



A Hybrid Approach of Factor Analysis and Decision Tree for Epilepsy Onset Prediction Using Public Datasets

Eka Sabna¹, Des Suryani², Oktavia Dewi³

¹Department of Computer Engineering, Faculty of Computer Science, Hang Tuah University of Pekanbaru, Pekanbaru, Indonesia

²Informatics Engineering, Islamic University of Riau, Pekanbaru, Indonesia

³Master of Public Health, Hang Tuah University of Pekanbaru, Pekanbaru, Indonesia

*ekasabna@htp.ac.id

Abstract

Early prediction of epilepsy onset is crucial for supporting timely clinical intervention and reducing seizure-related complications. However, the high dimensionality and complexity of Electroencephalogram (EEG) and clinical data remain major challenges for conventional predictive models. This study proposes an interpretable hybrid framework integrating Exploratory Factor Analysis (EFA) and Decision Tree classification for early epilepsy onset prediction using public datasets. EFA was employed to reduce 15 observed variables into five clinically meaningful latent factors, which were subsequently used as inputs for the Decision Tree model. According to experimental data, the suggested framework demonstrated steady classification performance with an accuracy of 88.14% and specificity of 88.73%, while its efficacy in identifying epilepsy beginning cases for early screening is highlighted by its sensitivity of 80.77%. The latent factor representing clinical neurological abnormalities was identified as the most influential predictor in the classification process. Compared with conventional black-box machine learning approaches, the proposed model provides transparent decision rules and clinically meaningful interpretation while maintaining reliable predictive capability. All things considered, the suggested architecture is a viable path for creating clinically interpretable prediction models that facilitate transparent and trustworthy early epilepsy screening.

Keywords: decision tree; EEG; epilepsy onset prediction; exploratory factor analysis; interpretable machine learning

How to Cite: E. Sabna, D. Suryani, and O. Dewi, "A Hybrid Approach of Factor Analysis and Decision Tree for Epilepsy Onset Prediction Using Public Datasets", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 4, pp. 984 - 995, Aug. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i4.7253>

Received: October 9, 2025

Accepted: May 23, 2026

Available Online: August 1, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Recurrent seizures brought on by aberrant electrical activity in the brain are the hallmark of epilepsy, a chronic neurological condition [1]. The World Health Organization (2019) reports that epilepsy affects over 50 million people globally, making it one of the most common neurological conditions and having a major negative influence on patients' quality of life [2], [3]. Early detection of epilepsy onset, defined as the initial occurrence of seizures, is critical for supporting clinical decision-making and minimizing the risk of severe complications. The onset of epilepsy is influenced by multiple interrelated factors, including perinatal, genetic, medical, and socio-economic conditions. Due to this complexity, accurate prediction of epilepsy onset remains a challenging yet essential task for enabling timely intervention and preventive strategies [4], [5]. In

clinical settings, electroencephalogram (EEG) data serve as a primary modality for identifying epileptic seizure patterns [6]. However, EEG data are inherently high-dimensional, complex, and non-linear, which limits the effectiveness of conventional analytical approaches [7], [8]. These difficulties underscore the necessity for stronger and more effective computational techniques to enhance prediction accuracy.

Over the past two decades, technological advancements have enabled more systematic approaches to epilepsy onset prediction. One widely utilized public resource is the Kaggle Epilepsy Onset Prediction Dataset, which contains electroencephalography (EEG) recordings along with associated clinical information, thereby providing a foundation for statistical and machine learning-based modeling. Despite its advantages, the dataset is characterized by high dimensionality,

complex structures, and significant inter-variable correlations, which present substantial challenges for conventional analytical methods [7]. To address these limitations, data mining techniques have been increasingly employed to extract meaningful patterns from large-scale medical datasets [9], [10]. Moreover, previous research has shown that information driven methods can represent intricate, non-linear connection that are frequently hard to recognize through conventional analysis methods [11]. Earlier research has mainly utilized individual classification methods, like Decision Trees, Support Vector Machines (SVM), and Random Forests classifiers for predicting epilepsy using EEG or clinical information. While these methods have shown competitive performance, they frequently experience elevated dimensionality [12]. In addition, model interpretability is not always explicitly addressed, despite its critical importance in medical decision-making.

Recently, deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have been investigated for tasks involving epilepsy prediction. Although these models can capture intricate patterns in high-dimensional dataset, they typically come with grater computational demands and reduced interpretability [13], [14]. These constraints emphasize the necessity for strategies that can proficiently balance predictive accuracy and interpretability. In this context, integrating dimensionality reduction techniques, such as Factor Analysis, with interpretable classifiers like Decision Trees offers a promising direction to reduce feature complexity while preserving explanatory capability.

Recent studies have explored hybrid modeling strategies that combine dimensionality reduction techniques with classification algorithms to address high-dimensional data and improve computational efficiency [14]. In addition, several works have investigated the integration of handcrafted features with deep learning models, reporting promising results in specific scenarios [9]. However, deep learning-based approaches are often associated with limited interpretability, which may constrain their applicability in clinical settings where transparency is essential [15].

