer-based architectures have recently become popular in biomedical signal processing, most implementations in EEG-based seizure detection do not incorporate both temporal and spectral features in a unified model. This results in a missed opportunity to leverage the full representational power of Transformers in capturing inter-domain interactions that could enhance seizure recognition performance.

Third, EEG signals in clinical settings are inherently noisy and prone to various artifacts. They also exhibit significant inter-patient variability due to physiological differences, electrode placement, and diverse neurological conditions. The reviewd models often struggle with these real-world imperfections, performing well in controlled environments but failing in practical deployments. Therefore, developing models that are robust to noise, patient variability, and signal imperfections is a critical, yet often overlooked.

Finally, although multimodal and hybrid deep learning models, which attempt to integrate multiple types of input features or network components, have shown encouraging results in recent studies, this area is still relatively underexplored. Few studies have designed architectures that explicitly learn from both time and frequency representations in a way that enables meaningful cross-modal interactions. Furthermore, the design of such fusion mechanisms remains an open research problem, particularly when considering EEG's multichannel structure and the dynamic relationships between brain regions.

Together, these gaps highlight the need for more comprehensive, robust, and clinically applicable models that go beyond accuracy metrics to incorporate modality fusion, noise tolerance, generalizability, and real-world usability in EEG-based seizure detection.

5. Research Methodology

This chapter outlines the methodological framework adopted in this study for EEG-based seizure detection and classification. It begins with a description of the data collection process and the dataset used in the experiments. Following that, the preprocessing steps applied to raw EEG signals are detailed, including filtering, segmentation, and transformation into spectrograms.

The chapter also introduces the tools and libraries employed for model development, followed by a comprehensive explanation of the proposed deep learning architecture. The model integrates both temporal and spectral modalities through LSTM, CNN, and Transformer components. Finally, the evaluation metrics used to assess model performance are described.

5.1. Data collection

The data used in this study is sourced from a Kaggle competition (Jin et al, 2024), and is provided under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License. The competition was hosted by the Sunstella Foundation, established in 2021 during the COVID-19 pandemic to support minority graduate students in technology by helping them overcome challenges and celebrate their achievements. Sunstella Foundation, in collaboration with Persyst, Jazz Pharmaceuticals, and the Clinical Data Animation Center (CDAC), aims to advance research focused on brain health. The dataset provided by these organizations has been instrumental in facilitating the analysis of EEG signals for the detection and localization of seizures in this research.

5.2. Dataset

The used dataset includes a total of 5,057 EEG recordings. The provided metadata allows us to extract the original EEG segments as annotated by a panel of experienced neurologists and clinical neurophysiologists specializing in EEG interpretation. To determine the final labels, a majority voting scheme was applied, assigning the class with the highest number of expert votes. Samples with tied votes between classes were excluded to ensure uniform expert agreement and label reliability.

The central 10 seconds of each sample were selected for labeling. Given a sampling rate of 200 Hz, this corresponds to 2,000 data points. To extract the central segment, we define the middle of the sequence, subtract 1,000 samples, and add 1,000 samples, ensuring that the selected window consists of the central 10 seconds (2,000 samples).

The metadata provided in the train.csv file contains key information used for model training. The column names are the names of the individual electrode locations for EEG leads. Each row corresponds to a 50-second EEG segment, with expert annotations focusing on the central 10 seconds. Table 5.1 summarizes the columns included in the metadata.

Table 5.1: Data Dictionary for the HMS - Harmful Brain Activity Classification Kaggle Competition Dataset (Jin et al., 2024)

Column Name	Description
eeg_id	Unique identifier for the full EEG recording.
eeg_sub_id	Identifier for the specific 50-second segment associated with the labels.
eeg_label_offset_seconds	Time offset from the start of the EEG recording to the labeled segment.
label_id	Unique identifier for the label set.
patient_id	Identifier for the patient who provided the EEG data.
expert_consensus	Final consensus label assigned by the expert annotators.
[seizure/lpd]_vote	Number of expert annotators who voted for a particular class (e.g., seizure or LPD).

5.3. Data Pre-processing

Data preprocessing is a critical step in the development of a model as it ensures that the input data is cleaned, structured, and transformed for optimal performance. The preprocessing phase focuses on preparing EEG signals and corresponding spectrogram data for input into the multimodal model. So, this section covers key aspects of data preprocessing, including noise reduction, filtering, transformation, and data augmentation.

5.3.1. Noise Reduction and Filtering

EEG signals are highly susceptible to noise and artifacts, originating from muscle activity, eye movements, electrode drift, and environmental interference. Without adequate preprocessing, these contaminants can obscure clinically relevant neural activity, leading to reduced model performance. Therefore, effective filtering is a critical preprocessing step to isolate brainwave activity and enhance the quality of both time-domain and frequency-domain representations.

In this study, a bandpass filter was applied with the following parameters: a low cut-off frequency of 0.5 Hz, a high cut-off frequency of 40 Hz, a sampling rate of 200 Hz, and a filter order of 4. This configuration is widely used in EEG preprocessing pipelines due to its ability to preserve essential neural oscillations while removing irrelevant signal components (Sanei & Chambers, 2013; Nunez & Srinivasan, 2006).

The lower bound of 0.5 Hz effectively removes slow baseline drifts and movement-related artifacts without attenuating clinically important slow-wave activity such as delta oscillations, which are particularly relevant in seizure detection and sleep analysis (Luck, 2014). The upper bound of 40 Hz is sufficient for eliminating high-frequency noise, including muscle artifacts and ambient electrical interference, which are not reliably captured by systems with a modest sampling rate of 200 Hz (Fatourechi et al., 2007). A fourth-order filter was selected to achieve a balance between frequency selectivity and temporal precision. Higher-order filters may introduce excessive phase distortion, while lower-order filters may fail to sufficiently attenuate noise. This choice minimizes phase shifts while maintaining the signal's temporal structure, which is crucial for accurately capturing seizure-related transients (Widmann, Schröger, & Maess, 2015).

Since the original dataset had not undergone artifact removal or noise filtering, this step was necessary to ensure the signal quality required for reliable model training and accurate seizure classification.

The bandpass filter is designed using the transfer function of a Butterworth filter, presented in Equation 5.1 as follows:

$$H(s) = \frac{1}{\sqrt{1 + \left(\frac{f_c}{f}\right)^{2n}}} \tag{5.1}$$

Where f_c is the cutoff frequency (either low-cut or high-cut), f is the input signal frequency, and n is the filter order.

5.3.2. Spectrogram Transformation

To capture the frequency information of EEG signals, each segment is transformed into a spectrogram using STFT, which provides a two-dimensional time–frequency image-like representation of the signal. This representation is particularly valuable in EEG analysis, as seizures are often characterized by transient, non-stationary neural oscillations that are not easily observable in the raw time domain (Acharya et al., 2018; Boonyakitanont et al., 2020).

The use of spectrograms allows the model to preserve the temporal evolution of spectral content, which is crucial for identifying pathological patterns such as spike-wave discharges, rhythmic slowing, or frequency shifts associated with ictal events. These patterns are especially important for distinguishing different seizure types and stages.

While several alternative time–frequency transforms exist, such as the Continuous Wavelet Transform (CWT), STFT-based spectrograms were chosen in this study due to their favorable balance between temporal and frequency resolution, interpretability, and computational efficiency (Craik et al., 2019). Unlike CWT, which requires more complex parameter tuning and is less efficient for large datasets, STFT provides a fixed time-frequency trade-off that aligns well with the objectives of this work.

For STFT computation, we set Sampling rate, Segment length (n_{perseg}), and $Overlap(n_{\text{overlap}})$, 200 Hz, 64 and $n_{\text{perseg}} \,//\, 2 \;=\; 32$, respectively. The STFT is computed using Equation 5.2:

$$X(m,k) = \sum_{n=0}^{N-1} x(n)w(n-mR)e^{-j2\pi kn/N} \tag{5.2}$$

Where $X(m,k)$ is the STFT output at time index m and frequency index k, x(n) is the input signal, w(n) is the window function, N is the FFT size, R is the hop size ($R = n_{\text{perseg}} - n_{\text{overlap}}$), j is the imaginary unit.

In Python, we implemented STFT using the scipy library, specifically its signal processing submodule (scipy.signal). The scipy library is a widely used open-source toolkit that provides efficient algorithms and functions for scientific and technical computing, including signal processing, optimization, and statistics (Virtanen et al., 2020).

5.3.3. Normalization and Standardization

Before feeding the data into the model, preprocessing steps such as normalization or standardization are essential to ensure numerical stability and effective learning. Normalization typically rescales the data to a fixed range (e.g., [0, 1]), while standardization transforms the data to have a mean of zero and a standard deviation of one. These transformations help prevent the model from being biased by input features with larger absolute values and ensure that all features contribute equally during training.

For multichannel EEG time-series data, normalization or standardization is typically performed per channel across time steps, allowing each channel to preserve its temporal dynamics while ensuring consistent scaling. Similarly, for spectrogram inputs, these transformations are applied per spectrogram or frequency band, depending on the network architecture. This preprocessing step is particularly important in deep learning models, where scale-sensitive operations such as convolution and attention can be significantly affected by unscaled input distributions.

5.3.4. Data Splitting

This step is fundamental in machine learning tasks, used to evaluate how well a model generalizes to unseen data. It involves splitting the dataset into separate subsets for training and validation, ensuring that performance metrics reflect real-world applicability.

In this study, the dataset was divided into training (80%) and validation (20%) sets, a commonly used split ratio that balances the need for sufficient training data with the requirement for meaningful validation. This division supports both the learning of robust feature representations and reliable performance evaluation.

Due to the multimodal nature of the data, comprising raw EEG signals and their corresponding spectrograms, special attention was given to maintaining pairwise consistency during the split. Each EEG segment and its associated spectrogram were kept together in the same subset to preserve data integrity across modalities. Moreover, stratified sampling was applied based on the two classes to ensure a balanced class distribution, which is essential for preventing model bias and ensuring valid comparisons.

This carefully designed preprocessing and splitting strategy ensures that the data fed into the model supports both accurate training and fair evaluation, ultimately improving the model's ability to classify and localize seizure activity from EEG signals.

5.4. Tools and Libraries

This study was implemented using Python as the main programming language due to its widespread use in machine learning and signal processing research, as well as its extensive ecosystem of scientific libraries. Several key Python libraries were employed to support various stages of the EEG-based deep learning pipeline, including data handling, preprocessing, model development, hyperparameter optimization, and visualization. The main libraries used in this thesis include PyTorch, Pandas, NumPy, SciPy, Librosa, Optuna, Matplotlib, and Seaborn. Table 5.2 is a brief description of the purpose and role of each library.

These tools collectively enable a comprehensive approach to processing, transforming, and analysing EEG signals for seizure detection, integrating both time-domain and frequency-domain information for improved classification.

Table 5.2: Software Libraries Used in This Study and Their Functional Roles

Libraries	Description
PyTorch	PyTorch served as the core deep learning framework for building, training, and evaluating neural network models. Its dynamic computation graph and modular structure make it especially suitable for developing advanced architectures such as LSTM, CNN, and Transformer models tailored to EEG data (Paszke et al., 2019).
Pandas	Essential for data management tasks, including reading and processing large EEG datasets stored in parquet and csv formats. It also facilitates easy access to metadata, such as eeg_id and spectrogram_id, which are necessary for model training.

NumPy	Used for efficient numerical computations, particularly for handling multi-dimensional EEG data arrays and preparing them for model input.
SciPy	Plays a key role in signal processing and filtering. It includes functions for designing filters, such as low-pass, high-pass, and band-pass filters, which are used to remove noise and isolate meaningful frequency bands in EEG data. These filters are crucial for preprocessing EEG signals and reducing artifacts.
Librosa	This library, along with scipy.signal, was used to transform time-domain EEG signals into spectrograms STFT and Mel-spectrograms. These time-frequency representations are crucial for extracting features that capture seizure-related dynamics and are particularly well-suited for CNN-based classification
Optuna	Used for hyperparameter tuning, allowing the automation of parameter selection for better model performance. Optuna helps optimize various parameters, including LSTM hidden sizes, CNN kernel sizes, and Transformer attention heads, enhancing classification results.
Matplotlib and Seaborn	These libraries provide visualization capabilities for data exploration, model performance evaluation, and hyperparameter optimization results, helping to illustrate model accuracy across different classes and refine model configuration.

