applied sciences

MDPI

Article

Effective Epileptic Seizure Detection with Hybrid Feature Selection and SMOTE-Based Data Balancing Using SVM Classifier

Hany F. Atlam [1,*], Gbenga Ebenezer Aderibigbe [2] and Muhammad Shahroz Nadeem [3]

[1] Cyber Security Centre, Warwick Manufacturing Group (WMG), University of Warwick, Coventry CV4 7AL, UK
[2] Department of Science and Engineering, Southampton Solent University, Southampton SO14 0YN, UK; ebenezer.aderibigbe@solent.ac.uk
[3] College of Computer and Information Systems, University of Suffolk, Ipswich IP4 1QJ, UK; s.nadeem3@uos.ac.uk
* Correspondence: hany.atlam@warwick.ac.uk

Abstract: Epileptic seizures, a leading cause of global morbidity and mortality, pose significant challenges in timely diagnosis and management. Epilepsy, a chronic neurological disorder characterized by recurrent and unpredictable seizures, affects over 70 million people worldwide, according to the World Health Organization (WHO). Despite significant advances in medical science, accurate and timely diagnosis of epileptic seizures remains a challenge, with misdiagnosis rates reported to be as high as 30%. The consequences of misdiagnosis or delayed diagnosis can be severe, leading to increased morbidity, mortality, and reduced quality of life for patients. Therefore, this paper presents a novel approach to enhancing epileptic seizure detection through the integration of Synthetic Minority Over-Sampling Technique (SMOTE) for data balancing and a Hybrid Feature Selection Technique—Principal Component Analysis (PCA) and Discrete Wavelet Transform (DWT). The proposed model aims to improve the accuracy and reliability of seizure detection systems by addressing data imbalance and extracting discriminative features from electroencephalograms (EEG) signals. Experimental results demonstrate substantial performance gains, with the Support Vector Machine (SVM) classifier achieving 97.30% accuracy, 99.62% Area Under the Curve (AUC), and 93.08% F1 score, which outperform the results of the existing studies from the literature. The results highlight the effectiveness of the proposed model in advancing seizure detection systems, highlighting the potential to improve diagnostic capabilities and patient outcomes.

Keywords: epilepsy; machine learning; epileptic seizures; EEG; seizure detection; SMOTE; PCA; DWT

Academic Editor: Emi Yuda

Received: 30 March 2025
Revised: 17 April 2025
Accepted: 22 April 2025
Published: 23 April 2025

Citation: Atlam, H.F.; Aderibigbe, G.E.; Nadeem, M.S. Effective Epileptic Seizure Detection with Hybrid Feature Selection and SMOTE-Based Data Balancing Using SVM Classifier. *Appl. Sci.* **2025**, *15*, 4690. https://doi.org/10.3390/app15094690

1. Introduction

Epilepsy, a chronic neurological disorder that affects the Central Nervous System (CNS), presents a formidable challenge due to its spontaneous recurrent seizures and the absence of a known cure [1,2]. These seizures not only pose immediate risks such as falls, fractures, and fatalities but also contribute to long-term neurological harm. Symptoms such as disorientation, unusual behaviour, and loss of consciousness often accompany seizures, further compounding the challenges faced by individuals with epilepsy [3].

Detecting impending seizures remains a daunting task due to their unpredictable nature, and most seizures occur without warning. Consequently, researchers have focused on developing methods for predicting seizures, leveraging classification algorithms as a viable approach [1].

Research in epilepsy detection has explored diverse paths, including the application of differential equations to model the dynamic behaviour of the brain's electrical activities [3]. Nonlinear Time Series Analysis (NTSA) has been instrumental in characterizing electroencephalograms (EEGs) based on brain activities [2]. Studies examining EEGs of patients with various conditions, including Parkinson's, depression, Alzheimer's, and epilepsy, have contributed significantly to understanding the complex dynamics of the human brain [4,5]. Epileptic seizures and EEGs of healthy individuals have played crucial roles in categorizing findings in related articles, aiding in identifying and modelling the non-linearity of human brain behaviour [5]. Even minor alterations to the dynamic system parameters of the brain can lead to diverse physiological states, potentially resulting in brain dysfunction or other disorders [5]. Consequently, predicting and understanding epileptic seizures continues to challenge researchers [6].

Improving the quality of life for individuals with epilepsy hinges on accurately detecting and predicting seizures. EEG-based seizure detection remains fundamental in understanding neuronal brain activity [7]. However, the complexity of EEG seizure detection, characterized by dynamic motion and perspective fluctuations, presents a significant challenge [3]. Moreover, the impact of epilepsy on affected individuals is substantial [8]. Seizures, which occur from abnormal electrical discharges of the brain, manifest in various forms, ranging from momentary lapses in attention to severe convulsions [9]. Furthermore, the effects of epilepsy on the CNS result in symptoms such as loss of awareness, unusual behaviour, and confusion, further increasing the risk of injuries [9]. The imperative to predict impending seizures arises from their sudden and unexpected onset [9]. Researchers have grappled with developing methods to anticipate seizures, driven by the potential for physical harm and, in extreme cases, fatality. According to recent data from the World Health Organization (WHO), epilepsy affects approximately 50 million people worldwide, emphasizing the urgent need for effective seizure detection and prediction methods to mitigate its impact [10].

Traditional approaches to seizure detection have relied heavily on visual analysis of EEG recordings by healthcare professionals. However, these methods suffer from subjectivity and time-intensive processes, often leading to delayed responses and missed early seizure indicators [11]. To address these challenges and enhance the efficiency and accuracy of seizure detection, researchers have turned to Machine Learning (ML). ML algorithms offer a promising target for automating seizure detection through EEG data analysis. They excel in processing the vast information contained within EEG recordings and learning to recognize patterns associated with seizures [12]. The integration of ML techniques aims to enable early and precise detection of epileptic seizures to reduce associated risks and minimize false positives [11].

This paper aims to propose a novel approach for epileptic seizure detection by integrating the Synthetic Minority Over-Sampling Technique (SMOTE) for data balancing and a hybrid feature selection method combining Principal Component Analysis (PCA) and Discrete Wavelet Transform (DWT). Unlike previous models that typically use either temporal or spectral features in isolation, our approach fuses both domains to enhance discriminative power. Also, existing models do not reflect real-world clinical data distributions. This leads to overfitting and poor generalization. Moreover, existing approaches often overlook feature redundancies and noise. Therefore, using either single-step feature extraction or limited dimensionality reduction is effective. Additionally, unlike traditional methods, our approach addresses the issue of data imbalance and leverages advanced

feature extraction techniques to enhance the accuracy and reliability of seizure detection systems. This layered pipeline is designed to improve both robustness and minority class sensitivity. Compared to existing literature, our method demonstrates significant performance improvements, achieving higher accuracy and robustness in detecting seizures from electroencephalogram (EEG) signals. The primary contributions outlined in this paper include the following:

- Proposing a novel approach consisting of a holistic pipeline for epileptic seizure detection using ML techniques;
- Addressing data imbalance through the integration of SMOTE to improve model performance;
- Proposing a hybrid feature selection technique combining PCA and DWT to extract discriminative features from EEG signals;
- Providing a comprehensive evaluation of the proposed approach, demonstrating significant performance gains and validating its effectiveness in enhancing seizure detection systems.

The rest of this paper is organized as follows: Section 2 provides the necessary background about the topic, Section 3 presents related work, Section 4 introduces the proposed epileptic seizure detection model followed by experimental results in Section 5, Section 6 presents the discussion and evaluation, and finally, Section 7 concludes the paper.

2. Background

Epileptic seizures represent a diverse array of neurological events characterized by abnormal electrical activity in the brain, often discernible through electroencephalogram (EEG) recordings [8]. These seizures manifest in various forms, ranging from tonic–clonic convulsions to absence spells (Petit Mal), each presenting unique EEG signatures and clinical manifestations. Understanding the complex interplay between epileptic seizures and EEG signals is paramount for developing effective diagnostic and therapeutic strategies in epilepsy management.

2.1. Types and Characteristics of Epileptic Seizures

Epileptic seizures encompass a spectrum of events with distinct clinical and electrographic features. Tonic–clonic seizures, the most recognizable type of seizures, are marked by sudden loss of consciousness, tonic muscle contractions, and clonic jerking movements. EEG recordings during tonic–clonic seizures typically reveal high-amplitude, generalized rhythmic oscillations known as spike-wave complexes. In contrast, absence seizures manifest as brief lapses in consciousness, often accompanied by generalized 3 Hz spike-and-wave discharges on the EEG, indicative of synchronous neuronal activity in thalamocortical networks.

2.2. Mechanisms Underlying Epileptic Seizures

Epileptic seizures arise from disruptions in normal neuronal excitability and synchronization, leading to hyperexcitability and hypersynchronous firing of neuronal populations. Aberrant network activity can propagate through cortical and subcortical structures, leading to characteristic EEG patterns associated with seizure onset, propagation, and termination. These electrographic signatures, ranging from spike-and-wave discharges to rhythmic delta or theta activity, provide valuable insights into the underlying pathophysiology of epileptic seizures.

2.3. Challenges in EEG-Based Seizure Detection

Despite the diagnostic utility of EEG signals in seizure detection, several challenges hamper their clinical application. Distinguishing pathological seizure activity from normal brain rhythms and artifacts poses a significant challenge, requiring sophisticated signal processing techniques and pattern recognition algorithms. In addition, artifacts such as muscle activity, electrode movement, and environmental interference can obscure genuine seizure events, necessitating robust artifact removal and noise reduction strategies. Additionally, the imbalanced nature of seizure datasets, where the number of seizure and non-seizure instances is disproportionate, presents challenges for training accurate and generalizable classification models. To overcome these challenges, researchers have employed advanced techniques and methodologies in EEG-based seizure detection. Hybrid feature selection methods, such as PCA combined with DWT, offer enhanced discriminative power by capturing both temporal and spectral features from EEG signals. Furthermore, techniques such as SMOTE enable the generation of synthetic samples to balance imbalanced datasets, ensuring robust model training and improved classification performance. Leveraging state-of-the-art classifiers, including SVM, DT, RF, and KNN, facilitates accurate classification of EEG segments into seizure and non-seizure classes, thereby enhancing the diagnostic accuracy of seizure detection systems.

Figure 1 depicts the classification of epileptic seizures into generalized and focal types. Generalized seizures involve widespread brain activity, while focal seizures originate in a specific brain region. Focal seizures can be simple or complex, with or without loss of consciousness. Tonic–clonic seizures have distinct phases of muscle stiffening and jerking, while atonic seizures cause sudden loss of muscle tone. Myoclonic seizures involve brief muscle contractions, clonic seizures feature rhythmic jerking movements, and tonic seizures involve sustained muscle contraction. Absence seizures (Petit Mal) are characterized by brief lapses in consciousness. This classification aids in diagnosis and treatment planning.