Recent studies on epileptic seizure prediction have increasingly adopted machine learning and hybrid frameworks that integrate feature extraction techniques with classification models. Typically, these methods adhere to a conventional workflow that includes signal preprocessing, feature extraction, and classification phases leveraging different algorithms like Support Vector Machines (SVM), ensemble techniques, and deep learning frameworks [16]. Specifically, tree-based and ensemble methods, such as Random Forest and Gradient Boosting Decision Trees, have shown impressive predictive capabilities when integrated with nonlinear EEG characteristics [17]. Furthermore, recent studies have explored advanced techniques such

as classifier stacking, concept drift adaptation, transformer-based deep learning models, and hybrid architectures to further improve prediction accuracy [18], [19]. Employing various algorithms in one study offers comparative insights into the performance traits of each approach, with Random Forest typically showing more consistent performance on intricate medical datasets, whereas SVM and Naïve Bayes are often effective on datasets featuring simpler structures. These results highlight the significance of choosing approaches that correspond with data traits when creating dependable predictive models for healthcare uses.

A separate study by Hairani Hairani et al. [20] explored the classification of anemia classification through Random Forest, Support Vector Machine (SVM), and Naïve Bayes algorithms. The research showed that machine learning methods can enhance the precision of disease diagnosis using clinical data by following structured phases of data preprocessing, model training, and performance assessment. Moreover, while feature reduction methods like Principal Component Analysis (PCA) and various statistical transformations are commonly used to enhance model efficiency and decrease data dimensionality, these techniques are typically employed as preprocessing phases without offering clear clinical interpretations of the altered feature [21]. As a result, the relationships between input variables and clinically meaningful risk factors are often not adequately represented in the final predictive models. In addition, most existing epilepsy prediction studies continue to rely heavily on raw or engineered features without explicitly modeling latent clinical structures that may better capture the underlying neurological, genetic, and developmental risk dimensions associated with epilepsy onset [22].

Exploratory Factor Analysis (EFA) with a Decision Tree classifier for epilepsy onset prediction. EFA is employed for feature reduction due to its ability to identify latent constructs underlying correlated clinical variables. In contrast to Principal Component Analysis (PCA), which emphasizes maximizing explained variance, EFA seeks to model the latent structure among observed variables, rendering it more appropriate for extracting clinically meaningful factors. This characteristic is particularly important in the medical domain, where interpretability and clinical relevance are essential for meaningful decision-support systems.

Decision Tree algorithm is selected due to its transparency and rule-based structure, which enables direct interpretability of decision paths. In contrast to other machine learning models such as Support Vector Machines, Random Forest, and deep neural networks, which are generally considered black-box approaches, Decision Trees provide explicit classification rules that are easier to understand and interpret [23]. This interpretability is particularly important in clinical applications, where model explanations are required to

build trust, ensure ethical use, and support the validation of predictive outcomes [23-25].

The novelty of this study does not lie in the use of individual methods alone, but in the structured integration of these techniques into a unified and interpretable predictive framework. This approach seeks to bridge the gap between high-performing machine learning models and the need for clinically meaningful interpretability, making it particularly relevant for early-stage epilepsy screening, where both predictive accuracy and explainability are critical in supporting medical decision-making systems. Therefore, this study proposes a hybrid approach that combines Factor Analysis as a dimensionality reduction technique with a Decision Tree classifier for epilepsy onset prediction using public datasets. This approach is designed to reduce feature redundancy, maintain model transparency, and achieve reliable predictive performance.

The contribution of this study lies in providing an empirical evaluation of the proposed hybrid framework in handling data complexity while maintaining model interpretability. Furthermore, this study offers insights into the potential of integrating dimensionality reduction with interpretable classification techniques to support the development of data-driven clinical decision support systems.

2. Methods

This study adopts a quantitative approach by applying data mining techniques to predict epilepsy onset through the integration of Factor Analysis and a Decision Tree algorithm. The research procedures were conducted systematically to ensure reproducibility and adherence to established scientific standards. The overall research process consists of several sequential stages, as illustrated in Figure 1.

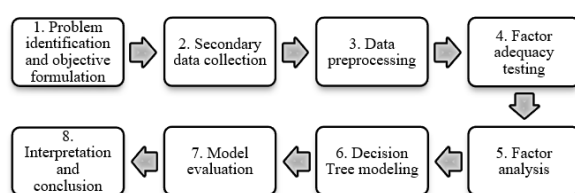


Figure 1. Research Stages

The research workflow began with (1) Problem Identification and Objective Formulation, which were conducted through a comprehensive literature review and preliminary exploration of the epilepsy dataset available on the Kaggle platform. Subsequently, (2) Secondary Data Collection was performed using the *Epilepsy Onset Dataset* obtained from Kaggle, which served as the primary data source in this study. The dataset comprised 7,000 records, with independent variables (X) including Socioeconomic Status (X1), Birth Asphyxia History (X2), Premature Birth History (X3), Low Birth Weight History (X4), Family History of Epilepsy (X5), Cerebral Malaria History (X6),

Meningitis/Encephalitis History (X7), Severe Head Trauma History (X8), Intellectual Disability (X9), Autism Spectrum Disorder (X10), Imaging Abnormalities (X11), Baseline EEG Abnormalities (X12), Stroke/Brain Tumor History (X13), Ratio Score (X14), and Epilepsy Diagnosis (X15). The dependent variable (Y) was epilepsy onset. The subsequent step involved (3) Data Preprocessing, which included cleaning the dataset to handle missing values and standardizing numerical variables to ensure readiness for factor analysis.