5.5. Proposed Models

The main proposed model is designed to integrate both temporal and spectral characteristics of brain activity by processing raw EEG signals and their corresponding spectrogram representations through two parallel branches. Each branch is optimized to capture complementary features from the EEG data, temporal dynamics and frequency-domain patterns, before fusing the outputs for final classification. For comparison, baseline models processing each modality independently are also developed and discussed in the following.

5.5.1. LSTM-Transformer Model

The current unimodal model processes raw EEG time-series data. Each EEG sample consists of 19 channels, with each channel representing a time-series signal sampled at 200 Hz. To capture the intra-channel temporal dynamics such as rhythmic discharges, transient spikes, and gradual signal evolution linked to seizure onset and offset, LSTM networks are employed.

LSTMs are well-suited for modeling time-dependent patterns due to their gated architecture, which helps retain relevant information over long sequences and mitigates vanishing gradient issues common in traditional RNNs (Hochreiter & Schmidhuber, 1997). However, standard LSTMs process each channel independently and lack the ability to model cross-channel dependencies, which are crucial in EEG analysis, as seizures often manifest across spatially distributed regions of the brain.

To address this, a Transformer module is introduced after the LSTM layers. The Transformer's self-attention mechanism enables the model to capture not only long-range temporal dependencies but also cross-channel interactions by attending to relevant features across all channels simultaneously. This ability is particularly important in multichannel EEG data where pathological patterns may emerge from coordinated activity across brain regions. Thus, the combined LSTM-Transformer architecture allows the model to leverage both local sequential information within each channel and global interdependencies across channels, improving its capacity to recognize complex seizure signatures.

Finally, the output from the Transformer is passed through a fully connected classification layer that maps the extracted features to classes. The architecture of this model is shown in Figure 5.2.

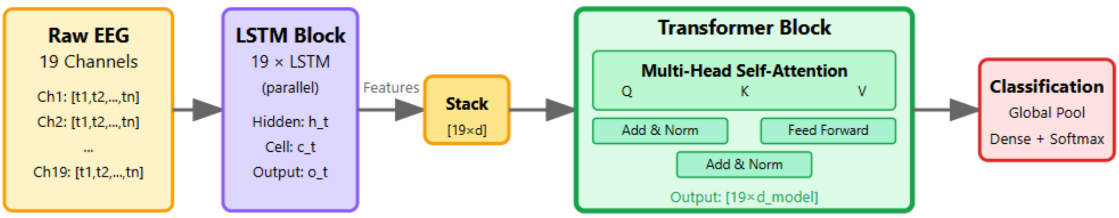

Figure 5.1: LSTM-Transformer Model

5.5.2. CNN-Transformer Model

This unimodal model utilizes spectrogram representations of EEG signals, obtained by applying the STFT to each EEG channel. These spectrograms provide a joint time–frequency representation, enabling the analysis of how frequency components evolve over time which is an essential characteristic for distinguishing various seizure types.

The resulting spectrograms are processed using a CNN, which excels at learning spatially localized features. This allows the model to detect clinically relevant patterns such as epileptiform spikes, bursts, and rhythmic discharges, which often manifest in the time–frequency domain. However, CNNs are inherently limited in their receptive field and thus struggle to capture long-range dependencies or global contextual information within the spectrogram.

To address this limitation, a Transformer module is integrated after the CNN layers. Leveraging its self-attention mechanism, the Transformer enhances the model's capacity to model global contextual relationships within the spectral domain as well as cross-channel interactions. This is particularly important given that seizure activity often involves temporally extended and spatially distributed features across different EEG channels.

Finally, the output of the Transformer is passed through a fully connected classification layer, which maps the learned high-level representations to the target seizure classes. The overall architecture of this model is illustrated in Figure 5.3.

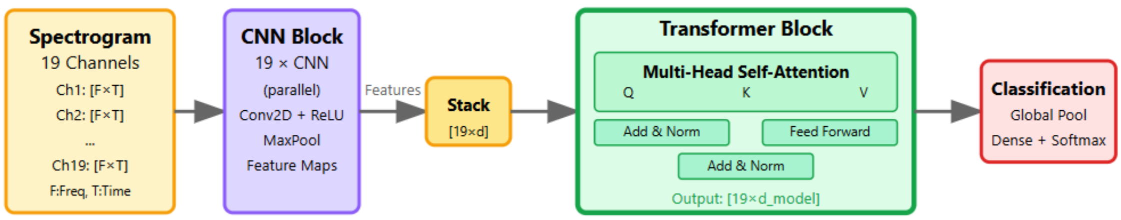

Figure 5.2: CNN-Transformer Model

5.5.3. Multimodal Fusion Model

To leverage both the temporal dynamics and spectral characteristics of EEG signals simultaneously, this study proposes a multimodal deep learning architecture that integrates features extracted from raw time-series data and their corresponding time–frequency image-like representations. This dual-branch design is motivated by the fact that seizures can manifest as both temporal patterns such as rhythmic discharges or transient spikes and spectral shifts, including power changes in specific frequency bands. By combining both modalities, the model aims to exploit their complementary strengths, improving classification performance, robustness to noise, and generalizability across seizure types.

In the temporal branch, each raw EEG segment consisting of 19 channels is processed using LSTM layers, which are well-suited for learning sequential dependencies and capturing intra-channel temporal dynamics. This enables the model to detect subtle changes over time, such as pre-ictal buildup or ictal waveform evolution, that may be lost in purely frequency-based approaches.

In parallel, the spectral branch transforms the same EEG segments into spectrograms using STFT, yielding a joint time–frequency representation that preserves the evolving spectral structure of brain activity. These spectrograms are then fed into a CNN, which is adept at identifying localized spatial patterns within the frequency domain.

The feature vectors obtained from both branches are flattened and concatenated, creating a unified embedding that encodes both temporal and spectral information. This fusion enables the model to benefit from both fine-grained sequential features and spatially structured spectral features, which are often indicative of seizure onset and propagation.

To model higher-order relationships within this joint representation, a Transformer module is applied post-fusion. The Transformer's self-attention mechanism allows the model to learn global contextual dependencies, including both cross-channel relationships and inter-modal interactions. This is particularly important for EEG-based seizure detection, where complex patterns may span multiple channels and emerge only when both time and frequency information are considered jointly.

Finally, the output of the Transformer is passed through a fully connected classification layer, which maps the learned representations to one of the target seizure categories. The complete architecture of the proposed multimodal model is illustrated in Figure 5.4.

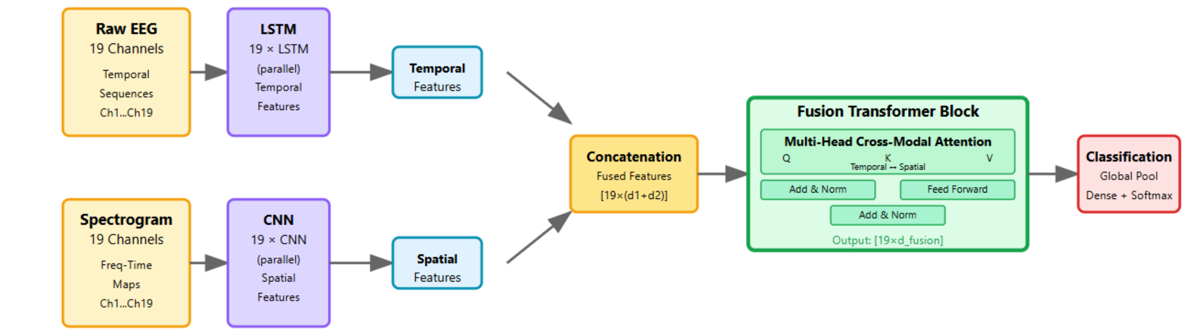

Figure 5.3: Multimodal Fusion Model

5.6. Performance Evaluation Criteria

To assess the effectiveness of the models in seizure detection, multiple evaluation metrics were considered, including Precision, Recall, F1-score, and Accuracy. These metrics provide a comprehensive understanding of the model's classification performance, particularly in handling imbalanced data and distinguishing between seizure and non-seizure states.

Precision: also known as Positive Predictive Value (PPV), it measures the proportion of correctly predicted positive cases out of all predicted positive cases. It is formally defined in Formula 5.3. A high precision score indicates fewer false positives, which is critical in medical applications where misclassification can have severe consequences.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.3}$$

TP (True Positives) = correctly predicted seizure instances,

FP (False Positives) = incorrectly predicted seizure instances.

Recall (Sensitivity or True Positive Rate): this criterion measures the proportion of actual positive cases that were correctly identified by the model, defined in Formula 5.4.

(5.4)

A high recall score ensures that most seizure events are detected, minimizing false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where FN (False Negatives) = actual seizures that were incorrectly classified as non-seizures.

F1-Score: it is a composite evaluation metric that balances the trade-off between precision and recall, especially in scenarios where class distributions are uneven. It is computed as the harmonic mean of precision and recall, placing more emphasis on their agreement. Unlike the arithmetic mean, the harmonic mean penalizes extreme values, ensuring that a high F1-Score can only be achieved when both precision (low false positives) and recall (low false negatives) are reasonably high. This makes it particularly suitable for classification tasks involving class imbalance, where relying solely on accuracy can be misleading. The F1-Score is formally presented in Formula 5.4.

$$F1\text{-}Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

A higher F1-score indicates better overall model performance, particularly when precision and recall need to be balanced.

Accuracy measures the overall correctness of the model by evaluating the proportion of correctly classified instances across all classes. While precision, recall, and F1-score are reported separately for each class (e.g., Seizure and LPD) to reflect class-specific performance, accuracy is a global metric that provides the overall correctness across all predictions. It is calculated as the ratio of all correct predictions (e.g., Seizure and LPD) to the total number of predictions. Therefore, accuracy is not defined on a per-class basis, and only one accuracy score is reported per model. Additionally, it may not always be reliable when dealing with imbalanced datasets, where precision and recall become more meaningful. The formal definition of accuracy is given in Formula 5.5.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.5}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

6. Results

This chapter presents the study's findings, detailing the performance of the proposed multimodal deep learning model for EEG-based seizure and LPD classification. It begins with a description of the experimental setup. It then details the results of three architectures: the LSTM-Transformer, which models temporal dependencies in raw EEG signals; the CNN-Transformer, which extracts and analyses spectral features from spectrograms; and finally, the Fusion Model, which integrates both temporal and spectral representations to leverage the strengths of both modalities. Finally, section 4.5 provides a comparative analysis of the three models, highlighting their respective accuracies and overall classification performance.

6.1. Experimental Setup and Configuration

All three models were trained and validated using an 80/20 split of the dataset over 200 epochs, providing sufficient iterations for convergence and performance monitoring. A consistent batch size of 32 was used across all experiments for both training and validation phases to ensure stable gradient updates and efficient computation. The AdamW optimizer was also selected for optimization. AdamW is an adaptive learning rate optimization algorithm that specifically addresses the issue of weight decay decoupling from the learning rate, often leading to better generalization performance than standard Adam. Training commenced with an initial learning rate of 0.0001. To further enhance training stability and performance, a learning rate scheduling strategy was implemented using LambdaLR, which dynamically adjusts the learning rate throughout the training process based on a predefined function. Furthermore, to effectively mitigate the risk of overfitting, an early stopping criterion was employed with a patience of 15 epochs. This meant that training would cease if the validation performance did not improve for 15 consecutive epochs, saving the model weights from the best performing epoch.