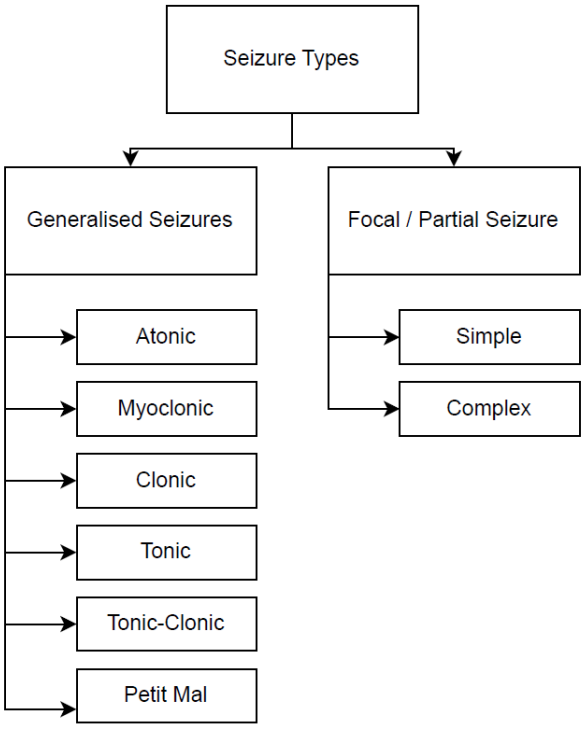

Figure 1. Types of seizures and subtypes.

3. Related Work

Several researchers have proposed techniques for seizure detection through EEG signal analysis. Such studies span advancing ML techniques. This section categorizes the reviewed papers into specific themes: early detection, feature engineering, classification techniques, and comprehensive reviews and methodologies. Several researchers proposed different techniques for early seizure detection. Shoeb [13] introduced a patient-specific seizure onset detection method leveraging Support Vector Machine (SVM) analysis of scalp EEG signals, achieving an accuracy of 96%. Similarly, Usman et al. [14] presented a preictal state detection model with promising accuracy but lacked information on false alarm rates, critical for assessing the model's reliability in real-world settings. Donos et al. [15] also presented an early seizure detection algorithm for implantable closed-loop stimulation devices, achieving high sensitivity with minimal detection delay. However, specific dataset details and implementation challenges were lacking.

Feature extraction and selection play a crucial role in enhancing the performance of seizure detection models. Anurag and Sisodia [16] proposed an automated method integrating frequency and time domain features with a flexible wavelet selection algorithm, achieving 97.08% accuracy using Random Forest classifiers. However, their work lacked significant details regarding computational efficiency and real-time feasibility. Nahzat and Yağanoğlu [17] examined feature selection and classification techniques for epileptic seizure prediction, emphasizing the trade-off between accuracy and computational time with the PCA algorithm. While achieving high accuracy without PCA, they noted the need for further optimization to balance accuracy and computational efficiency. Poorani and Balasubramanie [18] proposed two deep learning models for patient-specific seizure detection using CHB-MIT data: a 1D CNN and a CNN-LSTM hybrid. The models achieved 94.83% accuracy, 90.18% sensitivity, and 99.48% specificity. Despite improvements, the variability in performance across different patients highlights challenges in developing a generalized detection system.

Guo et al. [19] proposed an approach for detecting epileptic seizures in EEG signals using the line length feature extracted through DWT and a three-layer MLPNN for classification. Despite high classification accuracies of up to 99.60%, limitations regarding dataset preprocessing were acknowledged, necessitating further validation under real clinical conditions. On the other hand, Nicolaou and Georgiou [20] introduced Permutation Entropy (PE) and SVM for automated epileptic seizure detection, achieving high accuracy with a reported sensitivity of 94.38% and specificity of 93.23%. However, the focus on post-event classification without real-time detection challenges posed limitations.

Classification techniques form the backbone of seizure detection models. Wang et al. [21] proposed a postictal seizure detection method with impressive sensitivity and specificity, achieving 100% sensitivity and 98.5% specificity. However, its exclusive focus on postictal detection posed limitations in early intervention scenarios. Similarly, Sharmila and Geethanjali [7] proposed a framework for detecting epileptic seizures using DWT coupled with Naïve Bayes and KNN classifiers, achieving remarkable accuracy in classifying various epileptic dataset subtypes, with reported accuracies of over 99%. However, the study lacked exploration of enhancing overall seizure detection or tracking treatment effectiveness. Further, Hamad et al. [22] introduced a hybrid EEG classification approach using Grey Wolf Optimizer and SVMs, achieving an accuracy of 99%. However, specific accuracy and sensitivity metrics were lacking, hindering a comprehensive evaluation of the model's performance. Song and Liò [23] introduced a novel approach for automated epileptic seizure detection using sample entropy for feature extraction and an extreme ML for classification, emphasizing high accuracy and computational speed. However, specific dataset details were not specified.

Martis et al. [24] provided a comprehensive review of automated EEG signal classification methodologies, emphasizing non-linear features like entropy. While achieving notable classification accuracy, the research gap lies in the necessity for more diverse data and refined features to enhance accuracy. Similarly, Farooq et al. [25] conducted a systematic literature review examining ML techniques for epileptic seizure detection, focusing on various feature selection and classification techniques. Their taxonomy of state-of-the-art solutions provides insights into challenges and opportunities in the field, yet gaps in specific methodology details were noted. Chandel et al. [26] discussed ML-based classification models for epileptic EEG signals, comparing RF, DT, and Extra Tree classifiers. Despite achieving high accuracy, explicit limitations of the work were not mentioned.

In addition, Shoeb and Guttag [27] proposed an ML approach for constructing patient-specific classifiers for detecting epileptic seizures using scalp EEG data, showcasing accurate detection with a 96% detection rate. However, the reliance on patient-specific classifiers posed limitations. Pinto-Orellana and Cerqueira [28] presented a patient-dependent offline system for seizure detection in epilepsy diagnosis, achieving high specificity, sensitivity, and low false-positive rates. However, specific dataset details and limitations were not disclosed. Khurshid et al. [29] also proposed an approach for detecting epileptic seizures in EEG signals achieving a 96.25% detection accuracy. The study highlights the potential of deep learning models in improving seizure detection but notes limitations related to generalizability across different patient profiles. Raghu et al. [30] proposed deep-learning models for predicting epileptic seizures using iEEG signals. They developed a CNN and a CNN-LSTM hybrid model, achieving a maximum accuracy of 95.48%, sensitivity of 92.37%, and specificity of 96.18%. While the models demonstrated strong performance, challenges remain in handling variations in seizure patterns across different patients.

In addition, Zabihi et al. [31] proposed a patient-specific seizure detection method using phase space representation and time-delay embedding to analyze EEG dynamics. PCA was applied for dimensionality reduction, and key features were classified using LDA and Naïve Bayesian classifiers. The results achieved 88.27% sensitivity and 93.21% specificity. However, its reliance on patient-specific models may limit broader applicability. Also, Krishnan and Balasubramanian [32] proposed an autonomous epilepsy detection system using a Time–Frequency (TF) entropy measure to reduce EEG analysis time. The method computes the TF spectrum via S-transform and extracts entropy features, classified using an LSSVM classifier. The results achieved 86% accuracy with an AUC of 0.914. However, computational efficiency remains a consideration for real-time applications.

Moreover, Chen et al. [33] proposed a framework for epileptic focus localization using DWT and SVM. The method optimizes DWT parameters by analyzing seven wavelet families and selecting the best decomposition levels for feature extraction. The results achieved 83.07% accuracy. Also, Aarabi et al. [34] proposed an automatic seizure detection system for newborns, focusing on feature selection via relevance and redundancy analysis. Using correlation-based and ReliefF methods, key EEG features were ranked and optimized for classification with a backpropagation neural network. The results achieved a 93% overall detection rate with a false seizure detection rate of 1.17/h. The approach enhances neonatal seizure detection by accounting for age-specific EEG characteristics. Khan et al. [35] also proposed a wavelet-based seizure detection algorithm using DWT and a normalized coefficient of variation (NCOV).

This compilation highlights various approaches, techniques, and achievements in the EEG-based seizure detection domain, shedding light on their potential and limitations in clinical applications. Upon reviewing the existing literature, it becomes evident that there are significant limitations in the techniques and methodologies employed. Many studies

have proposed various approaches, yet several common shortcomings have been identified. Existing literature reveals limitations in feature extraction, selection methods, classification algorithms, and validation in clinical settings. Our study aims to bridge these gaps comprehensively by conducting a thorough feature extraction, implementing rigorous feature selection techniques, exploring hybrid feature selection methods, and utilizing SMOTE for data balancing. Table 1 presents a concise overview of the various feature selection techniques utilized in the reviewed studies. These techniques are essential for improving the effectiveness of ML models in detecting epileptic seizures from EEG data by identifying and extracting the most pertinent features from the signals.

Table 1. Feature Selection Techniques used in related papers.

Citation	SVM	PCA	DWT	RF	MLPNN	PE	NB	KNN	GWO	ELM
Shoeb [13]	✓	-	-	-	-	-	-	-	-	-
Usman et al. [14]	-	✓	-	-	-	-	-	-	-	-
Donos et al. [15]	-	-	-	✓	-	-	-	-	-	-
Anurag and Sisodia [16]	-	✓	✓	✓	-	-	-	-	-	-
Nahzat and Yağanoğlu [17]	-	✓	-	-	-	-	-	-	-	-
Guo et al. [19]	-	-	✓	-	✓	-	-	-	-	-
Nicolaou and Georgiou [20]	-	-	-	-	-	✓	-	-	-	-
Wang et al. [21]	✓	-	✓	-	-	-	-	-	-	-
Sharmila and Geethanjali [7]	-	-	✓	-	-	-	✓	✓	-	-
Hamad et al. [22]	✓	-	-	-	-	-	-	-	✓	-
Song and Liò [23]	-	-	-	-	-	-	-	-	-	✓
Martis et al. [24]	✓	✓	-	-	✓	-	-	-	-	-
Farooq et al. [25]	✓	-	-	-	-	-	-	✓	-	-
Garima et al. [26]	-	-	-	✓	-	-	-	-	-	-
Shoeb and Guttag [27]	✓	-	-	-	-	-	-	-	-	-
Pinto-Orellana and Cerqueira [28]	-	-	-	✓	-	-	-	-	-	-
Khurshid et al. [29]	-	-	-	-	✓	-	-	-	-	-
Raghu et al. [30]	-	-	✓	-	✓	-	-	-	-	-
Zabihi et al. [31]	-	✓	-	-	-	-	✓	-	-	-
Krishnan and Balasubramanian [32]	✓	-	-	-	-	-	-	-	-	-
Chen et al. [33]	✓	-	✓	-	-	-	-	-	-	-
Aarabi et al. [34]	-	-	-	✓	✓	-	-	-	-	-
Khan et al. [35]	✓	-	-	-	-	-	-	-	-	-

4. Proposed Epileptic Seizures Detection Model

Existing approaches relying on the visual inspection of EEG recordings by medical professionals often lead to delayed identification of early seizure indicators. This can result in subjective assessments and the potential for oversight [11]. Therefore, this paper aims to revolutionize the detection of epileptic seizures by integrating advanced ML methodologies. Epileptic seizures, characterized by erratic brain activity, pose a significant health concern requiring accurate and timely identification for effective intervention. To address this challenge, this research develops an automated and objective seizure detection system using sophisticated ML techniques. The goal is to substantially improve the accuracy and timeliness of seizure detection compared with conventional methods.