Following this, (4) Factor Adequacy Testing was performed using the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity to verify that the dataset met the requirements for dimensionality reduction [26]. The Measure of Sampling Adequacy (MSA) was also considered, where variables with MSA values below 0.50 were removed to improve model stability. The Measure of Sampling Adequacy (MSA) was also assessed, with variables exhibiting MSA values below 0.50 being removed to enhance model stability. The core analysis commenced with (5) Exploratory Factor Analysis (EFA) using Principal Component Analysis (PCA) as the extraction method. PCA was applied to reduce the large set of observed variables into a smaller number of principal factors while minimizing information loss, with eigenvalues greater than 1 serving as the extraction criterion. Varimax rotation was employed to produce more interpretable latent factors, clarifying the factor structure and facilitating subsequent interpretation. The extracted factors were then used as predictor variables in a binary logistic regression model to estimate the probability of epilepsy onset, thereby reducing data complexity without sacrificing critical information.

Next, (6) Decision Tree Model was constructed. The Decision Tree algorithm was chosen for its high interpretability and its suitability for generating rule-based predictions. By integrating factor analysis with Decision Tree modeling, this study aimed to develop a predictive framework that is not only accurate but also efficient and easily interpretable by medical practitioners. Subsequently, (7) Model Evaluation was conducted using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score, to assess the predictive performance of the model. Finally, (8) Interpretation And Conclusion were performed, in which the findings were thoroughly analyzed to provide insights relevant to both academic research and clinical practice. Through this systematic workflow, the integration of Factor Analysis and the Decision Tree algorithm is expected to yield a valid and robust model for epilepsy onset prediction, providing a foundation for data-driven decision-making in medical contexts.

3. Results and Discussions

3.1 Results

This study employed an integrated framework combining Factor Analysis and the Decision Tree

algorithm. The overall workflow of this integration is illustrated in Figure 2. In the first step, Factor Analysis was applied to reduce the dimensionality of the dataset and to identify latent factors representing clusters of clinically or biometrically related variables. This procedure not only simplified the data structure but also preserved the most informative components relevant for prediction. The extracted latent factors subsequently served as input variables for the Decision Tree classifier. By utilizing these latent factors as predictors, the Decision Tree model was able to generate classification rules that were both concise and interpretable. These rules provide insights into the contribution of various clinical and biometrical characteristics to the probability of epilepsy onset. The integration of Factor Analysis with the Decision Tree algorithm therefore resulted in a predictive model that is both accurate and transparent, offering clinically meaningful outputs aligned with the requirements of interpretable decision-support systems in medical practice. The workflow illustrates the sequential integration of Factor Analysis and the Decision Tree algorithm in constructing a predictive model for epilepsy onset. The process begins with data preprocessing and Factor Analysis, in which multiple EEG and clinical variables are condensed into a smaller set of latent factors through Principal Component Analysis (PCA) and Varimax rotation. These extracted factors subsequently serve as input features for the Decision Tree, which generates interpretable classification rules to distinguish between epileptic and non-epileptic onset. This integrated approach enables dimensionality reduction while maintaining predictive accuracy, providing a model that is both efficient and clinically interpretable.

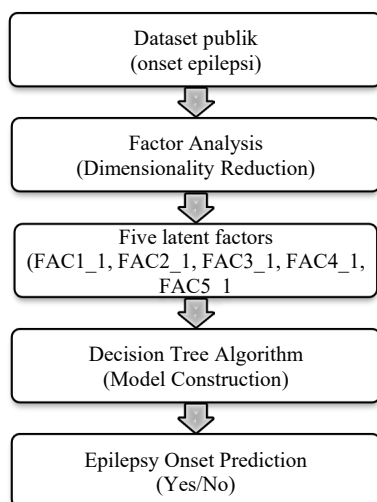


Figure 2. Workflow Of Integrating Factor Analysis with Decision Tree for Epilepsy Onset Prediction

The dataset used in this study was obtained from a publicly available epilepsy dataset hosted on the international Kaggle repository. The dataset comprises 7,000 samples containing EEG activity recordings, with independent variables X1–X15 and the dependent

variable Y representing epilepsy onset. The core stage of the analysis involves conducting Exploratory Factor Analysis (EFA) to reduce the large number of observed variables into key latent factors that effectively represent the underlying constructs relevant to epilepsy risk.

Factor Analysis is applied to group epilepsy onset risk variables into meaningful latent factors. Prior to the extraction process, data suitability is assessed using the Kaiser–Meyer–Olkin (KMO) measure, Bartlett’s Test of Sphericity, and the Anti-Image Correlation Matrix (MSA). These tests are conducted to ensure that the dataset is appropriate for factor analysis by evaluating sampling adequacy and the presence of sufficient correlations among variables. Only variables that meet the required statistical thresholds are retained for further analysis, thereby improving the reliability and interpretability of the resulting latent factors.