6.2. LSTM-Transformer Model

To analyse raw EEG signals, we employed a LSTM-Transformer architecture as one of the two processing modalities in our multimodal model. The EEG data was segmented using a sliding window approach with a window size of 200 samples (1 second at a 200 Hz sampling rate) and a step size of 50 samples. Each segment served as a sequential input to a LSTM network, which was configured with a hidden size of 16, operating in unidirectional mode to align with PyTorch's batch-wise processing. This LSTM layer was responsible for capturing the temporal dependencies in the EEG signals while maintaining computational efficiency.

The sequential output of the LSTM was passed to a Transformer encoder block to model more complex temporal dependencies between channels. The Transformer was configured with an embedding dimension of 32 (d_{model}=32), 8 attention heads to allow parallel attention to different representation subspaces, and a dropout rate of 0.5 to mitigate overfitting. This setup enabled the model to enhance its temporal awareness through self-attention mechanisms, allowing it to weigh the importance of different time steps dynamically.

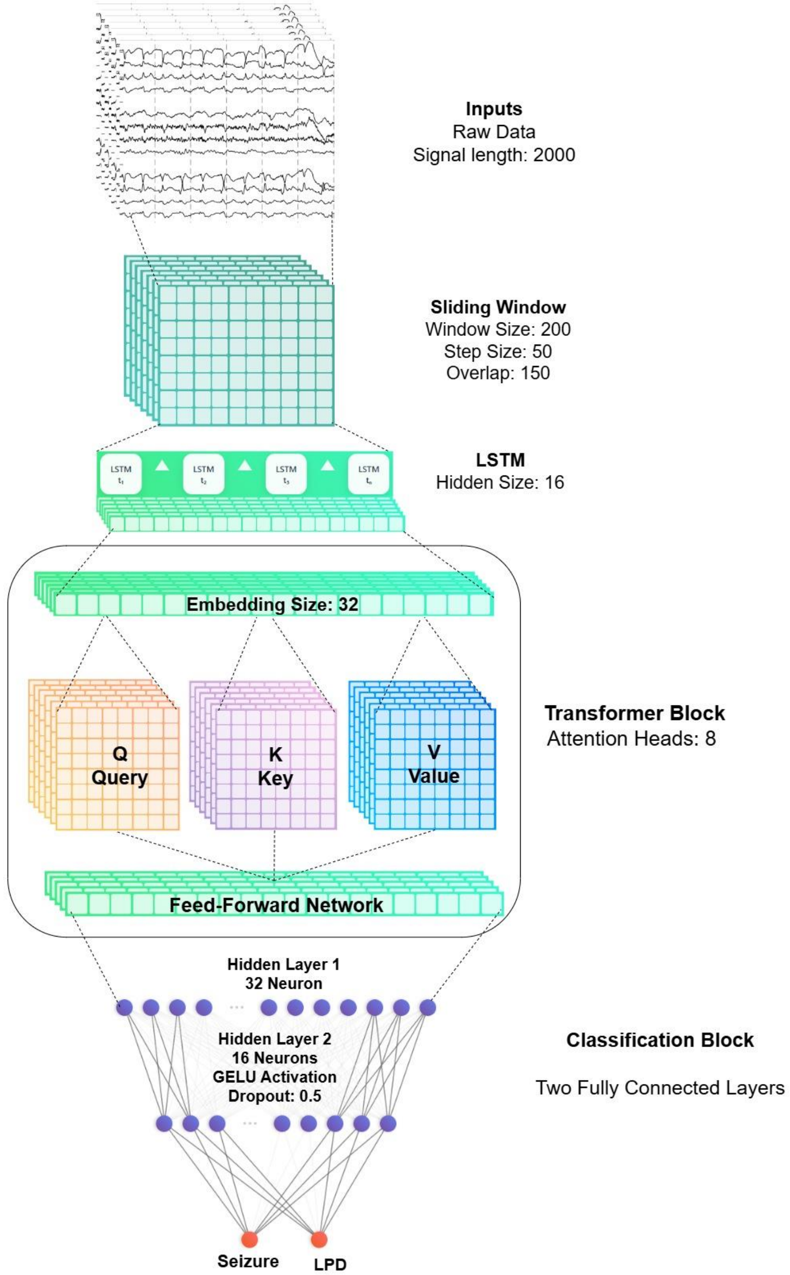

Figure 6.1: The schematic representation of the LSTM-Transformer architecture with key parameters

Following the Transformer, the extracted features were flattened and passed through a classification block consisting of two fully connected layers. The first linear layer reduced the dimension from $d_{\text{model}} = 32$ to dimension $d_{\text{model}}/2 = 16$, followed by a GELU (Gaussian Error Linear Unit) activation function and dropout for regularization. The final linear layer mapped the output to the number of classes, providing the prediction scores. This end-to-end pipeline allowed the model to learn hierarchical representations of raw EEG data and contributed significantly to the overall performance in detecting seizure-related patterns. Following this, the training and validation process was conducted for 200 epochs. The schematic representation of the architecture is illustrated in Figure 6.1.

Table 6.1 presents the performance metrics of the LSTM-Transformer model for classifying two distinct classes: Class 0 (LPD) and Class 1 (Seizure). The model's effectiveness is evaluated across several key metrics: Precision, Recall, f1-score, and Accuracy. The model demonstrates balanced performance across both classes, achieving an overall accuracy of 79.15%. This suggests that the model is generally effective in correctly identifying both LPD and Seizure conditions.

For Class 0 (LPD), the model exhibits a precision of 0.7823, indicating that approximately 78.23% of instances predicted as LPD are indeed correct. The recall for LPD is slightly higher at 0.8056, meaning the model successfully identifies about 80.56% of all actual LPD cases. The f1-score, which is the harmonic mean of precision and recall, stands at 0.7937, reflecting a good balance between these two metrics for the LPD class.

In contrast, for Class 1 (Seizure), the model's precision is 0.8012, indicating that a high proportion of its Seizure predictions are accurate. The recall for Seizure is 0.7776, suggesting that 77.76% of all true Seizure cases are correctly identified by the model. The f1-score for the Seizure class is 0.7892, which is very close to that of the LPD class, highlighting consistent performance across both categories.

It can be concluded that the model shows slightly better recall for LPD (0.8056 vs. 0.7776) and slightly better precision for Seizure (0.8012 vs. 0.7823). However, the f1-scores for both classes are remarkably similar (0.7937 for LPD and 0.7892 for Seizure), indicating that the LSTM-Transformer model maintains a robust and balanced predictive capability for identifying both LPD and Seizure conditions. This consistent performance across different evaluation metrics and classes underscores the model's reliability for the given classification task.

Table 6.1: The performance of LSTM-Transformer Model

Classes	Precision	Recall	f1-score	Accuracy
Class 0: LPD	0.7823	0.8056	0.7937	79.15
Class 1: Seizure	0.8012	0.7776	0.7892	

Figure 6.2 and Figure 6.3 illustrates the training and validation accuracy and loss curves, respectively. The best validation accuracy of 79.15% was achieved at epoch 33, with a learning rate of 0.000492 at that point. The accuracy curves indicate that the model was able to learn effectively during the early stages of training, with validation accuracy peaking at epoch 33. After this point, further training did not result in significant improvements, suggesting that the model had reached convergence, and prolonged training may have led to overfitting. Similarly, the training and validation loss curves show a sharp decrease during the initial epochs, indicating effective learning. The validation loss reaches its minimum around epoch 33, aligning with the peak in validation accuracy. Beyond this epoch, the validation loss plateaued while the training loss continues to decrease, suggesting the onset of overfitting. This divergence between training and validation loss is typical in deep learning and highlights the importance of early stopping or regularization strategies. Overall, the results suggest that the model successfully learned discriminative patterns from the EEG data and was able to generalize reasonably well to unseen samples. A more in-depth interpretation of these findings will be provided in the discussion chapter.

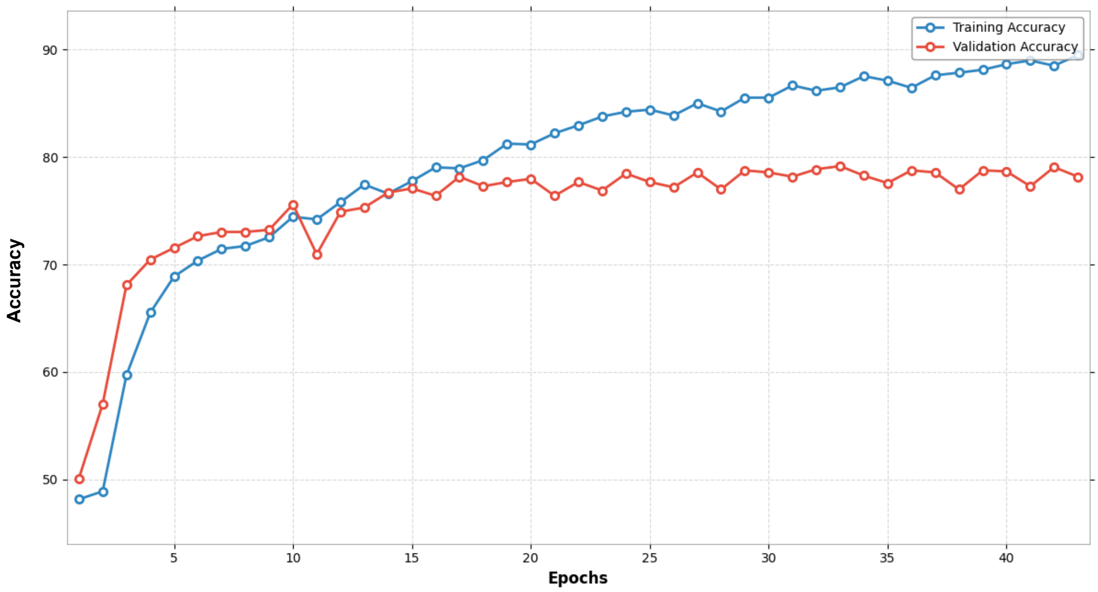

Figure 6.2: Training and validation accuracy of the LSTM-Transformer Model

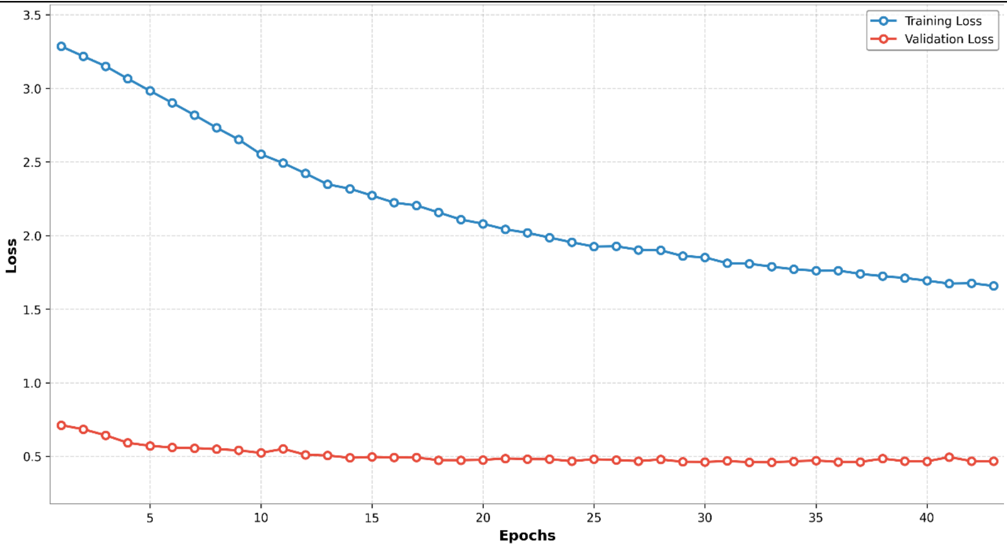

Figure 6.3: Training and validation Loss of the LSTM-Transformer Model

6.3. CNN-Transformer Model

To process the spectrogram representations of EEG data, CNN-Transformer model was developed and implemented. Figure 6.4 provides a summary of the Architecture block. Accordingly, spectrograms, being two-dimensional time–frequency representations with size 101×73, were treated as single-channel images (in_channels=1). The CNN block began with two convolutional layers: the first with 32 filters, a kernel size of 8, and padding of 3 to preserve spatial resolution, and the second with 64 filters, a kernel size of 6, and padding of 2 to extract deeper features. These convolutional operations enabled the network to capture localized time–frequency patterns important for seizure detection. To reduce the dimensionality and retain the most salient features, an Adaptive Average Pooling layer (AdaptiveAvgPool2d) was applied, yielding a fixed-size output regardless of input shape.