The significance of this research lies in its potential to greatly enhance the quality of life for individuals with epilepsy by reducing the risks associated with delayed or inaccurate seizure detection. Leveraging ML models holds the promise of improving diagnostic accuracy while streamlining the detection process. The proposed model tackles the issue of accurate epileptic seizure prediction through the integration of effective feature selection techniques and classification methods. This involves enhancing prediction accuracy while minimizing false detections. To achieve this, we employ PCA and DWT as hybrid

feature selection methods, chosen for their demonstrated effectiveness across various applications.

The integration of PCA and DWT as a hybrid feature selection technique enhances the discriminatory power of extracted features for epileptic seizure detection. Unlike traditional methods that either rely on time domain or frequency domain analysis separately, this approach effectively captures both spatial and temporal patterns in EEG signals. PCA reduces dimensionality while preserving essential seizure-related information, whereas DWT decomposes EEG signals into multiple frequency sub-bands to identify seizure-specific characteristics. This dual approach ensures that the most relevant features are retained, leading to a more efficient and accurate classification process. By systematically addressing feature redundancy and computational overhead, the proposed method offers a scalable and adaptable solution suitable for real-time seizure detection across diverse patient datasets.

Existing approaches relying on the visual inspection of EEG recordings by medical professionals often lead to delayed identification of early seizure indicators. This can result in subjective assessments and the potential for oversight [11]. The significance of this research lies in its potential to greatly enhance the quality of life for individuals with epilepsy by reducing the risks associated with delayed or inaccurate seizure detection. Leveraging ML models holds the promise of improving diagnostic accuracy while streamlining the detection process. The proposed model tackles the issue of accurate epileptic seizure prediction through the integration of effective feature selection techniques and classification methods. This involves enhancing prediction accuracy while minimizing false detections. To achieve this, we employ PCA and DWT as hybrid feature selection methods chosen for their demonstrated effectiveness across various applications.

While several studies have reported high classification accuracies using EEG signals, these results are often achieved under constrained or idealized conditions, such as balanced datasets, patient-specific models, or limited validation protocols. Such settings do not reflect real-world clinical environments, where seizure data is scarce, highly imbalanced, and heterogeneous across patients. Moreover, prior models tend to rely on either time domain or frequency domain features in isolation, which limits their ability to capture the full dynamics of EEG signals. This often results in overfitting and poor generalization of new, unseen data. Additionally, the use of imbalanced datasets without robust oversampling strategies reduces sensitivity to rare seizure events, failing to address the clinical priority of minimizing missed detection.

In contrast, our proposed approach introduces a novel and holistic pipeline that addresses these persistent limitations. By integrating SMOTE solely on the training set, we generate synthetic minority-class instances in a way that avoids data leakage and enhances generalization. The hybrid use of PCA and DWT combines dimensionality reduction with detailed time–frequency feature extraction, enabling the model to capture subtle and complex patterns associated with seizures. This fusion of feature spaces provides a more discriminative and noise-resilient representation than either method alone. Finally, the use of SVM, known for its robustness in high-dimensional spaces, allows for effective classification even under challenging data conditions. Together, this SMOTE + PCA + DWT + SVM combination offers a novel, clinically relevant, and generalizable solution to the unresolved challenges in seizure detection—bridging the gap between research accuracy and real-world utility.

Furthermore, this research makes a significant contribution by evaluating multiple ML classifiers (SVM, RF, DT, and KNN) to determine the optimal model for seizure prediction. Unlike prior studies that focus on patient-specific models, which often lack generalizability, this work emphasizes the development of a robust, patient-independent detection system. The proposed approach incorporates advanced data preprocessing

techniques, including outlier removal and dataset balancing, with the Synthetic Minority Over-Sampling Technique (SMOTE), to mitigate bias in training models. By streamlining the detection process and improving diagnostic reliability, this work represents a substantial advancement in EEG-based seizure detection, with the potential for real-world clinical deployment and integration into wearable seizure monitoring systems.

The workflow of the proposed model is depicted in Figure 2. It begins with the collection of EEG data, followed by preprocessing to handle missing values and balance the dataset. Feature extraction and dimensionality reduction are performed using the hybrid approach of DWT and PCA. The processed features are then fed into the selected classification algorithms (SVM, RF, DT, and KNN) to predict seizure events. This comprehensive approach ensures the model is well-equipped to handle the complexities of EEG data, providing accurate and reliable seizure detection. By systematically integrating these advanced techniques, the proposed model represents a significant advancement in the field of EEG-based epileptic seizure detection.

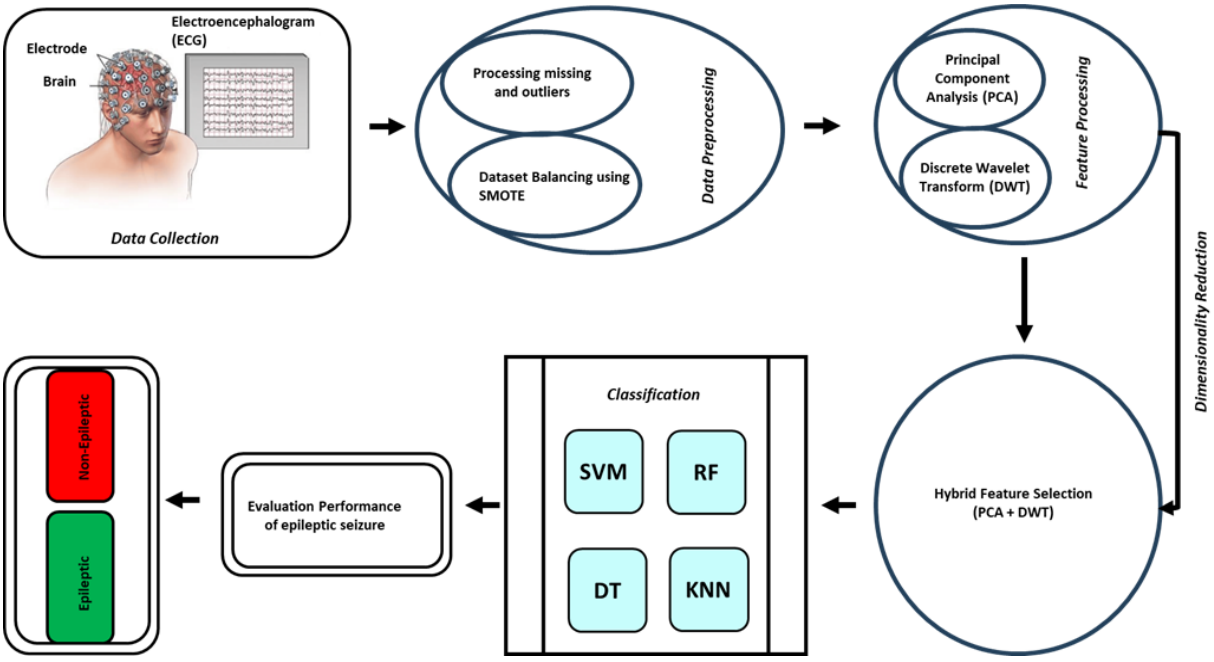

Figure 2. Proposed Epileptic Seizures Detection Model.

4.1. Data Collection—EEG Dataset

This research utilized the UCI Epileptic Seizure Recognition Dataset [1], a publicly available dataset widely acknowledged for its relevance and diversity in EEG recordings. This dataset served as the cornerstone for both the development and evaluation phases of our seizure detection models. Its comprehensive collection of EEG signals enabled the robust experimentation and validation of various ML algorithms tailored for seizure detection. The UCI Epileptic Seizure Recognition Dataset is structured to facilitate systematic analysis and interpretation of EEG data. It comprises five distinct folders, each representing a unique category delineating different physiological states. Within each category, there are precisely 100 files, each containing the EEG recordings of an individual.

The EEG recordings within these files are standardized to a duration of 23.6 s, providing a consistent temporal frame for analysis. These recordings are sampled into 4097 data points, ensuring granularity in capturing EEG patterns. To facilitate analysis at a finer temporal resolution, these 4097 data points are further segmented into 23 chunks, each spanning 1 s and containing 178 data points. Overall, the dataset encompasses EEG

recordings from 500 individuals, resulting in a total of 11,500 rows or instances of recorded data. These instances capture the EEG profile at various moments in time, offering insights into the dynamic nature of epileptic seizures and serving as the foundation for our study's analysis and model development. This dataset's structured organization and rich information content make it an invaluable resource for researchers seeking to investigate epileptic seizure detection using EEG signals. Its utilization in our study enabled rigorous experimentation and validation, contributing to the advancement of seizure detection methodologies.

The structure and composition of the UCI Epileptic Seizure Recognition Dataset comprises 5 distinct directories, each containing 100 files. Each file represents the recorded brain activity of a single subject, captured over a precisely timed duration of 23.6 s. The EEG recordings within each file consist of 4097 data points, capturing brain activity at discrete moments. These data points collectively represent observations from 500 individuals, where each person's recording spans 4097 data points corresponding to 23.5 s of EEG data. To effectively manage the sequential EEG data, the dataset is partitioned into 23 segments, each containing 178 data points, representing precisely 1-s intervals. Each data point within these segments signifies an EEG recording captured at distinct time instances.

As a result, the structured dataset comprises 11,500 rows, with each row consisting of 178 data points representing 1-s EEG intervals. The final column, positioned as the 179th column, serves as the 'y' label, indicating values 1, 2, 3, 4, 5. Here, 'y' represents the response variable, while the explanatory variables range from X1 to X178, facilitating the training and preparation of the dataset for subsequent modelling and analysis stages. In statistical modelling and ML contexts, the response variable 'y' is also known as the dependent variable, target variable, or outcome variable. The explanatory variables (X1 to X178) are commonly referred to as independent variables, predictor variables, features, or attributes in the dataset. The random illustration depicting a sample view of the Epileptic Seizure Recognition Dataset is represented in Figure 3.

	Unnamed	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
0	X21.V1.791	135	190	229	223	192	125	55	-9	-33	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	4
1	X15.V1.924	386	382	356	331	320	315	307	272	244	...	164	150	146	152	157	156	154	143	129	1
2	X8.V1.1	-32	-39	-47	-37	-32	-36	-57	-73	-85	...	57	64	48	19	-12	-30	-35	-35	-36	5
3	X16.V1.60	-105	-101	-96	-92	-89	-95	-102	-100	-87	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
4	X20.V1.54	-9	-65	-98	-102	-78	-48	-16	0	-21	...	4	2	-12	-32	-41	-65	-83	-89	-73	5

5 rows × 180 columns

Figure 3. A sample of the Epileptic Seizure Recognition Dataset [1].