Factor Analysis was applied to condense the risk variables associated with epilepsy onset into meaningful latent factors. Prior to factor extraction, preliminary statistical tests were conducted to ensure the suitability of the dataset for dimensionality reduction. These included the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy, Bartlett’s Test of Sphericity, and the Anti-Image Correlation Matrix (Measure of Sampling Adequacy, MSA). The results of the KMO and Bartlett’s tests are presented in Figure 3. The KMO measure produced a value of 0.502, which exceeds the minimum acceptable threshold of 0.5 [1], indicating that the dataset is adequate for factor extraction. Additionally, Bartlett’s Test of Sphericity yielded a chi-square value of 2946.391 with a significance level of $p < 0.001$, supporting the presence of sufficient correlations among variables to justify Factor Analysis [18]. The combination of a KMO value of 0.502 and a highly significant Bartlett’s Test result confirms that the dataset meets the minimum adequacy requirements and satisfies the fundamental assumptions necessary for conducting Exploratory Factor Analysis. Exploratory Factor Analysis (EFA). The adequacy of sampling and the presence of significant inter-variable correlations indicate that meaningful latent structures can be reliably extracted. Consequently, EFA can be effectively applied to reduce data complexity while preserving the underlying relationships among clinical variables.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Adequacy.	Measure of Sampling	.502
Bartlett's Test of Sphericity	Approx. Chi-Square	2946.391
	df	36
	Sig.	<.001

Figure 3. KMO and Bartlett's Test

The Measure of Sampling Adequacy (MSA) is used to evaluate the suitability of individual variables for inclusion in factor analysis. MSA values range from 0 to 1, with values below 0.50 considered inadequate for factor extraction. As recommended by Hair et al.

(2019), variables with $MSA < 0.50$ should be removed to enhance the stability and reliability of the factor model [5]. MSA is derived from the anti-image correlation matrix and reflects whether a variable exhibits sufficient correlation with other variables to justify its inclusion in the factor extraction process. In this study, variables X7, X8, and X14 were excluded due to MSA values below the 0.50 threshold. The remaining variables—X4 (0.503), X5 (0.501), X6 (0.501), X9 (0.598), X10 (0.567), X11 (0.501), X12 (0.501), X13 (0.663), and X15 (0.556)—all exceeded the 0.50 criterion and were retained for subsequent factor analysis. These results confirmed that the selected variables were adequate for extraction into

latent factors, ensuring both the stability and interpretability of the resulting factor model.

The Measure of Sampling Adequacy (MSA) values are used to evaluate the contribution of each variable to the overall factor structure. The MSA values are presented in Table 1. In this study, variables X7, X8, and X14 were removed based on low MSA values. This elimination process not only improves the statistical quality of the model but also simplifies the factor structure by removing variables that may introduce noise and reduce model stability. The retained variables exhibit stronger and more consistent correlation patterns, thereby supporting the formation of more stable and clinically interpretable latent factors.

Table 1. Anti-Image Matrices

		X4	X5	X6	X9	X10	X11	X12	X13	X15
Anti-image Covariance	X4	.999	2.959E-5	.007	.003	.014	.009	.000	-.006	.020
	X5	2.959E-5	.780	-.365	.012	.012	-.007	.015	.007	.002
	X6	.007	-.365	.780	.002	.009	-.003	-.008	-.001	.024
	X9	.003	.012	.002	.999	-.008	.005	.002	.011	-.007
	X10	.014	.012	.009	-.008	.999	-.004	.005	-.002	.024
	X11	.009	-.007	-.003	.005	-.004	.845	-.332	-.016	-.004
	X12	.000	.015	-.008	.002	.005	-.332	.845	-.016	-.001
	X13	-.006	.007	-.001	.011	-.002	-.016	-.016	.999	.003
	X15	.020	.002	.024	-.007	.024	-.004	-.001	.003	.998
	X4	.503 ^a	3.352E-5	.008	.003	.014	.009	.000	-.006	.020
Anti-image Correlation	X5	3.352E-5	.501 ^a	-.468	.013	.013	-.008	.018	.008	.002
	X6	.008	-.468	.501 ^a	.002	.010	-.003	-.009	-.001	.027
	X9	.003	.013	.002	.598 ^a	-.008	.005	.002	.011	-.007
	X10	.014	.013	.010	-.008	.567 ^a	-.004	.006	-.002	.024
	X11	.009	-.008	-.003	.005	-.004	.501 ^a	-.393	-.017	-.004
	X12	.000	.018	-.009	.002	.006	-.393	.501 ^a	-.017	-.002
	X13	-.006	.008	-.001	.011	-.002	-.017	-.017	.663 ^a	.003
	X15	.020	.002	.027	-.007	.024	-.004	-.002	.003	.556 ^a

a. Measures of Sampling Adequacy (MSA)

After confirming the adequacy of the dataset through the Kaiser–Meyer–Olkin (KMO) test, Bartlett’s Test of Sphericity, and the Anti-Image Matrix (Measure of Sampling Adequacy, MSA), factor extraction was performed. Factor extraction is performed using Principal Component Analysis (PCA). The communality values indicate that most variables are well represented within the extracted factors, suggesting that the underlying data structure is adequately captured by the model.

Varimax rotation is then applied to produce a more orthogonal and interpretable factor structure. This approach enables the identification of groups of variables with strong associations within the same latent dimension, thereby facilitating a more meaningful clinical interpretation of epilepsy risk factors.

Following extraction, communalities were examined to determine the proportion of variance in each variable explained by the extracted factors. This assessment ensured that the retained factors adequately represented the underlying variables. To enhance interpretability, factor rotation was subsequently applied using the Varimax method, yielding a clearer and more easily

interpretable factor structure. This step facilitated the identification and understanding of the latent constructs underlying the observed variables. The *Total Variance Explained* table, generated from PCA, summarizes the number of factors extracted and the proportion of total variance explained by each factor. This output provides an overview of the effectiveness of dimensionality reduction and the adequacy of the factor solution in representing the dataset.