Subsequently, the pooled features were then flattened and projected into a latent space with an embedding dimension of 128 (d_{model}=128). This process serves to compress the spatially distributed features into a compact, fixed-length vector representation, which is then compatible with the sequential processing requirements of the Transformer. The chosen embedding dimension of d_{model}=128 provides a sufficient capacity to encode meaningful information from the pooled features. Then, this feature representation was passed through a Transformer encoder to model global dependencies within the spectrogram, with self-attention mechanisms allowing the model to dynamically attend to informative frequency–time regions.

Finally, the output of the Transformer was fed into a classification block consisting of three fully connected layers. These layers progressively reduced the dimensionality from d_{model} (128) to $d_{\text{model}}/2$ (64), then to $d_{\text{model}}/4$ (32), and ultimately to the number of target classes. Each layer was followed by non-linear activation functions (GELU) and dropout for regularization. This design allowed the CNN-Transformer model to learn rich hierarchical features from spectrograms and contributed significantly to the model's ability to discriminate between seizure and non-seizure activity.

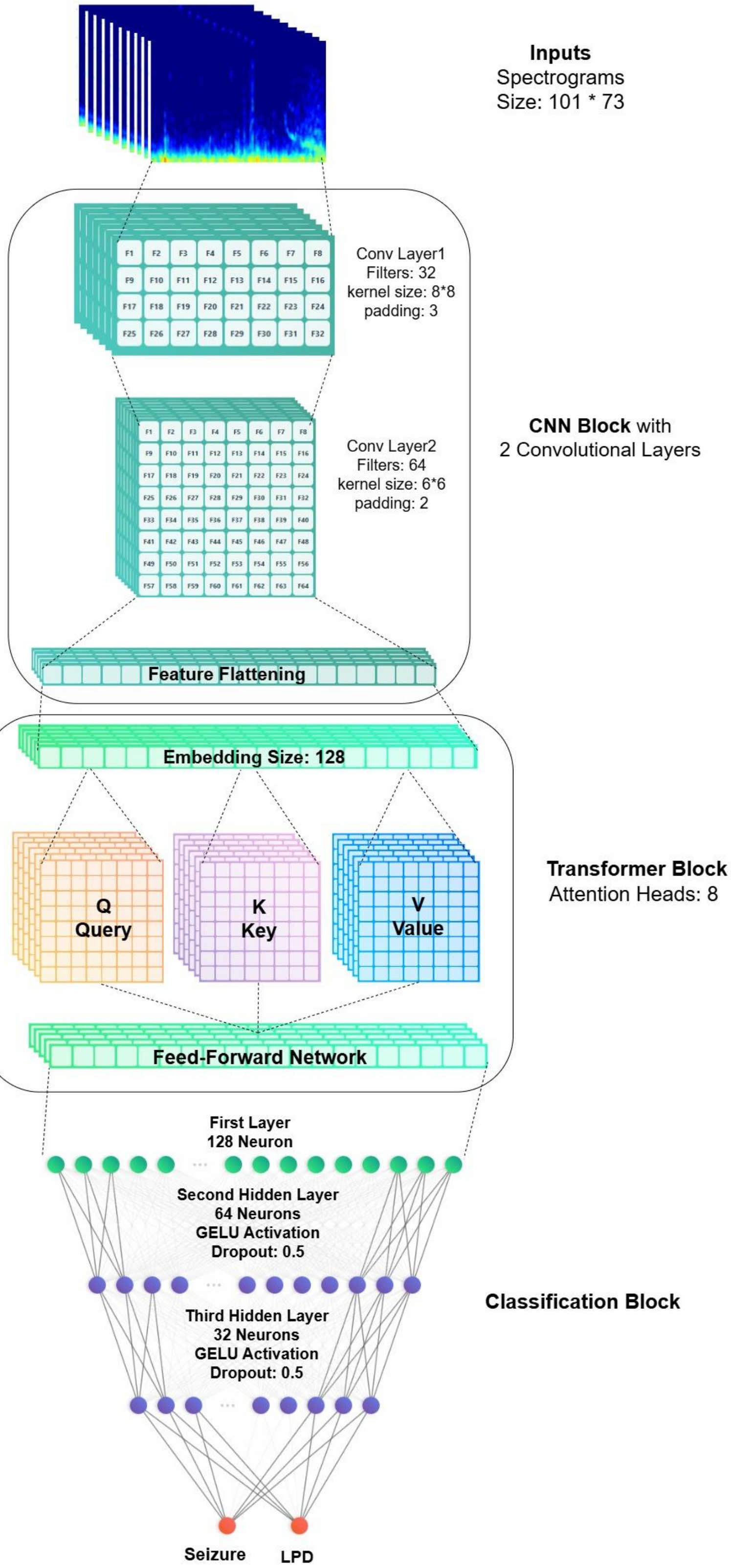

Figure 6.4: Schematic Representation of the CNN-Transformer Architecture with Key Parameters

Table 6.2 provides the performance metrics of the CNN-Transformer model for Class 0 (LPD) and Class 1 (Seizure). The model demonstrates exceptionally high and consistent performance across both classes, achieving an outstanding overall accuracy of 90.12%. This indicates a robust ability to correctly classify both LPD and Seizure with high confidence. For Class 0 (LPD), the model exhibits a precision of 0.9089, meaning that approximately 90.89% of instances predicted as LPD are genuinely LPD. The recall for LPD is also very high at 0.8909, indicating that the model successfully identifies about 89.09% of all actual LPD cases. The f1-score for LPD is 0.8998, reflecting an excellent balance between precision and recall, signifying a highly effective classification for this class. Similarly, for Class 1 (Seizure), the model's precision stands at 0.8938, demonstrating that a very high proportion of its Seizure predictions are accurate. The recall for Seizure is notably high at 0.9114, suggesting that the model is adept at identifying nearly 91.14% of all true Seizure cases. The f1-score for the Seizure class is 0.9025, further emphasizing the model's strong and well-balanced performance in detecting Seizures. When comparing the performance between the two classes, the CNN-Transformer model shows a slight advantage in precision for LPD (0.9089 vs. 0.8938) and a marginal lead in recall for Seizure (0.9114 vs. 0.8909). However, the f1-scores for both classes are remarkably close and exceptionally high (0.8998 for LPD and 0.9025 for Seizure). This near-perfect balance, coupled with the high overall accuracy, highlights the CNN-Transformer model's superior and reliable performance across both diagnostic categories, making it a highly promising tool for the classification of LPD and Seizure.

Table 6.2: The performance of CNN-Transformer model

Classes	Precision	Recall	f1-score	Accuracy
Class 0: LPD	0.9089	0.8909	0.8998	90.12
Class 1: Seizure	0.8938	0.9114	0.9025	

The training and validation curves are presented in Figures 6.5 and 6.6. it can be seen that the model exhibited consistent improvement during the initial training phase, with a notable rise in validation accuracy up to epoch 44, at which point it achieved its highest validation accuracy of 90.12%. The corresponding learning rate at this stage was 0.00047, indicating that the model had effectively converged. Beyond this point, the performance began to fluctuate, suggesting that the model had reached its optimal capacity and further training did not lead to meaningful improvements.

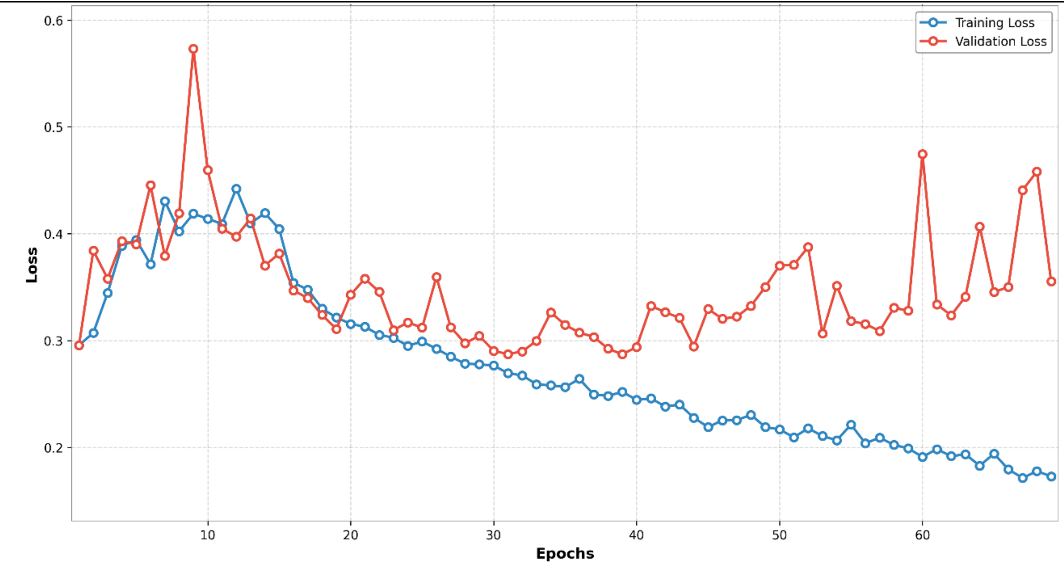

Figure 6.5: Training and validation accuracy of the CNN-Transformer Model

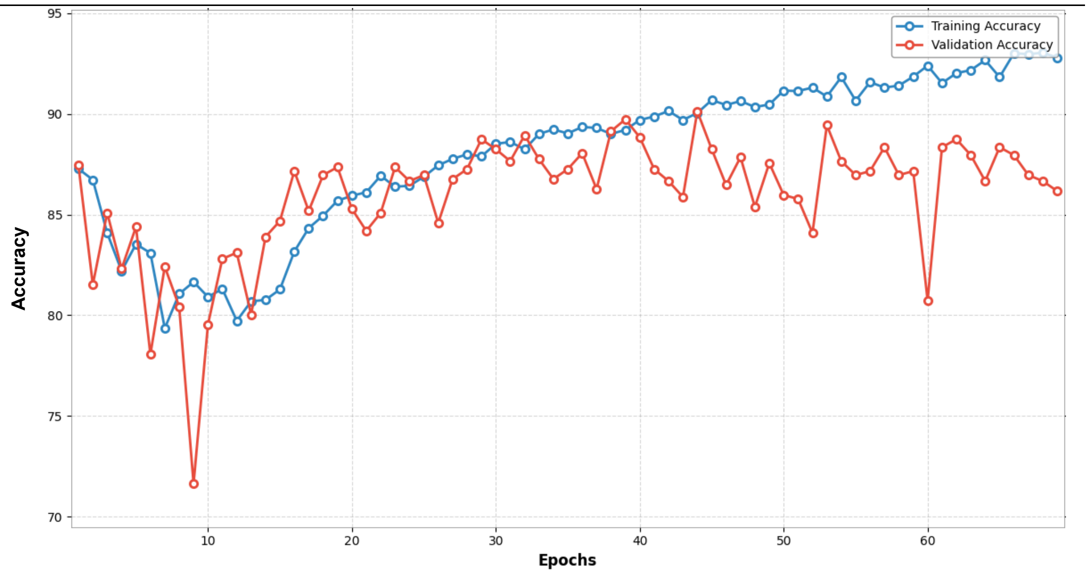

Figure 6.6: Training and validation accuracy of the CNN-Transformer Model

6.4. Fusion Model

To effectively capture both spectral and temporal characteristics of EEG signals, a Fusion Model was developed, integrating a CNN for spectrogram-based feature extraction and a LSTM for raw EEG signal processing. Figure 6.7 represents an overview of the architecture.