The structured organization of the dataset ensures that each instance captures the EEG profile at various moments in time, offering insights into the dynamic nature of epileptic seizures. The use of this dataset enabled rigorous experimentation and validation, contributing to the advancement of seizure detection methodologies. Preprocessing steps, including Binarization and the application of SMOTE, were applied to enhance data quality and rectify imbalances within the dataset, ensuring robustness in the models developed. This step was critical in ensuring that the classifier did not develop a bias toward the majority classes.

SMOTE was used to synthetically generate new instances only for the training set after the data had been split by subject to preserve patient independence and prevent data leakage into the test set. This approach ensured that no artificially generated samples influenced the model's evaluation on unseen data.

The EEG waveforms illustrated in Figure 4 are not specific to a single electrode or channel. Instead, they represent the average signal of all EEG data points across all samples within each class, computed using the average of the EEG vectors (X1 to X178) along the time axis. This approach highlights general waveform patterns typical of each class (e.g., ictal vs. healthy). The aim is to visually contrast the distinct temporal characteristics of the five EEG signal classes. These signals were originally captured using a single electrode, likely from the Fpz-Cz or Pz-Oz positions, although the exact montage is not specified in the dataset documentation. Therefore, multichannel data is not available in this version of the dataset; rather, each file contains one-dimensional EEG signals representing one channel. The five classes represent distinct brain states:

1. Class 1 (ictal)—EEG signals recorded during active epileptic seizures;
2. Class 2 (pre-ictal)—EEG signals recorded shortly before the onset of a seizure;
3. Class 3 (inter-ictal)—EEG signals between seizures (non-seizure periods);
4. Class 4 (healthy with eyes closed)—Baseline EEG from healthy individuals;
5. Class 5 (healthy with eyes open)—Baseline EEG from healthy individuals under normal alert conditions.

No behavioural tasks were conducted during EEG acquisition for the healthy participants. The healthy datasets were collected under resting-state conditions with eyes either open or closed. For seizure-related recordings, the original dataset documentation does not specify whether participants were under medication, the exact seizure type, or the environment during recording. As such, while this dataset supports classification tasks effectively, it has limitations in clinical metadata and channel diversity.

Each class represents distinct EEG signal patterns associated with specific physiological states or seizure types, as shown in Table 2. By visualizing these waveforms, we can observe the characteristic electrographic features of epileptic seizures and differentiate them from normal EEG patterns. This comparison aids in identifying unique signatures of epileptic activity, facilitating accurate seizure detection and classification.

The minority classes in the dataset were identified based on their original class distribution. Specifically, after segmentation and before balancing, the five classes—ictal, pre-ictal, inter-ictal, healthy (eyes closed), and healthy (eyes open)—were observed to have equal counts in the raw dataset. However, due to stratified subject-level train-test splitting, imbalances emerged in the training subset, particularly affecting seizure-related classes. SMOTE was therefore selectively applied to these underrepresented classes.

This targeted application of SMOTE improved the model's sensitivity to minority seizure classes while maintaining data integrity and reproducibility. The use of SMOTE only on the training set ensures that evaluation metrics reflect the model's true generalizability to real-world, imbalanced data scenarios.

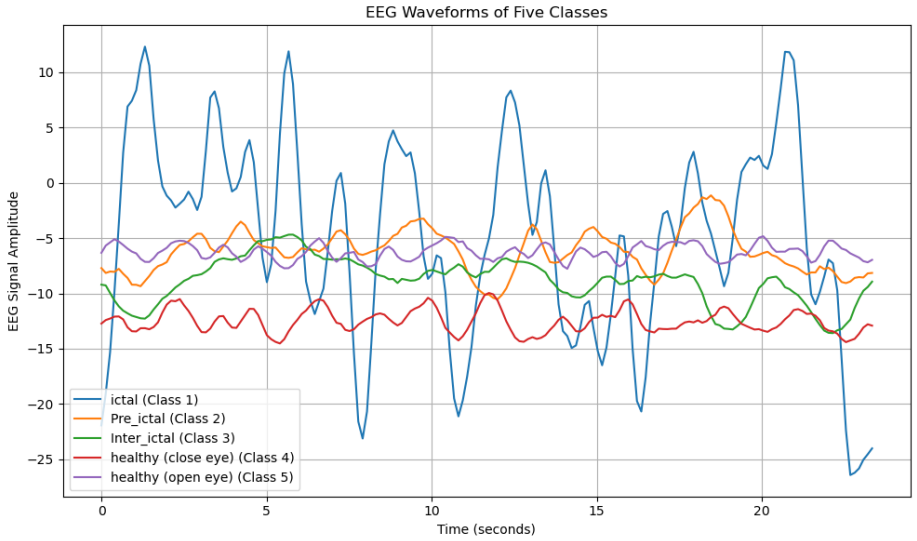

Figure 4. Average EEG waveforms for each signal class in the dataset: ictal, pre-ictal, inter-ictal, healthy (closed eyes), and healthy (open eyes). The plotted signals are the average of 178-sample EEG vectors across all instances in each class, providing a visual comparison of their distinct temporal features.

Table 2. Summary of EEG Signal classes and their descriptions [1].

Classes	Name	No of Samples	Output Classes (Labels)	Description of Classes
1	Ictal	2300	1	Signals recorded during seizures
2	Pre-ictal	2300	2	Signals recorded before the occurrence of a single seizure
3	Inter-ictal	2300	3	Signals recorded during the occurrence of consecutive seizures
4	Healthy (closed eyes)	2300	4	A healthy subject with closed eyes
5	Healthy (open eyes)	2300	5	A healthy subject with open eyes

4.2. Data Preprocessing

In this study, the following EEG preprocessing steps were applied in chronological order to prepare the dataset for ML analysis:

- **Data Import and Inspection**: The UCI Epileptic Seizure Recognition Dataset was loaded using Python (Pandas 2.2.3) and examined using '.head()', '.tail()', '.describe()', and '.info()' to ensure data integrity;
- **Feature Selection**: The 178 EEG signal features (X1 to X178) were extracted while excluding the subject identifier column. The label column ('y') was retained as the target variable;
- **Segmentation Review**: EEG signals in the dataset are already segmented into 1-s intervals of 178 points, offering consistency in sample size and reducing preprocessing needs;
- **Basic Data Cleaning**: No missing values or corrupted rows were found during inspection; hence no imputation was necessary;
- **SMOTE Balancing**: To address class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to balance the number of instances in each class before training.

It is important to note that the UCI dataset used in this study consists of pre-processed, artifact-free EEG recordings as confirmed in the dataset documentation. Therefore, additional steps, such as re-referencing, bandpass filtering, artifact rejection, and channel interpolation, were not required or applied. This decision aligns with evaluating

classifiers on clean, controlled EEG samples to establish baseline performance. However, future work will include experiments on raw EEG with full clinical preprocessing pipelines.

4.3. Feature Selection Techniques

Feature selection is pivotal in identifying informative EEG features while reducing dimensionality, which is crucial for improving the performance and interpretability of ML models. Our approach started with preprocessing steps, including Binarization and SMOTE to address class imbalance, followed by splitting the dataset into training and testing sets. PCA and DWT emerged as effective feature selection methods from our extensive review of previous studies. Each technique offers distinct advantages, which influenced our decision to incorporate them into our proposed model.

4.3.1. Principal Component Analysis (PCA)

PCA is a statistical procedure that transforms a set of correlated variables into a set of uncorrelated components ranked by the amount of variance they explain in the data. This technique is widely used for dimensionality reduction because it simplifies the complexity of the dataset while retaining the most critical information. We applied PCA for dimensionality reduction, selecting 50 components based on cumulative explained variance analysis. This approach ensures that the majority of the dataset's variance is captured with fewer features, thus simplifying the model and enhancing its computational efficiency. The decision to use PCA is rooted in its ability to mitigate the curse of dimensionality and improve model performance by reducing overfitting and enhancing generalization.

The selection of 50 principal components in PCA was guided by a cumulative explained variance analysis, which revealed that these components captured approximately 95% of the total variance present in the dataset. This threshold was chosen to ensure that most of the meaningful information in the original 178 EEG features was retained while reducing dimensionality to improve computational efficiency and model performance. This 95% cut-off is widely accepted in EEG-based machine learning studies, as it balances the trade-off between retaining discriminative signal features and eliminating redundant or noisy dimensions.

4.3.2. Discrete Wavelet Transform (DWT)

DWT is a powerful signal processing technique that decomposes EEG signals into different frequency components, allowing for the analysis of both time and frequency characteristics. This method is particularly effective in capturing the transient features of EEG signals, which are essential for detecting epileptic seizures. DWT generates coefficients that serve as significant features, capturing the nuances of EEG signal variations. The choice of DWT is motivated by its proven efficacy in retaining the most relevant features of non-stationary signals like EEG, which exhibit complex patterns and require sophisticated techniques to extract meaningful information.

4.3.3. Hybrid Feature Selection Approach (PCA + DWT)

To leverage the strengths of both PCA and DWT, we introduced a hybrid feature selection approach by merging PCA-transformed features with selected DWT coefficients. This hybrid approach aims to harness the dimensionality reduction capabilities of PCA and the detailed signal feature extraction of DWT. By combining these methods, we aim to create a robust feature set that maximizes the informative content while minimizing redundancy and noise. This dual strategy is designed to enhance the predictive accuracy and efficiency of the seizure detection model.

The selection of PCA and DWT, both individually and in combination, is based on their complementary strengths and established effectiveness in previous research. PCA's ability to reduce dimensionality and focus on the most significant components complements DWT's capability to capture essential time–frequency features of EEG signals. This synergy is crucial for creating a model that is not only accurate but also efficient and scalable. Moreover, the hybrid approach addresses the limitations of using either technique in isolation, providing a comprehensive feature set that improves model robustness. In summary, the application of PCA and DWT in our feature selection process is a strategic choice aimed at optimizing the performance of our ML models. By systematically reducing dimensionality and extracting key signal features, we enhance the model's ability to accurately detect epileptic seizures, ultimately contributing to better clinical outcomes and improved quality of life for individuals with epilepsy. This is shown in Figure 5.

```
# Start

# Step 1: Preprocessing Steps
# Binarization of target variable 'y'
Binarize target variable 'y'
Apply SMOTE for class balancing

# Step 2: Splitting the Dataset
# Split pre-processed dataset into training and
    testing sets (80:20 ratio)
Split data (80:20)

# Step 3: Feature Selection Techniques
# Perform PCA for Dimensionality Reduction:
Perform PCA for dimensionality reduction
Determine the optimal number of components using
    cumulative explained variance analysis
Set the number of components to 50 for both
    training and testing datasets

# Utilize DWT for Signal Feature Extraction:
Extract signal features from EEG data using
    Discrete Wavelet Transform

# Step 4: Hybrid Feature Selection (PCA + DWT)
# Merge PCA-transformed features and selected DWT
    coefficients:
Merge PCA-transformed features with DWT
    coefficients
Create a hybrid feature set combining PCA-
    transformed features and DWT coefficients

# End
```

Figure 5. Pseudocode for preprocessing, dataset splitting, and feature selection techniques.