The *Extraction Sums of Squared Loadings* column represents the amount of variance explained by each factor prior to rotation, whereas the *Rotation Sums of Squared Loadings* column indicates the variance explained after the factor rotation process. Factor rotation was applied to achieve a clearer and more interpretable factor structure. Based on the PCA results presented in Table 2, five principal factors were extracted from the twelve observed variables, as indicated by eigenvalues greater than 1. Factor 1 accounted for 16.374% of the total variance, Factor 2 explained 15.528%, Factor 3 contributed 11.406%, Factor 4 accounted for 11.318%, and Factor 5 explained 11.124% of the variance.

Table 2. Principal Component Analysis

Component	Total Variance Explained					
	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.474	16.374	16.374	1.473	16.369	16.369
2	1.397	15.528	31.902	1.395	15.495	31.864
3	1.027	11.406	43.308	1.023	11.365	43.230
4	1.019	11.318	54.626	1.015	11.279	54.508
5	1.001	11.124	65.749	1.012	11.241	65.749

Extraction Method: Principal Component Analysis.

Collectively, these five factors explained 65.749% of the total variance in the dataset. This cumulative variance surpasses the commonly accepted threshold of 60% for multivariate analyses [27], suggesting that the extracted factors adequately represent the underlying structure of the original variables. This indicates that the factor model is capable of capturing most of the information contained in the dataset without losing its underlying structural patterns. Consequently, the reduced factor representation maintains the essential variability of the original variables while enhancing model simplicity and interpretability. This balance between information retention and dimensionality reduction supports the effectiveness of the proposed approach in preparing clinically meaningful inputs for subsequent classification modeling. The remaining 34.25% of the variance is attributed to unique variance associated with individual variables or measurement error that is not captured by the extracted factors. In factor analysis, communalities indicate the proportion of variance in each variable that is explained by the extracted factors. The results of the factor extraction are presented in Table 3.

Table 3. Extraction

Communalities	
	Extraction
X4	.761
X5	.730
X6	.731
X9	.596
X10	.679
X11	.695
X12	.696
X13	.427
X15	.603

Extraction Method: Principal Component Analysis

For example, variable X4 exhibited a communality of 0.761, implying that 76.1% of its variance is accounted for by the factors, while the remaining 23.9% is due to unique variance or measurement error. According to Hair et al. (2019), higher communalities generally above 0.50 demonstrate that variables are well represented by the factor model. Most variables in this study had communalities exceeding 0.50, confirming that they were adequately captured by the latent factors. However, variable X13 had a communality of 0.427, indicating that it is less strongly represented by the extracted factors compared to the other variables. This may indicate that these variables exhibit more specific relationships or are not fully aligned with the extracted

latent dimensions. Such behavior suggests that they may capture unique or context-specific information that is not strongly represented within the common factor structure. Consequently, excluding these variables helps improve the coherence of the factor model and enhances the overall interpretability and stability of the resulting latent constructs.

The Rotated Component Matrix is a key output of factor analysis conducted using Principal Component Analysis (PCA) with Varimax rotation. This matrix allows for the identification of which observed variables load onto specific latent factors and facilitates the assessment of the strength of the association between each variable and its corresponding factor. Varimax rotation was employed to maximize high factor loadings while minimizing low loadings, thereby producing a factor structure that is more interpretable. By enhancing the clarity of variable-factor relationships, the rotated component matrix enables a better understanding of the underlying latent constructs represented by the observed variables. This step is crucial for accurately labeling and interpreting the five extracted factors, ensuring that each factor reflects a coherent combination of clinically or biometrically related variables.

Table 4. Rotated Component Matrix^a

	1	2	3	4	5
X4	-.032	-.004	-.064	.869	.027
X5	.854	-.009	-.014	-.009	.002
X6	.855	.012	.011	-.005	.003
X9	-.033	.038	.078	.004	.766
X10	-.057	-.002	.773	-.276	.040
X11	.013	.833	.002	-.013	-.017
X12	-.009	.834	-.008	.007	-.014
X13	-.032	.057	.073	-.006	-.646
X15	-.075	.007	-.640	-.428	.067

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Table 4 presents the Rotated Component Matrix. Based on the analysis, five latent factors were successfully extracted, each representing a distinct dimension of the underlying construct. Variables with absolute factor loadings greater than 0.50 were considered significant contributors, ensuring that only meaningful relationships were retained in the interpretation. The first factor is defined by variables X5 (0.854) and X6 (0.855), both of which exhibit strong loadings, indicating substantial contributions to the latent

construct. Similarly, the second factor comprises X11 (0.833) and X12 (0.834), demonstrating strong and consistent associations with the underlying dimension.