Table 6.3 provides a detailed breakdown of the architectural components used in the proposed fusion model, which integrates both temporal and spectral EEG information through a hybrid design involving CNN, LSTM, and Transformer blocks. The CNN block consists of two convolutional layers: the first layer applies 64 filters with a kernel size of 7 and padding of 3, followed by a second layer with 96 filters and a smaller kernel size of 5 with padding of 2. These layers are designed to extract localized spatial-frequency features from spectrogram representations of EEG signals. An adaptive average pooling layer is applied to reduce dimensionality while preserving the most informative activations. The LSTM block processes raw EEG time-series segments with an input size of 200 (corresponding to the windowed segment length). It utilizes a unidirectional LSTM with a hidden size of 32 to maintain compatibility with batch-based training. This recurrent component captures the temporal dynamics of EEG signals across time. The outputs from both the CNN and LSTM branches are fused and passed into a Transformer block to model complex cross-modal and cross-channel dependencies. The Transformer uses an embedding dimension of 128, 8 attention heads, and a dropout rate of 0.5 to enhance regularization and prevent overfitting. Finally, the classifier head consists of two linear layers. The first maps the Transformer output to a lower-dimensional space using a GELU activation and dropout for non-linearity and regularization. The second linear layer maps the reduced representation to the number of target classes for final prediction.

This architecture is designed to leverage the strengths of each component (CNN for spatial-frequency feature extraction, LSTM for temporal sequence modeling, and Transformer for capturing global contextual dependencies), thereby improving the model's overall capacity for seizure and LPD classification from EEG data.

Table 6.3: Summary of Fusion Model Architecture Components

Component	Layer / Parameter	Details
CNN Block	Conv Layer 1	64 filters, kernel size = 7, padding = 3
	Conv Layer 2	96 filters, kernel size = 5, padding = 2
	Pooling	Adaptive Average Pooling
LSTM Block	Input size	200 (windowed EEG segment)
	Hidden size	32
	Batch first	True
	Bidirectional	False (unidirectional)
Transformer	Embedding dimension d_{model}	128
	Attention heads	8
	Dropout rate	0.5
Classifier	Layer 1	Linear (d_{model}, $d_{\text{model}}/2$)→ GELU → Dropout
	Layer 2	Linear ($d_{\text{model}}/2$, num_classes)

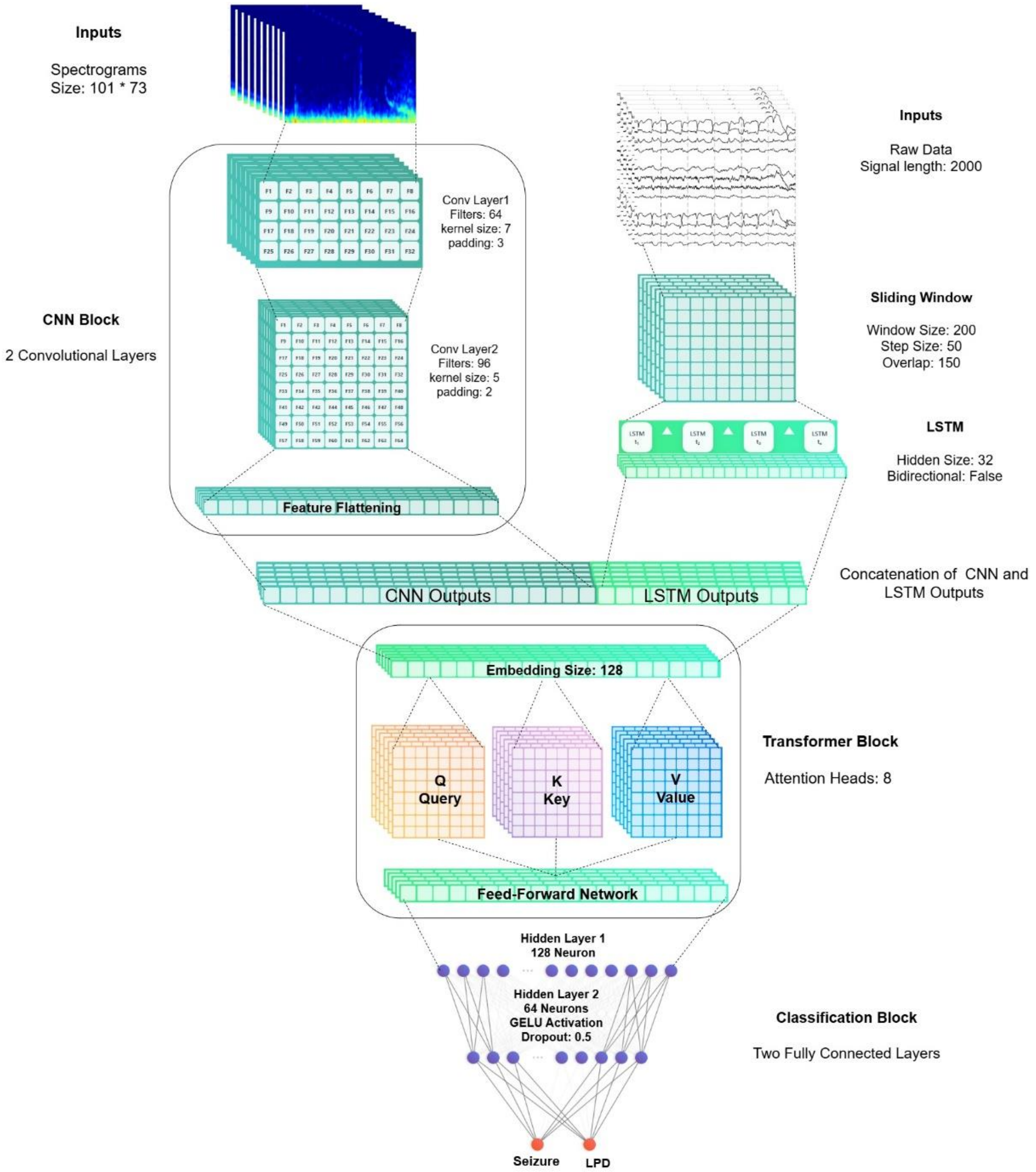

Figure 6.7: Schematic Representation of the Fusion Architecture with Key Parameters

Table 6.4 presents the performance metrics of the Fusion Model. The Model achieves an overall accuracy of 82.51%, indicating a generally strong capability in correctly classifying both LPD and Seizure conditions. For Class 0 (LPD), the Fusion Model demonstrates a precision of 0.7676, meaning that approximately 76.76% of instances predicted as LPD are accurate. Notably, the recall for LPD is exceptionally high at 0.9306, indicating that the model successfully identifies about 93.06% of all actual LPD cases. This suggests that the Fusion Model is particularly adept at minimizing false negatives for LPD. The f1-score for LPD is 0.8413, reflecting a good balance, albeit with a clear emphasis on recall. On the other hand, for Class 1 (Seizure), the model exhibits a very high precision of 0.9127, signifying that a large majority of its Seizure predictions are accurate. However, the recall for Seizure is considerably lower at 0.7205, meaning that only about 72.05% of all true Seizure cases are correctly identified by the model. This indicates a higher rate of false negatives for the Seizure class compared to LPD. The f1-score for the Seizure class is 0.8053, which is lower than that of the LPD class, largely due to the lower recall. In general, the Fusion Model shows a distinct imbalance in its performance characteristics. It prioritizes recall for LPD, effectively capturing a high proportion of LPD cases, even at the cost of some precision. Conversely, for Seizure, it prioritizes precision, making very accurate predictions when it does identify a Seizure, but missing a notable portion of actual Seizure cases. This characteristic suggests that while the Fusion Model is effective overall, its optimization appears to be skewed towards minimizing missed LPD diagnoses, potentially at the expense of recall for Seizure. The varying f1-scores between the classes further underscore these differences in performance emphasis.

Table 6.4: The performance of Fusion Model

Classes	Precision	Recall	f1-score	Accuracy
Class 0: LPD	0.7676	0.9306	0.8413	82.51
Class 1: Seizure	0.9127	0.7205	0.8053	

Figures 6.8 and 6.9 illustrate the training and validation accuracy and loss curves, respectively, recorded throughout the 200 training epochs. The model achieved its highest validation accuracy of 82.51% at epoch 12, with a corresponding learning rate of 0.000499. The learning curves indicate that the model converged early, with rapid improvement in performance within the first 15 epochs. After reaching peak validation accuracy at epoch 12, further training resulted in a gradual decline in validation performance, while training accuracy continued to increase. This divergence between training and validation trends suggests the onset of overfitting. Similarly, the validation loss reaches its minimum around epoch 12 and begins to rise thereafter, confirming that the model had learned the most generalizable features by that point.

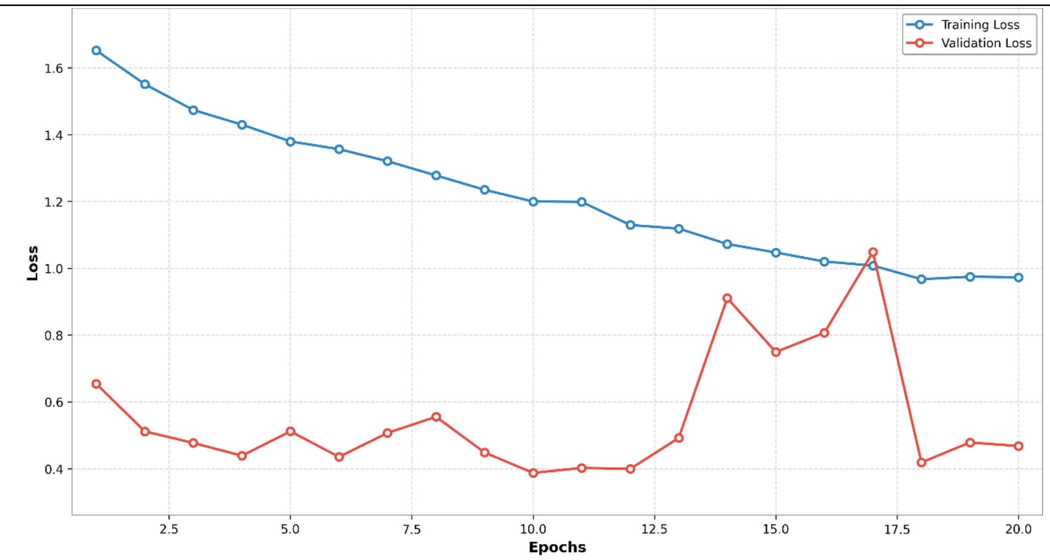

Figure 6.8: Training and validation Loss of the Fusion Model

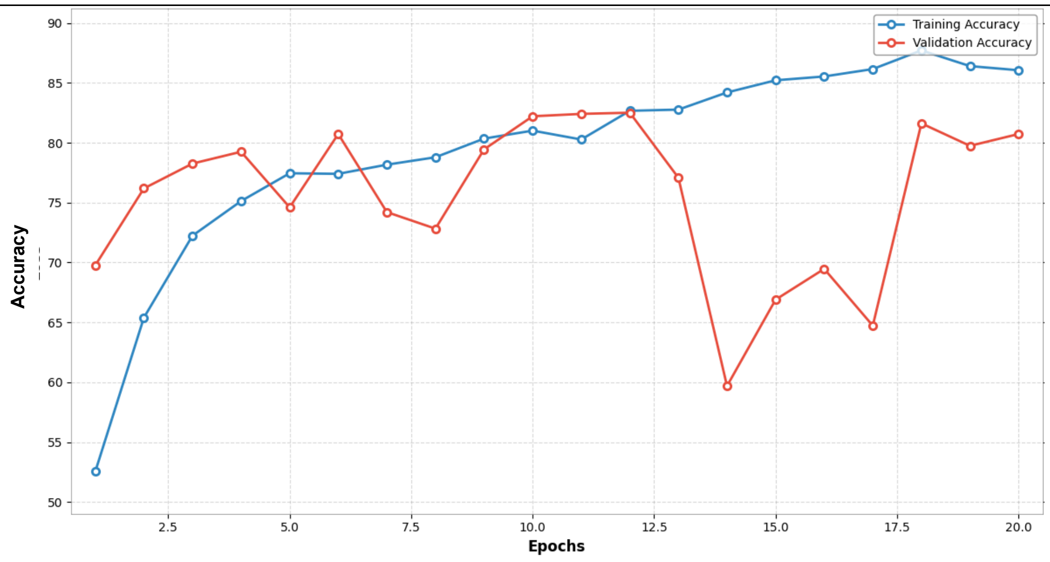

Figure 6.9: Training and validation Accuracy of the Fusion Model

6.5. Results Summary of Models

The performance of the three models, LSTM-Transformer, CNN-Transformer, and the Fusion Model, was evaluated based on their classification accuracy. The results are summarized in the Table 6.5.