The DWT was applied to each 1-s EEG segment (178 data points) to extract meaningful time–frequency domain features that capture transient patterns characteristic of epileptic activity. The DB4 wavelet was selected due to its effectiveness in EEG signal analysis, particularly in identifying seizure-related features. DWT was implemented using three levels of decomposition, which allowed the signal to be separated into

approximation and detail coefficients at different frequency bands. The decomposition levels were chosen based on the length of the signal and the sampling frequency (173.61 Hz), ensuring that the frequency bands covered typical EEG rhythms (delta to beta); details are shown in Table 3.

Table 3. Parameters and settings used for DWT algorithm.

Parameter	Value	Description
Wavelet Type	Daubechies 4 (db4)	Selected for its suitability in analyzing EEG signals with transient behaviour.
Number of Decomposition Levels	3	Allows decomposition into relevant EEG frequency bands (delta to beta).
Signal Length	178 data points per 1-s segment	Based on UCI dataset sampling rate (173.61 Hz).
Sampling Frequency	173.61 Hz	Defines the effective frequency resolution for DWT.
Features Used	Detail coefficients (D1, D2, D3)	Captures high-frequency signal changes linked to seizure events.
Implementation Library	PyWavelets (PyWT)	Python (Pandas 2.2.3) library used for wavelet transformation and feature extraction.

To support the hybrid use of PCA and DWT, we emphasize that each technique captures complementary aspects of EEG signals. DWT is well-suited for analyzing non-stationary signals by decomposing them into time–frequency components, enabling the extraction of transient features associated with seizure onset and progression. However, the resulting coefficient set may contain redundancy and noise, especially when multiple decomposition levels are used. PCA, as a linear dimensionality reduction method, is then applied not to extract new features but to refine the DWT-derived features by removing collinear and low-variance components. This enhances the discriminative quality of the feature space while retaining dominant seizure-related patterns.

Although PCA is a global transformation and does not preserve time-localized structure in raw signals, its application after DWT is beneficial because the DWT step has already captured localized features across different frequency bands. The PCA transformation then acts as a post-processing filter that emphasizes variance-rich patterns while reducing dimensionality and noise. Empirically, our results show that the hybrid approach outperforms either method alone in terms of classification accuracy, precision, and recall. This indicates that PCA and DWT are synergistic rather than conflicting in this context, where PCA acts as a feature refiner rather than a suppressor of time–frequency information.

4.4. Classification Techniques

Classification techniques are instrumental in categorizing EEG data into seizure and non-seizure classes, which is a critical step in the development of accurate and reliable seizure detection systems. We adopted four commonly used algorithms—SVM, DT, RF, and KNN—based on their demonstrated effectiveness and prevalence in related studies. Each of these algorithms brings unique strengths to the table, making them well-suited for our analysis.

4.4.1. Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm known for its effectiveness in high-dimensional spaces and its robustness against overfitting, especially in cases where the number of dimensions exceeds the number of samples. SVM works by finding the optimal hyperplane that separates the data into different classes with the maximum margin. This property is particularly useful in EEG data classification, where the distinction between seizure and non-seizure events can be subtle. SVM's ability to handle complex, non-linear boundaries through kernel functions further enhances its suitability for our task. Therefore, SVM stands out for its margin-maximizing decision boundaries, and kernel-based flexibility makes it highly suitable for sparse and imbalanced datasets. Additionally, SVM-based methods have demonstrated reliable performance across a variety of EEG domains [36]. For instance, in decoding motor intentions for predicting movement directions from EEG signals, SVM is employed to tackle noisy data while maintaining robustness. Similar to this, in brain–machine interfaces [37] SVM has been successfully applied to classify moment intention. The choice of SVM is driven by its high accuracy and ability to generalize from training data to unseen data.

4.4.2. Decision Tree (DT)

Decision Trees are intuitive and easy-to-interpret models that split the data into subsets based on the value of input features. Each node in the tree represents a feature, each branch represents a decision rule, and each leaf node represents an outcome. DTs are chosen for their simplicity and ease of visualization, which makes them an excellent tool for understanding the decision-making process of the model. Additionally, DTs are computationally efficient and can handle both numerical and categorical data, making them versatile for various types of EEG features. The interpretability of DTs is a significant advantage, as it allows for better insights into the factors contributing to seizure detection.

4.4.3. Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. This technique mitigates the overfitting problem associated with single decision trees by averaging multiple trees, thus improving generalization. RF's robustness and high accuracy make it a popular choice for EEG classification tasks. Its ability to handle large datasets with higher dimensionality, along with its built-in feature importance estimation, helps in identifying the most relevant features for seizure detection. The ensemble nature of RF ensures stable and reliable predictions, enhancing the overall performance of the model.

4.4.4. K-Nearest Neighbours (KNN)

KNN is a simple yet effective algorithm that classifies a data point based on how its neighbours are classified. It is a non-parametric method, meaning it makes no explicit assumptions about the form of the function mapping the inputs to the outputs. This makes KNN particularly flexible and easy to implement. The effectiveness of KNN in our context comes from its reliance on the local structure of the data, which can be advantageous in detecting patterns within EEG signals. The simplicity of KNN allows for quick and straightforward implementation, providing a baseline against which more complex models can be compared.

The selection of these four classification techniques is grounded in their proven track record and complementary strengths. SVM is chosen for its ability to handle high-dimensional data and its robustness against overfitting, making it suitable for complex EEG data. DT and RF are selected for their interpretability and ensemble capabilities, respectively,

which enhance model reliability and provide deeper insights into the feature importance. KNN is included for its simplicity and effectiveness in leveraging local data patterns. By presenting a visual depiction of the algorithmic process with flowcharts and pseudocode, particularly for SVM, we offer a clear, step-by-step illustration of how these techniques are implemented in our model. Similar approaches are applied to the other classification techniques, ensuring a comprehensive understanding of their application in seizure detection.

In summary, our choice of SVM, DT, RF, and KNN is driven by their individual strengths and their collective ability to provide a robust, accurate, and interpretable models for EEG-based epileptic seizure detection. This strategic selection aims to maximize detection accuracy while maintaining model simplicity and interpretability, ultimately contributing to more effective and reliable seizure detection systems. Our approach integrates advanced feature selection techniques and classification algorithms to develop a robust epileptic seizure detection model. Through careful implementation and evaluation, the aim is to enhance the accuracy and efficiency of seizure detection systems, ultimately improving patient outcomes in epilepsy management.

The classifiers used in this study were implemented using Scikit-learn (1.6.1). SVM was configured with a Radial Basis Function (RBF) kernel and enabled probability estimates. For DT, hyperparameter tuning was conducted using GridSearchCV with 10-fold cross-validation, exploring max_depth values of None, 5, 10, 15, and 20. The RF model was applied using Scikit-learn's default settings, including 100 estimators. The KNN classifier was optimized using GridSearchCV over the hyperparameters n_neighbors = [3, 5, 7, 9] and weights = ['uniform', 'distance'], with the best configuration found to be n_neighbors = 3 and weights = 'distance'. These configurations were selected to provide a balanced assessment of classifier performance. Although a full grid search across all models was not the primary goal of this study, fairness and consistency were ensured in preprocessing and evaluation protocols.

5. Experiment Setup and Metrics

This section describes the experimental setup, including the computing environment and patient-independent data splitting. It also defines the evaluation metrics used to assess the seizure detection model's performance.

5.1. Experiment Settings

Experiments were conducted on a high-performance desktop workstation running Windows 10 Professional, equipped with an Intel Core i7 processor, 16 GB of RAM, and a dedicated GPU, providing reliable performance for computational tasks, with Python as the primary programming language. Python libraries, such as NumPy (2.2.4), Pandas (2.2.3), Matplotlib (3.10.1), ScikitLearn (1.6.1), TensorFlow (2.19.0), and Keras (3.9.2), were utilized for data manipulation, model development, and evaluation.

To ensure the validity and generalizability of the proposed seizure detection model, a patient-independent classification protocol was adopted. This means that data segments from the same subject were not used in both the training and testing sets, thereby preventing data leakage and overly optimistic performance estimates.

As mentioned in Section 4.1, the dataset comprises 500 subjects, with each class containing EEG recordings from 100 distinct individuals. During preprocessing, each 23.6-s EEG recording was segmented into 23 1-s epochs, resulting in a total of 11,500 segments. To preserve patient independence, the data was split at the subject level rather than at the segment level. Specifically, 80% of the subjects were used for training, and the remaining 20% were reserved for testing. For example, if a subject's EEG file contributed 23 segments, all 23 segments were either in the training set or in the test set—never both. This

ensured that the classifier learned from EEG patterns of distinct individuals and was tested on completely unseen subjects.

This protocol better reflects real-world clinical scenarios, where seizure detection systems must generalize to new patients not seen during training. It also supports the claim of developing a patient-independent detection system as opposed to patient-specific models that often suffer from poor generalizability.

5.2. Evaluation Metrics

In assessing the models' performance, an array of metrics was employed to precisely measure the effectiveness and accuracy of the algorithms. These metrics encompass fundamental measures including accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC) derived from the Receiver Operating Characteristic (ROC) curve.

1. **Accuracy**: is a fundamental metric that measures the ratio of correctly predicted instances to the total number of instances in the dataset. A high accuracy score indicates the model's effectiveness in making correct predictions across all classes. It is computed using Equation (1):

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

2. **Area Under the ROC Curve (AUC):** The AUC metric is a valuable tool for assessing a model's ability to distinguish between classes. It is calculated by plotting the Receiver Operating Characteristic (ROC) curve and computing the area under this curve. A higher AUC value (ranging from 0 to 1) signifies better discrimination power of the model;
3. **Confusion Matrix**: offers a comprehensive tabular layout displaying True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. It aids in visualizing the model's performance across different classes, identifying misclassifications, and understanding the model's behaviour. These metrics collectively provide a thorough understanding of a classification model's performance, allowing for informed decisions and enhancements in model training and optimization strategies;
    - **True Positive (TP):** This represents the number of instances where the model correctly predicted the positive class (or the condition being tested) as positive. In medical terms, this could be the number of correctly identified patients with a disease;
    - **True Negative (TN)**: This represents the number of instances where the model correctly predicted the negative class (or the absence of the condition being tested) as negative. In medical scenarios, this could indicate the number of correctly identified healthy individuals who do not have the disease;
    - **False Positive (FP):** This indicates the number of instances where the model incorrectly predicted the positive class when it was actually negative. In medical scenarios, it could signify the number of healthy individuals incorrectly identified as having the disease;
    - **False Negative (FN):** This refers to the number of instances where the model incorrectly predicted the negative class when it was actually positive. In medical contexts, this might represent the number of individuals with the disease incorrectly identified as healthy by the model.
4. **F1 Score**: The F1 score represents the harmonic mean of precision and recall. The F1 score provides a balanced evaluation of a model's precision and recall, making it a robust metric for binary classification tasks. It is computed using Equation (2):

$$F1\ Score = 2\ \text{x}\ \frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

5. **Precision**: It measures the ratio of correctly predicted positive instances to the total instances predicted as positive. A high precision score indicates the model's ability to minimize false positives, ensuring that most of the predicted positive instances are relevant. It is computed using Equation (3):

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

6. **Recall**: It measures the ratio of correctly predicted positive instances to the actual total positive instances in the dataset. A high recall score indicates the model's effectiveness in capturing most of the relevant instances while minimizing false negatives. It is computed using Equation (4).