The third factor includes X10 (0.773) and X15 (−0.640). Although X15 shows a negative loading, this reflects an inverse relationship with the factor while still representing a meaningful contribution. In contrast, the fourth factor is predominantly characterized by X4 (0.869), which exhibits a high loading and represents a clear and dominant dimension within the construct. The fifth factor consists of X9 (0.766) and X13 (−0.646), where X13 displays a negative loading, indicating an inverse yet relevant association with the latent factor. Despite the presence of negative loadings in certain variables, all included indicators contribute significantly to their respective factors. Overall, these findings demonstrate that the extracted factors are stable, interpretable, and well-defined. Each factor is supported by variables with substantial loadings, confirming the adequacy of both the factor extraction and rotation processes. Furthermore, the application of Varimax rotation effectively clarifies the factor structure, thereby facilitating a more precise interpretation of the relationships between observed variables and their underlying latent dimensions. Table 5 summarizes the loadings of the five extracted latent factors :

Table 5. Results of Factor Analysis of Epilepsy Onset Determinants

Factor	Constituent Variables	Factor Label
Factor 1 (FAC1_1)	X5=Family History of Epilepsy X6=History of Cerebral Malaria	Family History & Brain Infection
Factor 2 (FAC2_1)	X11=Imaging Abnormalities X12=Baseline EEG Abnormalities	Clinical Neurological Abnormalities
Factor 3 (FAC3_1)	X10=Autism Spectrum Disorder X15=Epilepsy Probability	Developmental Comorbidities & Clinical Risk
Factor 4 (FAC4_1)	X4=Low Birth Weight History	Perinatal Risk Factor for Epilepsy
Factor 5 (FAC5_1)	X9= Intellectual Disability X13=History of Stroke/ Brain Tumor	Cognitive & Neuropathological History

The emergence of five latent factors reflects the multifactorial nature of epilepsy onset, which is influenced by genetic, neurological, developmental, and perinatal conditions. This finding is consistent with prior epidemiological studies indicating that epilepsy risk is rarely driven by a single factor, but rather by the interaction of multiple clinical dimensions.

The emergence of five latent factors reflects the multifactorial nature of epilepsy, where disease onset is not caused by a single determinant but rather by a complex interaction of genetic, neurological, and environmental influences. From a clinical perspective, this factor structure provides a strong conceptual framework for understanding patient risk stratification. For instance, the clinical neurological abnormality factor is directly associated with brain dysfunction that may increase susceptibility to seizures, whereas the

perinatal factor reflects long-term risks originating from complications during birth.

The factor analysis identified five latent constructs that collectively represent the major domains influencing epilepsy onset, encompassing clinical, developmental, and neuropathological dimensions. Each factor reflects a distinct yet interrelated aspect of epilepsy risk, providing a comprehensive understanding of the underlying mechanisms. The first factor represents Family History and Brain Infection, highlighting the combined influence of genetic susceptibility and environmental exposure. A family history of epilepsy is widely recognized as a strong predictor of seizure risk, reflecting inherited predispositions as well as shared environmental factors. In addition, central nervous system infections, such as cerebral malaria, contribute to acquired epilepsy, particularly in endemic regions. These findings underscore the dual role of hereditary and infectious determinants in shaping epilepsy risk, especially in low- and middle-income settings [28]. Central nervous system (CNS) infections, such as cerebral malaria, are also well recognized as causes of acquired epilepsy in endemic regions [29]. These findings support WHO [2] reports that both hereditary predisposition and infectious diseases contribute substantially to the global epilepsy burden, especially in low and middle income countries. The second factor, Clinical Neurological Abnormalities, captures structural and functional irregularities in the brain as identified through neuroimaging and electroencephalography (EEG). Such abnormalities are central to established classifications of epilepsy etiology and are strongly associated with seizure onset. Structural lesions, cortical malformations, and functional dysrhythmias provide clinically relevant evidence supporting the importance of this factor in understanding epilepsy development [30]. Structural brain lesions, cortical malformations, or functional dysrhythmias captured through these modalities are strongly associated with seizure onset [31]. The third factor reflects Developmental Comorbidities and Clinical Risk, emphasizing the interplay between neurological development and seizure susceptibility. The co-occurrence of autism spectrum disorder (ASD) and epilepsy is well documented, with a substantial proportion of individuals with ASD experiencing seizures. Moreover, the clustering of cognitive and behavioral comorbidities in newly diagnosed epilepsy cases further supports the interpretation of this factor as a key developmental risk domain [32]. Morales et al [33] further highlight the clustering of cognitive and behavioral comorbidities in new-onset epilepsy, reinforcing the interpretation of this factor as a developmental risk domain. The fourth factor, Perinatal Risk Factors, highlights the role of early-life conditions in influencing long-term neurological outcomes. Low birth weight and perinatal complications are consistently linked to an increased risk of epilepsy, often due to hypoxic-ischemic injury and delayed brain development. These findings emphasize the importance

of perinatal care and early interventions in reducing the likelihood of epilepsy onset [34]. WHO [2] also identifies birth-related injuries as a key preventable risk factor, underscoring the importance of perinatal health interventions. Finally, the fifth factor represents Cognitive and Neuropathological History, encompassing conditions such as intellectual disability and acquired brain injuries. Intellectual disability frequently co-occurs with epilepsy, suggesting shared underlying neuropathological mechanisms. Additionally, conditions such as stroke and brain tumors are well-established structural causes of epilepsy, particularly in later life [1]. These conditions align with this latent construct, reinforcing its clinical plausibility. Together, the five extracted factors explain 65.7% of the total variance and provide a robust multidimensional framework for understanding epilepsy onset. They capture the interplay of hereditary, infectious, developmental, perinatal, and neuropathological domains, supporting a holistic model consistent with recent epidemiological perspectives [2], [33].

The dimensionality reduction achieved through factor analysis not only enhanced the efficiency of subsequent analyses but also facilitated the development of a predictive model using the Decision Tree algorithm. The extracted latent factors served as input variables for constructing the Decision Tree, enabling a more streamlined and informative modeling process. In this study, the Decision Tree algorithm was employed as a classification method to predict epilepsy onset based on a publicly available dataset comprising 7,000 records. This algorithm was selected due to its capacity to generate interpretable models and to produce decision rules in the form of a tree structure, which can be readily understood by both researchers and clinical practitioners.