According to the table, the CNN- Transformer model achieved the best performance with an accuracy of 90.12%, indicating that spectrogram-based representations, when processed with a convolutional network and Transformer, provide highly discriminative features for classification. The superior performance of CNN-Transformer suggests that frequency-domain features extracted from EEG spectrograms are more informative for seizure classification compared to raw time-domain EEG signals. The LSTM-

Transformer model obtained the lowest accuracy of 79.15%, highlighting the limitations of using only raw EEG signals in a sequential learning framework. While LSTM can capture temporal dependencies, EEG signals are inherently non-stationary and complex, making it challenging to extract robust patterns solely from the raw time series. Additionally, LSTM lacks the spatial feature extraction capability of CNN, which may explain its lower performance compared to CNN- Transformer.

Table 6.5: The performances of the models

Models	Accuracy	Recall	Precision	F1-score
LSTM-Transformer	79.15	0.7916	0.7918	0.7915
CNN-Transformer	90.12	0.9012	0.9014	0.9012
Fusion	85.51	0.8256	0.8402	0.8233

The Fusion Model, which combines CNN and LSTM features using Transformer, achieved an accuracy of 85.51%, outperforming LSTM-Transformer but falling short of CNN-Transformer. This result indicates that although integrating both spectrogram-based and raw EEG-based features provides additional information, the fusion approach did not surpass the CNN-Transformer model. One possible reason could be that the spectrogram features dominate the classification task, and adding raw EEG features may introduce redundant or less discriminative information.

7. Discussion

7.1. Overview

This chapter provides an in-depth discussion of the findings, their implications, the limitations of the study, and directions for future research. The primary focus of this research was to detect seizure from EEG signals using a multimodal deep learning approach that integrates LSTM for temporal feature extraction, CNN for spectrogram analysis, and Transformers for channels interdependency. The results have demonstrated the feasibility of using such an approach for classification of seizures and LPD. The study also developed two more deep learning architectures, including LSTM-Transformer, CNN-Transformer, and then compared their result with the proposed multimodal fusion model to find out which has better performance.

7.2. Conclusions

The models are compared based on accuracy, precision, recall, and F1-score. In addition, a broader comparison is made with findings from recent studies in the literature to contextualize the results and highlight the potential contributions and limitations of the current work. This comparative perspective not only underscores the strengths of the proposed architectures but also reveals important directions for future improvement in multimodal EEG classification systems.

Table 7.1: Comparative analysis of performance metrics across models

Model	Class	Precision	Recall	F1-Score	Accuracy (%)
LSTM-Transformer	LPD	0.78	0.81	0.79	79.15
	Seizure	0.80	0.78	0.79	
CNN-Transformer	LPD	0.91	0.89	0.90	90.12
	Seizure	0.89	0.91	0.90	
Fusion Model	LPD	0.77	0.93	0.84	82.51
	Seizure	0.91	0.72	0.81	

Table 7.1 presents a summary of the classification results, highlighting the superior performance of the CNN-Transformer architecture. This model achieved the highest accuracy of 90.12%, with F1-scores of 0.90 for LPD and 0.90 for Seizure detection. Its higher precision and recall metrics across both classes indicate that integrating spatial-frequency features (via CNN) with temporal modeling (Transformer) results in a more discriminative representation of EEG patterns. This strongly supports the hypothesis that frequency-domain representations, when coupled with convolutional and attention

mechanisms, provide highly discriminative features for seizure classification. Spectrograms encapsulate both temporal and spectral properties of EEG signals, and CNNs are particularly well-suited for extracting localized patterns within these two-dimensional representations. When enhanced by Transformers which capture long-range dependencies and contextual relationships, the result is robust.

In contrast, The LSTM-Transformer model achieved a balanced performance across both classes (LPD and Seizure), with an overall accuracy of 79.15%, f1-scores of 0.79 (LPD) and 0.79 (Seizure), and strong precision-recall trade-offs. This demonstrates that the architecture is capable of effectively capturing temporal dependencies in EEG data, particularly useful in distinguishing subtle differences between seizure-like and periodic discharges. However, it attained a lower accuracy compared to CNN-Transformer model. This discrepancy reveals key limitations of time-domain modeling alone, especially in the presence of the non-stationary and high-noise characteristics of EEG signals. Although LSTMs can learn sequential dependencies, their expressiveness in extracting localized spectral features is inherently limited. This aligns with prior studies (e.g., Albaqami et al., 2023) which also found LSTM models to be effective but not optimal when used in isolation for EEG classification.

Moreover, The Fusion Model offered an intermediate performance of 82.51% with f1-scores of 0.84 LPD and 0.81 for Seizure, outperforming the LSTM-Transformer but falling short of the CNN-Transformer. This suggests that while multimodal fusion adds complementary information, the selection of modality weights and fusion techniques significantly impacts performance. The performance gain is not automatic and heavily depends on how the features from different modalities are integrated. A simple fusion strategy, such as feature concatenation, may not adequately balance the contributions of each modality, especially when one (such as CNN-derived frequency features) is more informative. Moreover, without an adaptive weighting or attention mechanism to emphasize the most relevant features from each domain, the added complexity may introduce noise or redundancy, thereby limiting the potential performance gains. These findings highlight the importance of carefully designed fusion architectures and weighting strategies in multimodal EEG classification tasks.

While the CNN-Transformer model achieved the highest overall, it is important to look beyond accuracy and consider the balance between precision and recall for each class, as these metrics provide deeper insight into the model’s clinical applicability and robustness.

LPD Class: The Fusion Model demonstrates a notably higher recall for LPD at 0.93, surpassing both the CNN-Transformer (0.89) and the LSTM-Transformer (0.81) models. This suggests that the Fusion Model is more effective in correctly identifying true LPD cases, which is critical in clinical settings where missing such events could have serious consequences. Although its precision for LPD (0.77) is slightly lower than CNN-Transformer's (0.91), the increased recall indicates that the Fusion Model minimizes false negatives more aggressively, potentially leading to better patient safety by reducing missed LPD diagnoses.

Seizure Class: Conversely, the Fusion Model exhibits the highest precision for Seizure classification at 0.91, exceeding both CNN-Transformer (0.89) and LSTM-Transformer (0.80). This indicates that when the Fusion Model predicts a Seizure event, it is more likely to be correct, which is essential for avoiding false alarms and unnecessary interventions. However, the recall for Seizure in the Fusion Model is lower (0.72) compared to CNN-Transformer's 0.91, reflecting a trade-off where the model prioritizes precision over sensitivity for seizures.

F1-Score and Accuracy: Despite these class-specific advantages, the Fusion Model's overall accuracy is 82.51%, which is less than CNN-Transformer's 90.12%. This can be attributed to the imbalanced trade-offs between precision and recall in different classes. The Fusion Model achieves a higher f1-score for LPD (0.84) compared to LSTM-Transformer but lower than CNN-Transformer's (0.90), reflecting its moderate balance between precision and recall for this class. For Seizure, the Fusion Model's f1-score (0.81) is lower than CNN-Transformer's (0.90), again highlighting the trade-off in recall.

The Fusion Model's superior recall for LPD and superior precision for Seizure illustrate its specialized strengths in clinical seizure detection. It is more sensitive in detecting LPD cases (fewer missed diagnoses) and more confident when flagging Seizure events (fewer false alarms). This suggests that combining temporal (LSTM) and spectral (CNN) features through the fusion architecture may enable the model to leverage complementary information, improving detection robustness for specific classes even if the overall accuracy is not the highest.

In practical terms, the Fusion Model might be preferable in scenarios where minimizing missed LPD detections and reducing false seizure alarms are prioritized, despite a slightly lower overall accuracy. This nuanced performance profile highlights the importance of multi-metric evaluation and the potential clinical relevance of the fusion approach.

Building upon the evaluation of the proposed models, it is also important to contextualize these findings by comparing them with results reported in prior studies. Although the majority of existing works have reported their performance on the same dataset using Kullback–Leibler (KL) divergence, rather than the classification metrics employed in this study, a meaningful comparison of approaches can still be established. This allows for an approximate assessment of the relative effectiveness of the proposed models, particularly in terms of seizure detection accuracy and robustness across different classes

The results of the current model are consistent with prior studies that highlight the effectiveness of deep learning approaches in EEG signal classification (Saadoon et al, 2023). For instance, Singh & Lobiyal (2025) used a CNN-LSTM hybrid with STFT input and achieved 94.5% accuracy and F1-score of 0.9376. Their model, however, does not incorporate raw EEG data. Similarly, Qi et al. (2025) employed a hybrid 3D+2D CNN on spectral depth maps and achieved 98.43% accuracy with 98.58% sensitivity, showing the effectiveness of advanced spectral representations, but again without temporal-spectral fusion.

Moreover, Wei et al. (2025) used 3D-CNNs directly on raw EEG and reported >90% accuracy, while Yang et al. (2025) incorporated attention mechanisms in a CNN-LSTM framework to reach 92.07% accuracy, illustrating that time-domain modeling remains powerful. However, these methods often lack spectral context or cross-modal synergy. In contrast, the Fusion model in this study uniquely captures both modalities, showing clear gains in class-specific precision and recall metrics.

It's also worth noting that Zhu et al. (2025) combined Transformer and GRU models to model spectral dynamics and achieved a 98.24% sensitivity, underscoring the growing popularity of Transformer-based models. However, few studies have attempted the multimodal integration of both time and frequency features in a unified neural architecture, as proposed here.

some earlier works reported high accuracy (e.g., Ramos-Aguilar et al. with 100% accuracy using MLP on spectrograms), but they often lack transparency in how performance is measured across classes and do not generalize well to real-world datasets like CHB-MIT. Additionally, models such as those in Liu et al. (2025) and Assali et al. (2025) focus on boosting sensitivity without detailing class-specific trade-offs—limiting their clinical reliability.

The insights derived from the current study, in conjunction with findings from previous research, confirm that both time- and frequency-domain features contribute significantly to the performance of EEG-based classification models. This evidence supports the conclusion that a multimodal fusion approach holds substantial potential for improving both the specificity and robustness of EEG-based classification models. The results suggest that when a well-designed and properly implemented fusion strategy is employed, the complementary strengths of temporal dynamics and spectral characteristics can be leveraged to achieve more accurate and reliable detection of abnormal brain activity, outperforming approaches that rely solely on a single data modality.

Despite the promising results achieved in this study, several critical limitations and ongoing challenges must be acknowledged, particularly concerning generalizability, data complexity, computational resources, and interpretability.

7.3. Limitations and Challenges

Despite the promising results achieved in this study, several critical limitations and ongoing challenges must be acknowledged, particularly concerning generalizability, data complexity, computational resources, and interpretability.