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

6. Experimental Results

The outcomes derived from employing four separate algorithms, SVM, DT, RF, and KNN, have been meticulously recorded. The performance metrics include accuracy, AUC, F1 score, and the average of these metrics. The results are summarized in Tables 4 and 5.

The application of data balancing and hybrid feature selection (PCA + DWT) techniques generally improved the performance of all four classifiers. The most notable improvements were observed in the SVM and RF models. Before the application of data balancing and hybrid feature selection techniques, the SVM and RF models already demonstrated strong classification capabilities, with SVM achieving an accuracy of 97.00% and an AUC of 99.40%, while RF followed closely with an accuracy of 96.30% and an AUC of 99.20%. In contrast, the DT and KNN models exhibited relatively lower performance, with DT showing an accuracy of 93.40% and an AUC of 81.50%, and KNN achieving an accuracy of 93.60% with an AUC of 90.30%. These initial results suggest that, while DT and KNN can serve as viable classifiers, they are less robust compared to SVM and RF.

Table 4. Initial Model Performance Without Application of Data Balancing.

Model	Accuracy	AUC	F1 Score
SVM	97%	99.4%	92.3%
DT	93.4%	81.50%	82.5%
RF	96.3%	99.2%	90.2%
KNN	93.6%	90.3%	80.6%

Table 5. Model Performance After Application of Data Balancing and Hybrid.

Model	Accuracy	AUC	F1 Score
SVM	97.3%	99.62%	93.08%
DT	94.48%	83.24%	85.87%
RF	97.17%	99.56%	92.72%
KNN	92.13%	88.48%	76.03%

After implementing data balancing and feature selection techniques, noticeable improvements were observed across most models. The SVM classifier experienced a marginal enhancement in all metrics, with accuracy increasing to 97.30%, AUC rising to 99.62%, and the F1 score improving from 92.30% to 93.08%. This indicates the model's

high reliability in distinguishing between seizure and no seizure events. Similarly, RF showed significant gains, achieving an accuracy of 97.17%, an AUC of 99.56%, and an F1 score of 92.72%, reinforcing its strong predictive capability.

The DT model demonstrated moderate improvements, with accuracy increasing from 93.40% to 94.48%, the AUC improving from 81.50% to 83.24%, and the F1 score rising from 82.50% to 85.87%. Despite these enhancements, DT continued to lag behind SVM and RF, suggesting potential overfitting issues and limitations in handling complex patterns. On the other hand, KNN displayed mixed results, with its accuracy declining from 93.60% to 92.13%, the AUC decreasing from 90.30% to 88.48%, and the F1 score dropping from 80.60% to 76.03%. This reduction in performance indicates that KNN may not be well-suited to high-dimensional data or imbalanced class distributions, highlighting its sensitivity to changes in feature selection and data balancing.

Further evaluation was conducted through confusion matrices, providing insights into each model's classification errors and overall predictive accuracy. The confusion matrices, as illustrated in Figure 6, reveal that SVM and RF exhibit the lowest misclassification rates, further validating their superiority in seizure detection. The DT and KNN models, however, displayed a higher rate of false positives and false negatives, suggesting their relative inefficiency in precise classification. Additionally, the ROC curve analysis, as shown in Figure 7, further substantiates the findings. The SVM classifier exhibited an outstanding AUC of 0.996, positioning it as the best-performing model alongside RF, which also achieved an AUC of 0.996. The DT classifier displayed moderate performance with an AUC of 0.832, while KNN achieved an AUC of 0.885, indicating reasonable but comparatively weaker discrimination ability. A comparative ROC curve analysis of all classifiers (Figure 8) underscores the dominance of SVM and RF over DT and KNN in distinguishing between seizure and non-seizure events.

The final performance comparison, visualized in Figure 9, consolidates these observations by illustrating the models' accuracy, AUC, and F1 score. The SVM and RF models consistently outperformed DT and KNN across all metrics. The superior performance of SVM can be attributed to its ability to find an optimal hyperplane that maximizes class separation, making it particularly effective for high-dimensional data such as EEG signals. RF, benefiting from its ensemble learning approach, successfully reduces overfitting and captures complex patterns, contributing to its high predictive power.

In summary, the application of data balancing and hybrid feature selection techniques resulted in overall performance improvements, with SVM and RF emerging as the most effective models for seizure detection. DT showed moderate gains but remained less reliable, while KNN exhibited performance degradation, highlighting its limitations in handling high-dimensional, imbalanced data. These findings provide valuable insights into the selection of ML models for EEG-based seizure detection, with SVM and RF proving to be the most robust and efficient classifiers in this context.

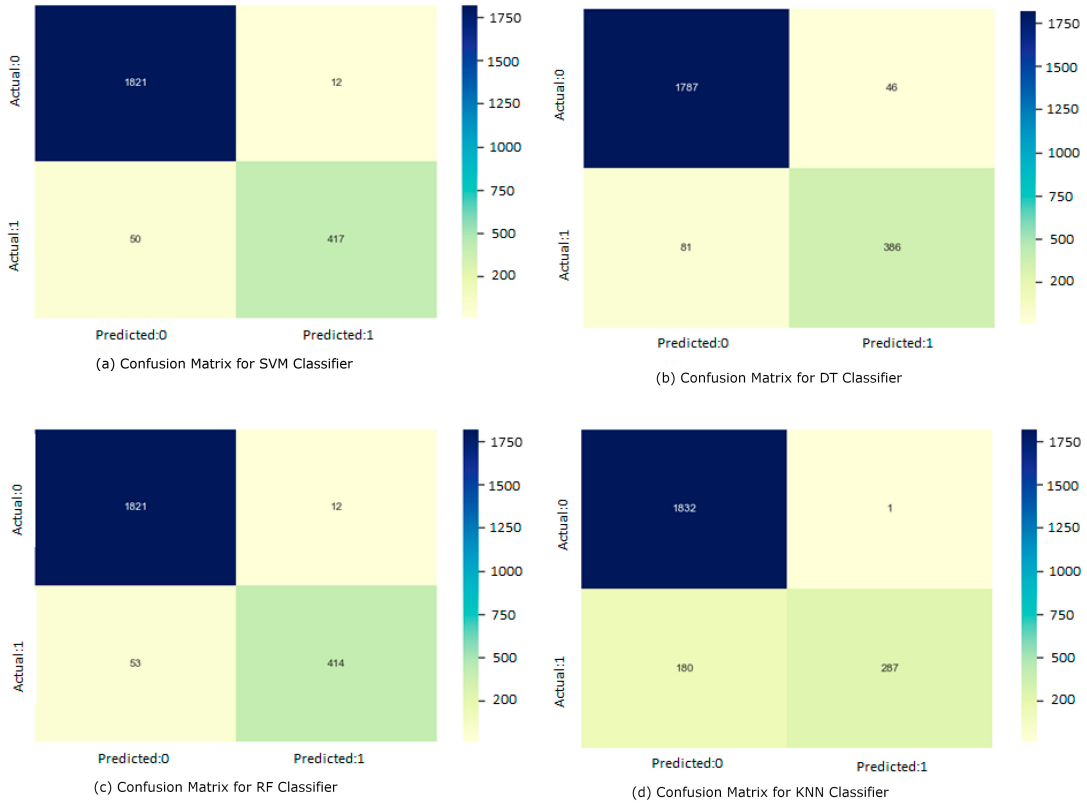

Figure 6. Confusion matrices for the evaluated classifiers (SVM, Random Forest, Decision Tree, and KNN). Rows represent true classes, and columns represent predicted classes. Class 0 corresponds to non-seizure activity, and Class 1 represents seizure events.

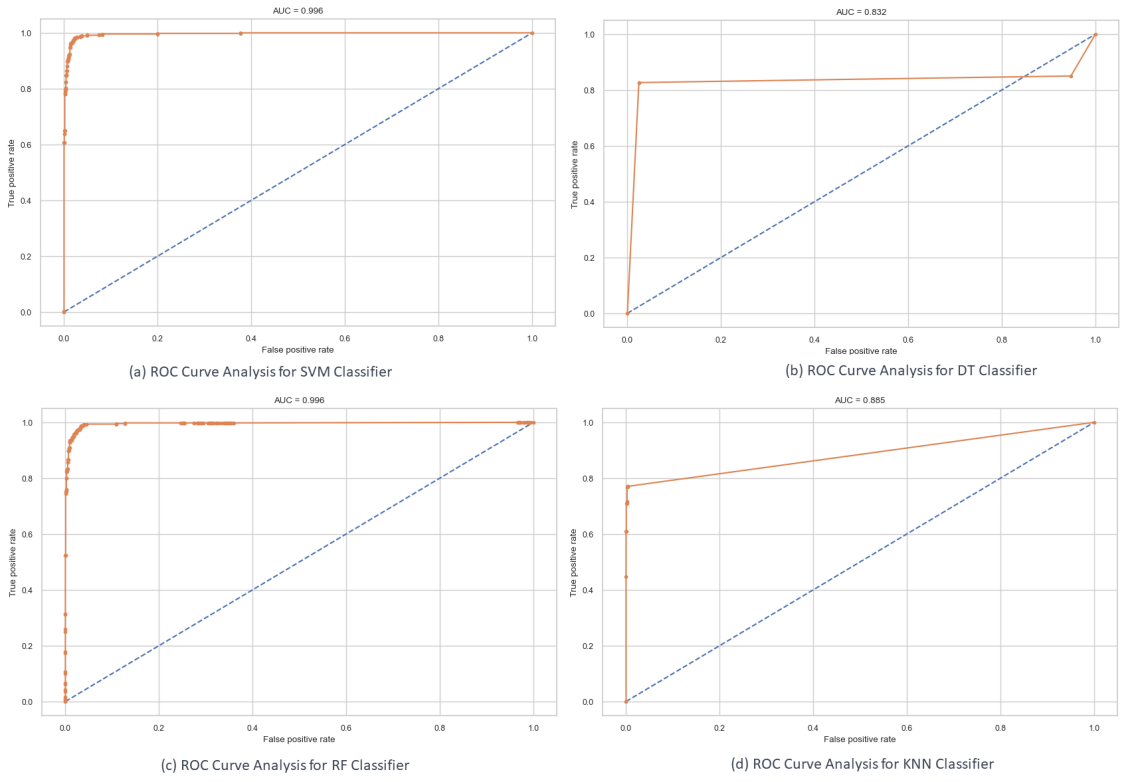

Figure 7. ROC for each Classifier.