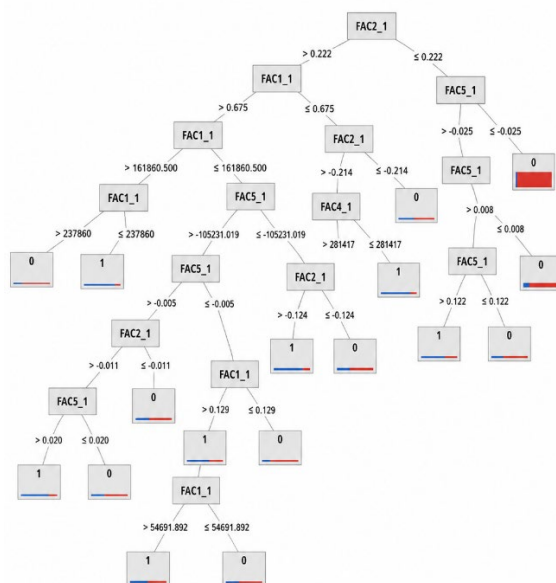


Figure 4. Decision Tree Model

The dataset was divided into a training set, consisting of 90% of the records (6,300 samples), and a testing set, comprising the remaining 10% (700 samples) to evaluate model performance objectively. During tree construction, the algorithm selected the root node based on the attribute exhibiting the highest information gain, ensuring optimal partitioning at each decision point. After training, the model was evaluated using the testing data to assess its predictive performance in classifying epilepsy onset into two categories: Yes (1) or No (0). The resulting Decision Tree structure, illustrating the hierarchical decision rules derived from the latent factors, is presented in Figure 4. This approach provides both accurate predictions and interpretable rules, supporting clinical decision-making and offering insights into the factors most strongly associated with epilepsy onset.

The Decision Tree model is constructed based on the highest information gain values, enabling the selection of the most informative attribute as the root node. The resulting tree structure is hierarchical and inherently interpretable, making it highly suitable for clinical applications that require transparency in decision-making processes. Each decision path represents a clear set of rule-based conditions, allowing clinicians to trace and understand how predictions are derived from the input factors. The results of the decision tree in Figure 3 show that the initial root node is the latent factor FAC2_1, indicating that this variable provides the highest information gain in classifying the onset of epilepsy. This finding underscores the central role of FAC2_1 as the most influential predictor in the hierarchical decision-making process of the model. The results show that the root node is the latent factor FAC2_1 (Clinical Neurological Abnormalities), which has the highest information gain. This indicates that this factor is the most dominant predictor in determining epilepsy onset. From a clinical perspective, this finding is consistent with the understanding that neurological impairments serve as direct indicators of brain dysfunction closely associated with seizure occurrence. Consequently, the decision tree structure further reinforces the clinical relevance of this factor as a primary determinant in the early classification of epilepsy risk.

The predictive performance of the Decision Tree model was assessed using the testing dataset through a confusion matrix. Key evaluation metrics, including accuracy, precision, recall (sensitivity), specificity, and F1-score, are summarized in Table 6 [35].

Table 6. Presents The Evaluation Metrics Of The Decision Tree Classification Model.

	true 1	true 0	class precision
pred.1	42	10	80.77%
pred.0	73	575	88.73%
class recall	36.52%	98.29%	

Performance measures were calculated based on standard Equations 1-5.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1_Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (5)$$

Table 7 presents the performance evaluation results. The evaluation results indicate that the model achieved an accuracy of 88.14% and a specificity of 88.73%, suggesting that the Decision Tree algorithm is reliable in correctly identifying individuals without epilepsy onset (negative class). This reflects a relatively low rate of false positives among healthy cases. However, the precision was 36.52%, indicating that a considerable proportion of cases predicted as epilepsy onset were not true positives. The relatively low precision may be attributed to class imbalance in the dataset, where the number of non-epilepsy cases dominates. As a result, the model tends to over-predict positive cases, increasing false positives. This trade-off is common in medical screening models that prioritize sensitivity over precision. This trade-off reflects a common characteristic in medical screening models, where sensitivity is prioritized over precision to minimize the risk of missed positive cases. In the context of epilepsy screening, such an approach is considered acceptable because false negatives may have more severe clinical consequences than false positives. Therefore, although precision is reduced, the model remains clinically useful as a preliminary screening tool that ensures high detection capability for potential epilepsy cases.

Table 7. Evaluation performance results

Metric	Value
Accuracy	88.14
Precision	36.52
Recall (Sensitivity)	80.77
Specificity	88.73
F1-Score	50.30

This reflects a notable rate of false-positive classifications. In contrast, the recall of 80.77% demonstrates that the model correctly identified the majority of true epilepsy onset cases, which is critical in a clinical setting, as minimizing false negatives is paramount to avoid missing actual epilepsy cases. The F1-score of 50.30% represents a moderate balance between precision and recall, indicating that while the model effectively detects positive cases, further optimization is warranted. Improvements may include selecting more representative predictor variables or implementing ensemble learning techniques to enhance precision without compromising recall. The findings indicate that the integration of factor analysis with the Decision Tree algorithm is capable of generating a predictive model with high sensitivity. However, the model exhibits limitations in accurately identifying positive cases. Accordingly, this approach is well-

suited for use as an initial screening tool, yet it should be complemented by additional methods to improve the precision of positive case detection.