One of the primary challenges encountered in this thesis was the computational intensity of processing and training multimodal EEG data, particularly when applying deep architectures such as LSTMs and Transformers. Although Optuna was used to systematically tune key hyperparameters, the large memory footprint and long training times associated with larger LSTM hidden sizes posed a significant constraint. In particular, the combination of high-dimensional EEG data and long sequence lengths made some configurations computationally infeasible within the resource and time limits of the university's high-performance computing (HPC) cluster. As a result, the hyperparameter search space had to be narrowed, and the final models were trained with a limited range of configurations. While these trade-offs were necessary to ensure feasible experimentation, they may have limited the potential for achieving even higher performance. Future work could explore larger-scale hyperparameter searches and more complex architectures if sufficient computational resources become available.

Additionally, although the original dataset included six seizure classes, this study focused on binary classification, which was not only computationally more feasible but also aligned with the primary objective of the study to investigate whether integrating temporal and spectral information through a multimodal deep learning architecture improves seizure classification performance. The decision was motivated by the high-dimensional nature of the multimodal input, consisting of 2,000 time steps per channel across 19 EEG channels, along with their corresponding spectrograms. This complexity made multi-class training prohibitively resource-intensive within the limits of the available computing infrastructure.

Another major challenge encountered during this study was the computational intensity of generating time–frequency representations for the multimodal deep learning architecture. In the early stages of experimentation, CWT was considered as an alternative to STFT due to its ability to provide rich, adaptive time–frequency analysis with better resolution at lower frequencies. However, applying CWT across the entire dataset, especially in combination with high-dimensional EEG data and deep neural networks, proved to be computationally prohibitive. The processing time, memory consumption, and storage requirements of CWT significantly exceeded available resources, making large-scale model training infeasible. As a result, the study opted for STFT, which offers a computationally efficient and interpretable method for spectrogram generation while still capturing essential spectral dynamics relevant to seizure activity. Although CWT may offer superior frequency localization, STFT provided a practical balance between signal resolution and resource efficiency, aligning well with the study's focus on evaluating multimodal integration under real-world constraints.

Generalizability across patient cohorts also remains a major unresolved issue in EEG-based deep learning since EEG signals are highly individualistic, and model performance often deteriorates when applied across patients with differing brain morphologies, underlying pathologies, or acquisition protocols. This raises concerns about the external validity of the model and its applicability in real-world clinical settings.

Another notable challenge concerns the feature fusion strategy employed. In this study, a straightforward concatenation of LSTM and CNN outputs was used before feeding the features into Transformer layers. While this approach establishes a useful baseline, it implicitly assumes equal importance for both time-domain and frequency-domain modalities.

Finally, interpretability remains a substantial barrier to clinical translation. CNN and Transformer layers, especially when operating on spectrogram inputs, act as black-box systems, offering limited insight into the decision-making process.

In summary, while this study contributes meaningfully to the field of seizure detection through multimodal modeling, it also highlights key areas requiring further investigation. Improving generalization, optimizing computational efficiency, adopting more advanced fusion strategies, and enhancing model interpretability are all essential next steps for translating deep learning models from experimental settings into clinically viable solutions.

7.4. Future Work

Given the promising yet limited scope of this study, several avenues for future research remain open. One important direction is to expand the classification task beyond binary categories to include all six clinically relevant classes, such as Generalized Periodic Discharges (GPD), Lateralized Periodic Discharges (LPD), Lateralized Rhythmic Delta Activity (LRDA), and Generalized Rhythmic Delta Activity (GRDA), among others. Doing so would significantly enhance the clinical relevance and diagnostic utility of the model.

In addition to task expansion, adopting more advanced temporal modelling techniques could further enhance model performance. While LSTMs were chosen for in this study, future work could explore more expressive architectures, such as GRU or Temporal Convolutional Networks (TCNs), which are better suited to capture long-range temporal dependencies in EEG data without incurring excessive computational costs.

Future work could also explore the use of CWT for time–frequency representation. Although it provides richer spectral resolution than STFT, it was excluded in this study due to its high computational cost. With more advanced hardware or optimized implementations, CWT could improve the model's ability to detect subtle seizure patterns.

Furthermore, the current results highlight the importance of effective multimodal fusion strategies. Future efforts should focus on developing more sophisticated fusion mechanisms, such as modality-specific attention layers or cross-modal transformers, that dynamically adjust the importance of temporal and spectral features based on context. This could lead to more adaptive and accurate models.

Finally, to ensure generalizability and reproducibility, future research must prioritize evaluation on external datasets collected across multiple institutions. Addressing heterogeneity in EEG acquisition protocols, including differences in sampling rates, electrode placements, and annotation standards, will be essential. The development and adoption of standardized, large-scale, and demographically diverse EEG datasets would also enable more reliable benchmarking and facilitate model replication.

Collectively, these directions point toward the development of clinically viable, generalizable, and interpretable deep learning models capable of seizure detection and prediction, ultimately supporting better patient care and timely intervention.

References

Abhang, Bharti W. Gawali, Suresh C. Mehrotra, Chapter 2 - Technological Basics of EEG Recording and Operation of Apparatus, Editor(s): Priyanka A. Abhang, Bharti W. Gawali, Suresh C. Mehrotra, Introduction to EEG- and Speech-Based Emotion Recognition, Academic Press, 2016, Pages 19-50, ISBN 9780128044902, https://doi.org/10.1016/B978-0-12-804490-2.00002-6.

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Computers in Biology and Medicine, 100, 270–278. https://doi.org/10.1016/j.compbiomed.2017.09.017

Albaqami, H., Hassan, G. M., & Datta, A. (2023). EEG-based seizure detection using LSTM and wavelet transform. IEEE Transactions on Biomedical Engineering.

Aluvalu, R., Aravinda, K., Maheswari, V. U., Kumar, K. A. J., Rao, B. V., & Prasad, K. M. V. V. (2024). Designing a cognitive smart healthcare framework for seizure prediction using multimodal convolutional neural network. Cognitive Neurodynamics. Advance online publication. https://doi.org/10.1007/s11571-023-10049-x.

A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks", Nature, vol. 542, no. 7639, pp. 115, 2017.

A. Ng, "Machine learning yearning: Technical strategy for ai engineers in the era of deep learning", 2019.

Ashit Kumar, A., Zhang, Y., & Li, X. (2024). Deep learning for EEG seizure detection: A comparative study of CNN and hybrid architectures. Journal of Neural Engineering.

Assali, A., Rekik, I., & Soltanian-Zadeh, H. (2025). EEG-based seizure prediction using deep neural networks. Neurocomputing.

Başar, E., Başar-Eroglu, C., Karakaş, S., & Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International Journal of Psychophysiology*, **39**(2–3), 241–248.

Bancaud, J., Hecaen, H., and Lairy, G. C., ‘Modification de la reactivite EEG, troubles des functions symboliques et troubles con fusionels dans les lesions hemi-spherigues localisees’, Electroencephalogr. Clin. Neurophysiol., 7, 1955, 179.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ digital medicine, 3(1), 118.

Bengio, Y., Simard, P., & Frasconi, P. (1994). "Learning long-term dependencies with gradient descent is difficult." IEEE Transactions on Neural Networks.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten*, **87**, 527–570.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). "Language Models are Few-Shot Learners." arXiv preprint arXiv:2005.14165.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. Science, 304(5679), 1926–1929.

Buzsaki, G. (2006). Rhythms of the Brain. Oxford university press.

C. Metz, Turing Award Won by 3 Pioneers in Artificial Intelligence, New York, NY, USA:New York Times, pp. B3, 2019.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Clarke, S., Karoly, P. J., Nurse, E., Seneviratne, U., Taylor, J., Knight-Sadler, R., ... & D'Souza, W. (2021). Computer-assisted EEG diagnostic review for idiopathic generalized epilepsy. Epilepsy & Behavior, 121, 106556.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding." NAACL-HLT.

Dieleman, S., Oord, A. V. D., & Simonyan, K. (2018). "The challenge of realistic music generation: modelling raw audio at scale." International Society for Music Information Retrieval Conference (ISMIR).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint arXiv:2010.11929.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929.

Dutta, A. K., Raparthi, M., Alsaadi, M., Bhatt, M. W., Dodda, S. B., Sandhu, M., & Patni, J. C. (2024). Deep learning-based multi-head self-attention model for human epilepsy identification from EEG signal for biomedical traits. Multimedia Tools and Applications, 1-23.

Emotiv. (n.d.). *Understanding the 10-20 System of EEG Electrode Placement*. Retrieved July 7, 2025, from *https://www.emotiv.com/blogs/how-to/understanding-the-10-20-system-of-eeg-electrode-placement?srsltid=AfmBOoqhIyf1o3kjhTnl0Z201AXg_pnihmV2Zjn0nkARbBHI1f-6sSxs*

Fatourechi, M., Bashashati, A., Ward, R. K., & Birch, G. E. (2007). EMG artifact removal in BCI systems: A comparative study. *Clinical Neurophysiology, 118*(3), 480–494.

Feng, T., Ni, J., Gleichgerrcht, E., & Jin, W. (2025). SeizureFormer: A Transformer Model for IEA-Based Seizure Risk Forecasting. arXiv preprint arXiv:2504.16098.

Fisher, R. S., Boas, W. V. E., Blume, W., Elger, C., Genton, P., Lee, P., & Engel Jr, J. (2005). Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia, 46(4), 470-472.

Fraiwan L, Alkhodari M (2020) Classification of focal and non-focal epileptic patients using single channel EEG and long short-term memory learning system. IEEE Access 8:7725577262. https://doi.org/10.1109/ACCESS.2020.2989442

F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain", Psychol. Rev., vol. 65, pp. 386-408, 1958.

Jasper, H.H., & Andrews, H.L. (1938). Brain potentials and voluntary muscle activity in man. *Journal of Neurophysiology*, **1**, 87–100.

Jin Jing, Zhen Lin, Chaoqi Yang, Ashley Chow, Sohier Dane, Jimeng Sun, and M. Brandon Westover. HMS - Harmful Brain Activity Classification . https://kaggle.com/competitions/hms-harmful-brain-activity-classification, 2024. Kaggle.

G. Cybenko, "Approximation by superpositions of a sigmoidal function", Math. Control Signals Syst., vol. 2, no. 4, pp. 303-314, 1989.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). "Learning to forget: Continual prediction with LSTM." Neural Computation, 12(10), 2451-2471.

Gloor, P. (1977). *The EEG and Differential Diagnosis of Epilepsy*. Masson.

Graves, A., Mohamed, A. R., & Hinton, G. (2013). "Speech recognition with deep recurrent neural networks." ICASSP.

Graves, A., & Schmidhuber, J. (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures." Neural Networks, 18(5-6), 602-610.

Graves, A. (2013). "Generating sequences with recurrent neural networks." arXiv preprint arXiv:1308.0850.

Han, K., Liu, C., & Friedman, D. (2024). Artificial intelligence/machine learning for epilepsy and seizure diagnosis. Epilepsy & Behavior, 155, 109736.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hirsch, L. J., Fong, M. W., Leitinger, M., LaRoche, S. M., Beniczky, S., Abend, N. S., ... & Gaspard, N. (2021). American clinical neurophysiology society's standardized critical care EEG terminology: 2021 version. *Journal of clinical neurophysiology*, *38*(1), 1-29.

Hochreiter, S., & Schmidhuber, J. (1997). "Long short-term memory." Neural Computation, 9(8), 1735-1780.

Hopfield, J. J. (1982). "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." Proceedings of the National Academy of Sciences.

Hussein, R.; Lee, S.; Ward, R. Multi-Channel Vision Transformer for Epileptic Seizure Prediction. Biomedicines 2022, 10, 1551. https://doi.org/10.3390/biomedicines10071551.

K. Hornik, "Approximation capabilities of multilayer feedforward networks", Neural Netw., vol. 4, no. 2, pp. 251-257, 1991.