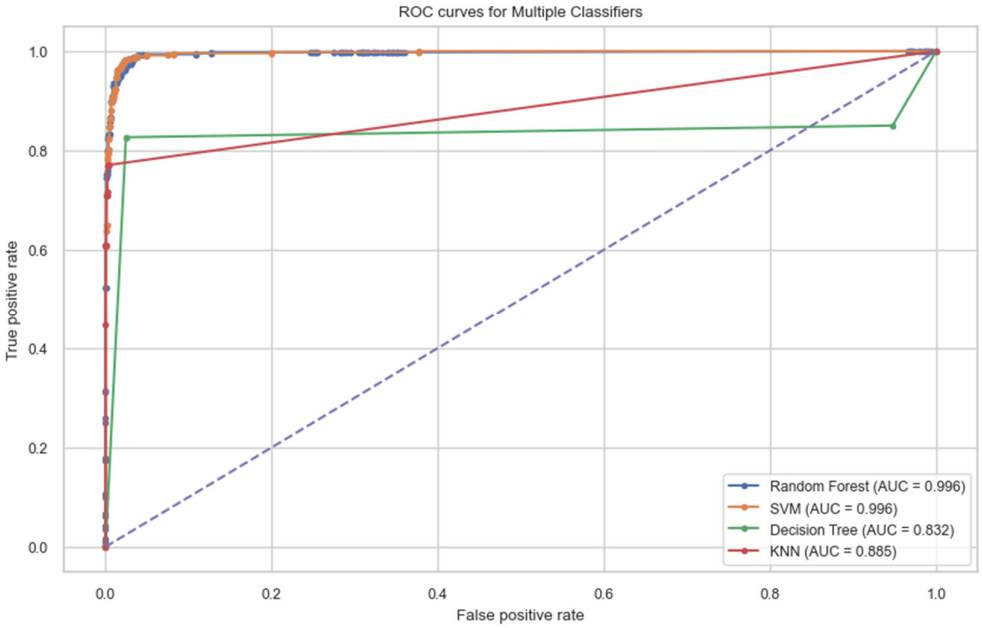

Figure 8. Comparison of ROC Curves for all Classifiers.

The evaluation of our proposed model reveals substantial advancements in seizure detection accuracy through EEG signal analysis. By integrating the SMOTE for data balancing and employing a hybrid feature selection approach combining PCA and DWT, our model demonstrates remarkable improvements across several algorithms, including SVM, RF, DT, and KNN.

Our proposed model demonstrates significant performance improvements in seizure detection through the integration of SMOTE and hybrid feature selection techniques. The SVM and Random Forest models achieved the highest accuracy (97.30% and 97.17%, respectively) and near-perfect AUC scores (0.996), showcasing their superior classification capabilities. The Decision Tree model exhibited moderate improvements, while the KNN model experienced slight declines in accuracy and F1 score, indicating challenges with high-dimensional data. The ROC curve analysis confirmed the robustness of SVM and Random Forest, with high true positive rates and minimal false positives. These findings highlight the effectiveness of our approach in enhancing EEG signal analysis and seizure detection, setting a new benchmark for future research and clinical applications.

Our proposed model's performance has been benchmarked against a range of recent studies using the Bonn (UCI) EEG dataset, as shown in Table 6. The results demonstrate meaningful improvements in both classification accuracy and recall, validating the effectiveness of our hybrid approach. Krishnan and Balasubramanian [32] developed an EEG seizure detection method based on time–frequency entropy and an LSSVM classifier, achieving 86% accuracy and 68% recall. Similarly, Chen et al. [33] employed optimized DWT-based features for seizure localization, reporting 88% accuracy and 91.52% recall. In contrast, our model integrates both DWT and PCA for a richer feature representation, improving overall accuracy to 97.3%.

More recent studies have aimed to improve performance using various classifiers and preprocessing strategies. Wang et al. [38] presented a hardware-oriented multiclass SVM system, achieving 93.9% accuracy and 94.7% recall, whereas Kabir et al. [39] utilized logistic model trees, with a performance of 95.33% accuracy and 95% recall. Wang et al. [40] proposed a symlet wavelet-based pipeline combined with gradient boosting, reaching 96.5% accuracy and 95.8% recall. While these studies show strong performance, our model

exceeds all in classification accuracy, while maintaining competitive recall. Our model outperforms this by leveraging a more comprehensive hybrid feature selection process and multiple classifiers (SVM, RF, DT, KNN), ultimately enhancing predictive performance. These comparisons collectively highlight the advancements introduced by our methodology, particularly through effective feature selection, dimensionality reduction, and data balancing techniques, leading to a more reliable and accurate seizure detection system.

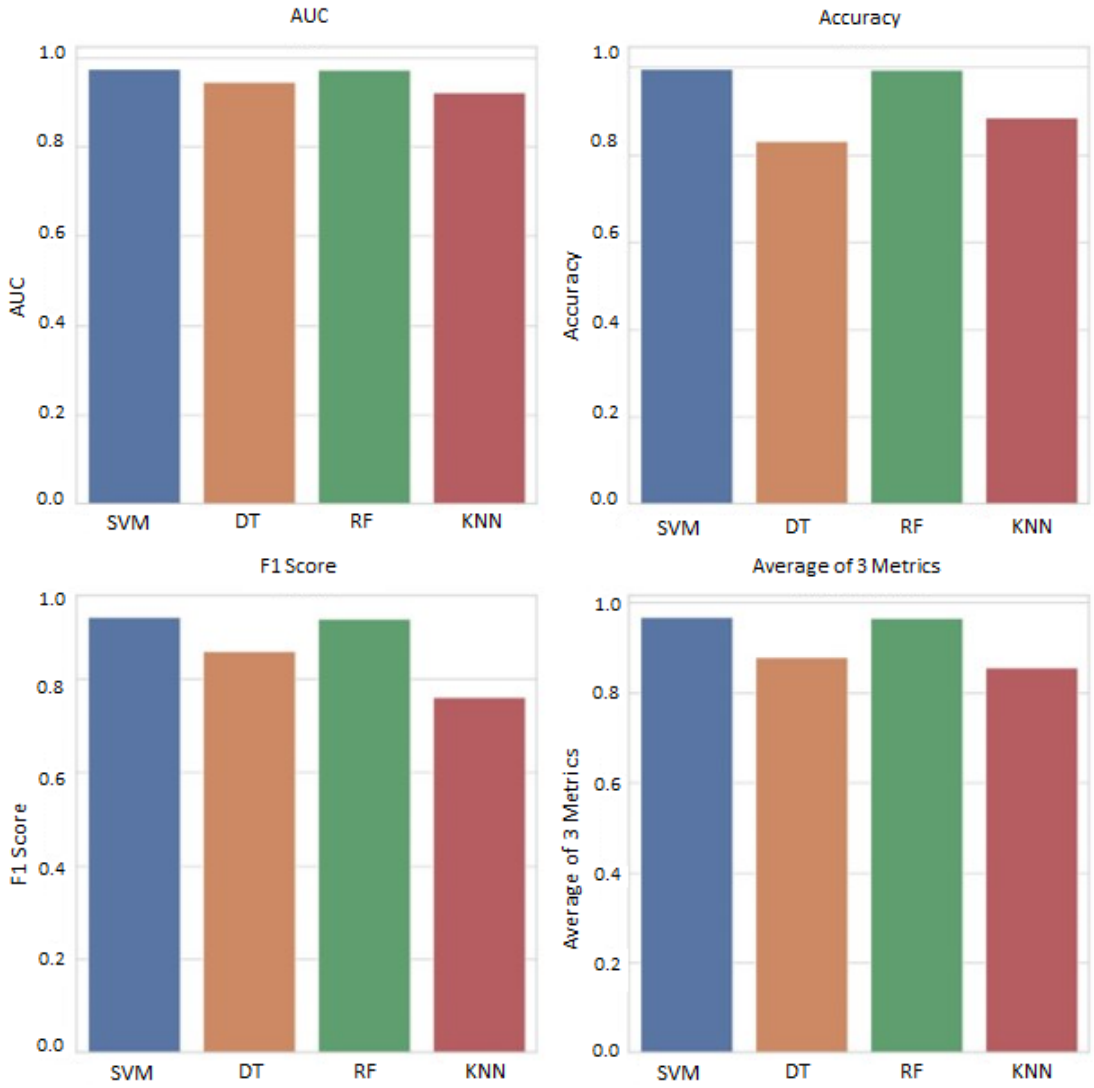

Figure 9. Performance Metrics Comparison Across Models.

Our findings contribute to the ongoing advancements in EEG signal analysis and seizure detection technologies. The integration of SMOTE and hybrid feature selection techniques offers a robust framework for improving seizure detection accuracy. By enhancing model performance and reliability, our approach sets a new benchmark for future research and clinical applications in this field, as shown in Table 6. The improvements in accuracy and performance metrics have significant implications for clinical practice. Enhanced seizure detection systems can lead to more reliable and timely identification of seizures, contributing to better patient management and treatment outcomes. The robustness of SVM and RF in our study suggests that these models could be effectively deployed in clinical settings to assist in real-time seizure monitoring and diagnosis.

Table 6. Evaluation of the proposed work against existing work from the literature.

Citation	Accuracy	Recall
Krishnan and Balasubramanian [32]	86%	68%
Chen et al. [33]	88%	91.52%
Wang et al. [38]	93.9%	94.7%
Kabir et al. [39]	95.33%	95%
Wang et al. [40]	96.5%	95.8%
Proposed Model	**97.3%**	**93.08%**

While our study demonstrates promising results, there are some limitations. First, the proposed model was evaluated on a single, publicly available dataset, which contains pre-segmented and artifact-free EEG signals recorded under controlled conditions. This limits the ecological validity of the findings, as the model may not perform equally well on raw, noisy, or multi-channel EEG data collected in clinical settings. Additionally, the use of a single-channel signal restricts the ability to capture spatial features across brain regions, which are often informative in seizure detection. These constraints may affect the model's generalizability across different patient demographics, hardware configurations, and seizure types. Also, the current model does not yet support real-time processing or adaptive learning, which are essential for integration into wearable or bedside monitoring systems. Moreover, although the hybrid feature selection method improves accuracy, it may introduce computational overhead that limits scalability in low-power or embedded environments. Future work will address these concerns by validating the model on multi-centre, real-world EEG datasets, incorporating online learning or transfer learning strategies, and optimizing the pipeline for real-time, resource-constrained applications. Exploring hybrid architectures that integrate deep temporal models with traditional classifiers may also enhance robustness and interpretability in dynamic seizure detection scenarios. Future research will also focus on incorporating stratified k-fold cross-validation to enhance the robustness of model evaluation. This will allow us to assess performance variability across different data partitions and provide statistically grounded metrics.

Another key challenge for real-world clinical deployment is robustness to artifacts such as muscle movements, eye blinks, and electrical interference, which are common in clinical EEG recordings. Since the UCI dataset is pre-cleaned, our current model has not been tested under such noisy conditions. Furthermore, the model has not yet been evaluated in a patient-specific or longitudinal context, where intra-subject variability over time can impact performance. Future studies will focus on testing the model with raw, multi-channel clinical EEG and assessing performance consistency across individual patients to support personalized seizure monitoring.