Kellaway, P., & Crawley, J. W. (1953). EEG patterns associated with raised intracranial pressure. *Neurology*, **3**, 198–208.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." Advances in Neural Information Processing Systems (NeurIPS), pp. 1097-1105.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11), 2278-2324.

Li, C., Huang, X., Song, R., Qian, R., Liu, X., & Chen, X. (2022). EEG-based seizure prediction via Transformer guided CNN. Measurement, 203, 111948.

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzell, R. (2016). "Learning to diagnose with LSTM recurrent neural networks." arXiv preprint arXiv:1511.03677.

Liu, Y., Zhou, W., Yuan, Q., & Chen, S. (2025). Automated EEG classification using deep learning: Challenges and opportunities. Computers in Biology and Medicine.

L. Xia et al., "Hybrid LSTM–Transformer Model for the Prediction of Epileptic Seizure Using Scalp EEG," in IEEE Sensors Journal, vol. 24, no. 13, pp. 21123-21131, 1 July1, 2024, doi: 10.1109/JSEN.2024.3401771.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Marco Zurdo, J., Martinez-de Dios, J. R., & Conde, C. (2023). Transformer-based models for EEG signal processing: A review. Sensors.

Makeig, S., Bell, A., Jung, T. P., & Sejnowski, T. J. (1995). Independent component analysis of electroencephalographic data. Advances in neural information processing systems, 8.

M. I. Jordan and T. M. Mitchell, "Machine learning: Trends perspectives and prospects", Science, vol. 349, no. 6245, pp. 255-260, 2015.

M. Minsky and S. A. Papert, Perceptrons: An Introduction to Computational Geometry Expanded Edition, Cambridge, MA, USA:MIT Press, pp. 258, 1969.

McLeod, A. J., et al. (1983). A brief historical account of EEG frequency bands. *Neuroscience Bulletin*, **6**, 153–157.

McLeod, A. J., and Li, W. K., ‘Diagnostics checking ARMA time series models using squared residual autocorrelations’, J. Time Series Analysis, 4, 1983, 269–273.

Niedermeyer, E., & da Silva, F. L. (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (5th ed.). Lippincott Williams & Wilkins.

Ni J, Young T, Pandelea V, Xue F, Cambria E (2022) Recent advances in deep learning based dialogue systems: a systematic survey. Artif Intell Rev 56:1–101.

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of EEG* (2nd ed.). Oxford University Press.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. ACM Comput. Surv. 56, 10, Article 264 (October 2024), 42 pages. https://doi.org/10.1145/3656580.

Pellinen, J., French, J., & Knupp, K. G. (2021). Diagnostic delay in epilepsy: the scope of the problem. Current Neurology and Neuroscience Reports, 21, 1-6.

Prenger, R., Valle, R., & Catanzaro, B. (2019). "WaveGlow: A flow-based generative network for speech synthesis." ICASSP 2019.

P. J. Werbos, "‘Beyond Regression:’ New tools for prediction and analysis in the behavioral sciences", 1975.

Priyanka A. Abhang, Bharti W. Gawali, Suresh C. Mehrotra, Chapter 2 - Technological Basics of EEG Recording and Operation of Apparatus, Editor(s): Priyanka A. Abhang, Bharti W. Gawali, Suresh C. Mehrotra, Introduction to EEG- and Speech-Based Emotion Recognition, Academic Press, 2016, Pages 19-50, ISBN 9780128044902, https://doi.org/10.1016/B978-0-12-804490-2.00002-6.

Qi, J., Wang, H., & Li, Y. (2025). *3D-CNN for EEG spectral depth mapping in seizure detection*. IEEE Journal of Biomedical and Health Informatics.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). "Improving language understanding by generative pre-training." OpenAI.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al. (2021). "Learning transferable visual models from natural language supervision." arXiv preprint arXiv:2103.00020.

Ramos-Aguilar, J., Guzmán-Martínez, M., Hernández-Hernández, M., & García-Ramírez, E. (2025). EEG Signal Analysis for Epileptic Seizure Detection:

Algorithms, Challenges, and Opportunities. *Journal of Education for Pure Science*, *15*(1), 101-112.

Rhee, E. J. (2018). A deep learning approach for classification of cloud image patches on small datasets.

Rodriguez Ruiz, A. , Vlachy, J. , Lee, J. W. , Gilmore, E. J. , Ayer, T. , Haider, H. A. , Gaspard, N. , Ehrenberg, J. A. , Tolchin, B. , Fantaneanu, T. A. , Fernandez, A. , Hirsch, L. J. & LaRoche, S. (2017). Association of Periodic and Rhythmic Electroencephalographic Patterns With Seizures in Critically Ill Patients. JAMA Neurology, 74 (2), 181-188. doi: 10.1001/jamaneurol.2016.4990.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5), 051001. https://doi.org/10.1088/1741-2552/ab260c

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). "Learning representations by back-propagating errors." Nature.

R. C. V, K. H, K. G and K. S. P, "Identification of Epileptic Seizures using Recurrent Neural Networks and Time Series Transformer," 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2024, pp. 1546-1553, doi: 10.1109/ICCPCT61902.2024.10673329.

Saadoon, Yousif A., Mohamad Khalil, and Dalia Battikh. "Machine and Deep Learning-Based Seizure Prediction: A Scoping Review on the Use of Temporal and Spectral Features." *Applied Sciences (2076-3417)* 15.11 (2025).

Sanei, S., & Chambers, J. A. (2013). EEG signal processing. John Wiley & Sons., Introduction to EEG, 2007, EEG Signal Processing, 1-34.

Sheykhivand S, Rezaii TY, Mousavi Z, Delpak A, Farzamnia A (2020) Automatic identification of epileptic seizures from EEG signals using sparse representation-based classification. IEEE Access 8:138834–138845. https://doi.org/10.1109/ACCESS.2020.3011877

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., et al. (2018). "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions." ICASSP 2018.

Sharbrough, F. W., ‘Nonspecific abnormal EEG patterns’, Chapter. 12, in Electroencephalography, Basic Principles, Clinical Applications, and Related Fields, Eds E. Niedermeyer and F. Lopes Da Silva, 4th edn., Lippincott, Williams and Wilkins, Philadelphia, Pennsylvania, 1999.

Shoeb, A. (2009). Application of machine learning to epileptic seizure onset detection and treatment (Doctoral dissertation, MIT). *https://dspace.mit.edu/handle/1721.1/54602*

Simonyan, K., & Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556.

Smith, S. J. M. (2005). EEG in the diagnosis, classification, and management of patients with epilepsy. Journal of Neurology, Neurosurgery & Psychiatry, 76(suppl 2), ii2–ii7. *https://doi.org/10.1136/jnnp.2005.069245*

Smith, D., Defalla, B. A., & Chadwick, D. W. (1999). The misdiagnosis of epilepsy and the management of refractory epilepsy in a specialist clinic. Qjm, 92(1), 15-23.

Singh, P., & Lobiyal, D. K. (2025). Seizure Detection in Medical IoT: Hybrid CNN-LSTM-GRU Model with Data Balancing and XAI Integration. *Sensors*, *18*(2), 77.

Steriade, M., McCormick, D.A., & Sejnowski, T.J. (1993). Thalamocortical oscillations in the sleeping and aroused brain. *Science*, **262**(5134), 679–685.

Stam, C. J. (2005). Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. Clinical neurophysiology, 116(10), 2266-2301.

Sun, C., Xu, C., Li, H., Bo, H., Ma, L., & Li, H. (2024). A novel multi-feature fusion attention neural network for the recognition of epileptic EEG signals. Frontiers in Computational Neuroscience, 18, 1393122. https://doi.org/10.3389/fncom.2024.1393122

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to sequence learning with neural networks." NeurIPS.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Tan, M., & Le, Q. V. (2019). "EfficientNet: Rethinking model scaling for convolutional neural networks." arXiv preprint arXiv:1905.11946.

Tian X et al (2019) Deep multi-view feature learning for eeg-based epileptic seizure detection. IEEE Trans Neural Syst Rehabil Eng 27(10):1962–1972. https://doi.org/10.1109/TNSRE.2019.2940485

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). "Training data-efficient image transformers and distillation through attention." arXiv preprint arXiv:2012.12877.

Tsiouris, K. M., Pezoulas, V. C., Zervakis, M., Konitsiotis, S., Koutsouris, D. D., & Fotiadis, D. I. (2018). A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals. Computers in biology and medicine, 99, 24–37. https://doi.org/10.1016/j.compbiomed.2018.05.019.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). "WaveNet: A generative model for raw audio." arXiv preprint arXiv:1609.03499.

van den Oord, A., et al. (2016). "WaveNet: A generative model for raw audio." arXiv preprint arXiv:1609.03499.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Graves, A., et al. (2018). "Parallel WaveNet: Fast high-fidelity speech synthesis." arXiv preprint arXiv:1711.10433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention is all you need." Advances in Neural Information Processing Systems (NeurIPS).

Vaswani, A., et al. (2017). "Attention is all you need." NeurIPS.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

Virtanen, P., Gommers, R., Oliphant, T.E., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walter, W.G. (1936). The location of cerebral tumors by electro-encephalography. *Lancet*, **1**, 305–308.

Wei, J., Li, X., Wang, Y., & Zhang, H. (2025). Predicting Epileptic Seizures Using EfficientNet-B0 and SVMs: A Deep Learning Methodology for EEG Analysis. Journal of Imaging, 12(2), 109.

Wei, X., Zhou, L., Zhang, Z., Chen, Z., & Zhou, Y. (2019). Early prediction of epileptic seizures using a long-term recurrent convolutional network. Journal of neuroscience methods, 327, 108395. https://doi.org/10.1016/j.jneumeth.2019.108395.

Westmoreland, B., and Klass, D., 'Asymmetrical attention of alpha activity with arithmetical attention', Electroencephalogr. Clin. Neurophysiol., 31, 1971, 634–635.

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data – A practical approach. *Journal of Neuroscience Methods, 250*, 34–46.

WHO. Epilepsy. WHO; 2024. *https://www.who.int/news-room/fact-sheets/detail/epilepsy*.

Wolter, J.R., & Dovey, V.J. (1944). Theta waves in human EEG: Their occurrence and clinical significance. *Electroencephalography and Clinical Neurophysiology*, **4**, 123–132.

Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", Nature, vol. 521, no. 7553, pp. 436-444, May 2015.

Y. Sun et al., "Continuous Seizure Detection Based on Transformer and Long-Term iEEG," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 11, pp. 5418-5427, Nov. 2022, doi: 10.1109/JBHI.2022.3199206.

Yang, Y., Zhang, R., Wang, L., & Li, X. (2025). A Multi-Channel Feature Fusion CNN-Bi-LSTM Epilepsy EEG Classification and Prediction Model Based on Attention Mechanism. arXiv preprint arXiv:2306.09633.

Yuan, S., Yan, K., Wang, S., Liu, J.-X., & Wang, J. (2024). EEG-Based Seizure Prediction Using Hybrid DenseNet–ViT Network with Attention Fusion. Brain Sciences, 14(8), 839. https://doi.org/10.3390/brainsci14080839mdpi.com

Zhiwen Xiong, Yang Liu & Peng Jiang. (2024) An EEG-Based Seizure Recognition Method Using Dynamic Routing. IEEE Access 12, pages 74054-74068.

Zhu, T., Ma, K., Wu, S., & Chen, X. (2025). Multi-branch GAT-GRU-transformer for explainable EEG-based finger motor imagery classification. Frontiers in Human Neuroscience, 19, 1599960.

Zurdo-Tabernero, M., Canal-Alonso, Á., de la Prieta, F., Rodríguez, S., Prieto, J. & Corchado, J. (2023). An overview of machine learning and deep learning techniques for predicting epileptic seizures. Journal of Integrative Bioinformatics, 20(4), 20230002. https://doi.org/10.1515/jib-2023-0002.

Appendix 1

The source code and implementation details for this study are available on *GitHub*