7. Conclusions

Epileptic seizure detection is crucial for providing timely intervention, improving patient safety, and enhancing the quality of life for individuals living with epilepsy. This paper introduced an advanced epileptic seizure detection model that addresses key challenges in existing methodologies, including data imbalance and feature selection. Traditional approaches often suffer from misdiagnosis and delayed detection, affecting patient outcomes. By integrating SMOTE for dataset balancing and a hybrid feature selection technique (PCA + DWT), our proposed model enhances the reliability of seizure classification using EEG signals. Experimental results demonstrated significant performance gains, with the SVM classifier achieving 97.30% accuracy, 99.62% AUC, and 93.08% F1 score, surpassing existing methods. The proposed approach can serve as a robust decision-support tool for neurologists, reducing the risk of delayed or inaccurate diagnoses.

A comparative evaluation with prior studies highlights the model's effectiveness in optimizing feature extraction, dimensionality reduction, and classification performance. Unlike patient-specific approaches, our proposed model enhances generalization while mitigating computational inefficiencies. While the results are promising, future work will focus on enhancing real-time implementation, optimizing computational efficiency, and exploring deep learning architectures to improve generalization across diverse datasets.

Author Contributions: Conceptualization, H.F.A. and G.E.A.; methodology, H.F.A.; software, G.E.A.; validation, H.F.A., G.E.A., and M.S.N.; formal analysis, H.F.A.; investigation, G.E.A.; resources, M.S.N.; data curation, G.E.A.; writing—original draft preparation, H.F.A. and G.E.A.; writing—review and editing, M.S.N.; visualization, M.S.N.; supervision, H.F.A.; project administration, H.F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in https://www.ukbonn.de/epileptologie/arbeitsgruppen/ag-lehnertz-neurophysik/downloads/ (accessed 28 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Andrzejak, R.G.; Lehnertz, K.; Mormann, F.; Rieke, C.; David, P.; Elger, C.E. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Phys. Rev. E* **2001**, *64*, 061907.
2. Fisher, R.S.; Cross, J.H.; French, J.A.; Higurashi, N.; Hirsch, E.; Jansen, F.E.; Lagae, L.; Moshé, S.L.; Peltola, J.; Perez, E.R.; et al. Operational classification of seizure types by the international league against epilepsy: Position paper of the ilae commission for classification and terminology. *Epilepsia* **2017**, *58*, 522–530.
3. Rayi, A.; Murr, N.I. *Electroencephalogram*; StatPearls Publishing: Tampa, FL, USA, 2024. Available online: https://www.ncbi.nlm.nih.gov/books/NBK563295/ (accessed 28 March 2025).
4. Houmani, N.; Vialatte, F.; Gallego-Jutglà, E.; Dreyfus, G.; Nguyen-Michel, V.-H.; Mariani, J.; Kinugawa, K. Diagnosis of alzheimer's disease with electroencephalography in a differential framework. *PLoS ONE* **2018**, *13*, e0193607.
5. Fergus, P.; Hussain, A.; Hignett, D.; Al-Jumeily, D.; Abdel-Aziz, K.; Hamdan, H. A machine learning system for automated whole-brain seizure detection. *Appl. Comput. Inform.* **2016**, *12*, 70–89.
6. Mursalin, M.; Zhang, Y.; Chen, Y.; Chawla, N.V. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing* **2017**, *241*, 204–214.
7. Sharmila, A.; Geethanjali, P. Dwt based detection of epileptic seizure from eeg signals using naive bayes and k-nn classifiers. *IEEE Access* **2016**, *4*, 7716–7727.
8. Fisher, R.S. The new classification of seizures by the international league against epilepsy 2017. *Curr. Neurol. Neurosci. Rep.* **2017**, *17*, 1–6.
9. Kwan, P.; Brodie, M.J. Early identification of refractory epilepsy. *New Engl. J. Med.* **2000**, *342*, 314–319.
10. World Health Organization. *Epilepsy: A Public Health Imperative*; World Health Organization: Geneva, Switzerland, 2019.
11. Ulate-Campos, A.; Coughlin, F.; Gaínza-Lein, M.; Fernández, I.S.; Pearl, P.; Loddenkemper, T. Automated seizure detection systems and their effectiveness for each type of seizure. *Seizure* **2016**, *40*, 88–101.
12. Subasi, A. Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.* **2007**, *32*, 1084–1093.
13. Shoeb, A.H. Application of machine learning to epileptic seizure onset detection and treatment. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.
14. Usman, S.M.; Latif, S.; Beg, A. Principle components analysis for seizures prediction using wavelet transform. *arXiv* **2020**, arXiv:2004.07937.
15. Donos, C.; Dümpelmann, M.; Schulze-Bonhage, A. Early seizure detection algorithm based on intracranial eeg and random forest classification. *Int. J. Neural Syst.* **2015**, *25*, 1550023.

16. Anuragi, A.; Sisodia, D.S. Empirical wavelet transform based automated alcoholism detecting using EEG signal features. *Biomed. Signal Process. Control.* **2020**, *57*, 101777. Available online: https://www.sciencedirect.com/science/article/pii/S1746809419303581 (accessed 27 March 2025).
17. Nahzat, S.; Yağanoğlu, M. Classification of epileptic seizure dataset using different machine learning algorithms and pca feature reduction technique. *J. Investig. Eng. Technol.* **2021**, *4*, 47–60.
18. Poorani, S.; Balasubramanie, P. Deep learning based epileptic seizure detection with eeg data. *Int. J. Syst. Assur. Eng. Manag.* **2023**, 1–10. https://doi.org/10.1007/s13198-022-01845-5.
19. Guo, L.; Rivero, D.; Dorado, J.; Rabunal, J.R.; Pazos, A. Automatic epileptic seizure detection in eegs based on line length feature and artificial neural networks. *J. Neurosci. Methods* **2010**, *191*, 101–109.
20. Nicolaou, N.; Georgiou, J. Detection of epileptic electroencephalogram based on permutation entropy and support vector machines. *Expert Syst. Appl.* **2012**, *39*, 202–209.
21. Wang, Z.; Liu, F.; Shi, S.; Xia, S.; Peng, F.; Wang, L.; Ai, S.; Xu, Z. Automatic epileptic seizure detection based on persistent homology. *Front. Physiol.* **2023**, *14*, 1227952. https://doi.org/10.3389/fphys.2023.1227952.
22. Hamad, A.; Houssein, E.H.; Hassanien, A.E.; Fahmy, A.A. A hybrid eeg signals classification approach based on grey wolf optimizer enhanced svms for epileptic detection. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, *Cairo*, *Egypt*, *9–11 September 2017*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 108–117.
23. Song, Y.; Liò, P. A new approach for epileptic seizure detection: Sample entropy based feature extraction and extreme learning machine. *J. Biomed. Sci. Eng.* **2010**, *3*, 556.
24. Martis, R.J.; Acharya, U.R.; Lim, C.M.; Mandana, K.; Ray, A.K.; Chakraborty, C. Application of higher order cumulant features for cardiac health diagnosis using ecg signals. *Int. J. Neural systems* **2013**, *23*, 1350014.
25. Farooq, M.S.; Zulfiqar, A.; Riaz, S. Epileptic seizure detection using machine learning: Taxonomy, opportunities, and challenges. *Diagnostics* **2023**, *13*, 1058.
26. Chandel, G.; Saini, S.K.; Sharma, A. Epileptic eeg signal classification using machine learning based model. In *Proceedings of the 2023 International Conference on Disruptive Technologies (ICDT)*, *Greater Noida*, *India*, *11–12 May 2023*; IEEE: New York, NY, USA, 2023; pp 733–739.
27. Shoeb, A.H.; Guttag, J.V. Application of machine learning to epileptic seizure detection. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010. Available online: https://api.semanticscholar.org/CorpusID:11141395 (accessed on 27 March 2025).
28. Pinto-Orellana, M.; Cerqueira, F. Patient-dependent epilepsy seizure detection using random forest classification over one-dimension transformed EEG data. *bioRxiv* **2016**, 70300. https://doi.org/10.1101/070300.
29. Khurshid, D.; Wahid, F.; Ali, S.; Gumaei, A.H.; Alzanin, S.M.; Mosleh, M.A. A deep neural network-based approach for seizure activity recognition of epilepsy sufferers. *Front. Med.* **2024**, *11*, 1405848.
30. Raghu, S.; Sriraam, N.; Temel, Y.; Rao, S.V.; Kubben, P.L. Deep learning models for predicting epileptic seizures using ieeg signals. *Electronics* **2022**, *11*, 605.
31. Zabihi, M.; Kiranyaz, S.; Rad, A.B.; Katsaggelos, A.K.; Gabbouj, M.; Ince, T. Analysis of high-dimensional phase space via Poincaré section for patient-specific seizure detection. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2015**, *24*, 386–398.
32. Krishnan, P.T.; Balasubramanian, P. Automated EEG seizure detection based on S-transform. In *Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, *Chennai*, *India*, *15–17 December 2016*; IEEE: New York, NY, USA, 2016; pp. 1–5.
33. Chen, D.; Wan, S.; Bao, F.S. Epileptic focus localization using discrete wavelet transform based on interictal intracranial EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2016**, *25*, 413–425.
34. Aarabi, A.; Wallois, F.; Grebe, R. Automated neonatal seizure detection: A multistage classification system through feature selection based on relevance and redundancy analysis. *Clin. Neurophysiol.* **2006**, *117*, 328–340.
35. Khan, Y.U.; Rafiuddin, N.; Farooq, O. Automated seizure detection in scalp EEG using multiple wavelet scales. In *Proceedings of the 2012 IEEE International Conference on Signal Processing, Computing and Control*, *Solan*, *India*, *15–17 March 2012*; IEEE: New York, NY, USA, 2021; pp. 1–5.
36. Kim, H.; Yoshimura, N.; Koike, Y. Classification of movement intention using independent components of premovement EEG. *Front. Hum. Neurosci.* **2019**, *13*, 63.
37. Pfurtscheller, G.; Neuper, C.; Flotzinger, D.; Pregenzer, M. EEG-based discrimination between imagination of right and left hand movement. *Electroencephalogr. Clin. Neurophysiol.* **1997**, *103*, 642–651.

38. Wang, Y.; Li, Z.; Feng, L.; Bai, H.; Wang, C. Hardware design of multiclass SVM classification for epilepsy and epileptic seizure detection. *IET Circuits Devices Syst*. **2018**, *12*, 108–115. https://doi.org/10.1049/iet-cds.2017.0216.
39. Kabir, E.; Siuly; Zhang, Y. Epileptic seizure detection from EEG signals using logistic model trees. *Brain Inform.* **2016**, *3*, 93–100. https://doi.org/10.1007/s40708-015-0030-2.
40. Wang, X.; Gong, G.; Li, N. Automated recognition of epileptic EEG states using a combination of symlet wavelet processing, gradient boosting machine, and grid search optimizer. *Sensors* **2019**, *19*, 219. https://doi.org/10.3390/s19020219.