3.2 Discussions

The findings of this study indicate that the integration of Factor Analysis and the Decision Tree algorithm has the potential to yield satisfactory predictive performance for epilepsy onset. The observed accuracy of 88.14% demonstrates that the proposed model is capable of correctly classifying a substantial proportion of the dataset. This performance may be attributed to the contribution of Factor Analysis in reducing data dimensionality and alleviating multicollinearity among variables, an issue frequently reported in medical datasets characterized by highly correlated features [8]. The recall value of 80.77% suggests that the model is relatively effective in identifying epilepsy cases. This result is consistent with prior studies emphasizing the importance of sensitivity-oriented models in early detection tasks within the medical domain, where failure to identify positive cases may lead to serious clinical consequences. Furthermore, the specificity of 88.73% indicates that the model maintains a balanced capability in correctly recognizing non-epilepsy instances, aligning with findings from previous research employing Decision Tree-based approaches for similar classification problems. However, the relatively low precision of 36.52% reveals that the model generates a considerable number of false positive predictions. This limitation is likely associated with class imbalance in the dataset, which has been widely acknowledged in the literature as a key factor contributing to reduced precision in biomedical classification tasks [36]. From a clinical perspective, the proposed model is more suitable as an early screening tool, where sensitivity is prioritized over precision. However, the relatively high number of false positives indicates the necessity of further diagnostic confirmation to validate the predicted outcomes. To improve model performance, particularly precision, advanced techniques such as data balancing methods (e.g., resampling strategies) or ensemble-based approaches (e.g., Random Forest) may be considered. In addition, incorporating additional clinical variables or multimodal data sources may further enhance the model's discriminative ability and overall robustness in real-world clinical applications.

Similar challenges have been addressed in previous studies through the application of data balancing techniques, such as modified Synthetic Minority Over-sampling Technique (SMOTE) combined with noise filtering and distance-based approaches [37]. These methods have been shown to improve class distribution and reduce the impact of noisy samples, thereby enhancing precision and overall classification robustness. In contrast, the absence of explicit data balancing in the present study may explain the persistence of false positive predictions, despite the model achieving relatively high recall. In addition,

prior research has demonstrated that combining dimensionality reduction techniques, such as Principal Component Analysis (PCA), with ensemble learning methods like Random Forest optimized with K-Means SMOTE-ENN can lead to improved performance, particularly in terms of precision and F1-score [38]. These approaches benefit from both balanced data representation and the robustness of ensemble classifiers. However, such methods often sacrifice interpretability, as PCA-derived components are less directly interpretable in clinical terms, and ensemble models typically operate as black-box systems. In contrast, the proposed approach emphasizes interpretability by employing Factor Analysis to extract clinically meaningful latent factors and a Decision Tree model to provide transparent and traceable classification rules. From an interpretability standpoint, this hybrid framework offers a distinct advantage, as it enables both the identification of underlying data structures and the explicit examination of the decision-making process. This level of transparency is particularly important in medical applications, where explainability is often as critical as predictive accuracy. Consequently, although the proposed model does not yet achieve optimal precision compared to SMOTE-enhanced or ensemble-based approaches, it provides a balanced framework that prioritizes sensitivity and interpretability. This trade-off may be considered acceptable in the context of preliminary medical screening, where minimizing missed cases is often more critical than reducing false positives.

Overall, the results suggest that the proposed hybrid approach represents a promising exploratory framework. Future research should consider integrating

data balancing techniques, such as SMOTE or its variants, as well as exploring ensemble-based extensions to improve precision and F1-score while preserving the interpretability of the model. Additionally, validation using balanced and external datasets is necessary to further enhance the robustness and generalizability of the proposed approach.

From an interpretability perspective, the proposed hybrid approach offers a potential advantage over black-box models that are commonly used in medical prediction tasks. The latent factors derived from Factor Analysis provide an interpretable representation of dominant underlying variables associated with epilepsy risk, thereby offering insights into the structure of the data rather than relying solely on raw feature importance. At the same time, the Decision Tree structure enables transparent tracing of classification rules, allowing the decision-making process of the model to be examined and understood. This level of interpretability is particularly relevant in medical data-driven applications, where model transparency is often considered as important as predictive performance. Consequently, the proposed model is not only indicative of competitive predictive capability but also supports result interpretability, which may facilitate its adoption in clinical decision support systems. Nevertheless, further validation involving domain experts and external datasets is required to assess the practical usefulness and clinical relevance of the extracted factors and decision rules. When compared with previous studies, several patterns can be observed. To further contextualize the results, a comparison between the proposed approach and previous studies is summarized in Table 8.

Table 8. Comparison between Previous Studies and the Proposed Method

No	Author & Year	Method	Dataset	Comparison with This Study
1	Costa et al. (2024)	General Machine Learning	EEG	This study is simpler and more interpretable, making it more suitable for clinical applications
2	Pontes et al. (2024)	SVM + Concept Drift	Dynamic EEG	This study is more stable and easier to implement in general clinical settings
3	Zhang et al. (2024)	CNN-GRU + Attention Mechanism	EEG	This study offers higher transparency and better interpretability for clinical decision-making
4	Yuan et al. (2024)	DenseNet + Vision Transformer	EEG	This study is computationally lighter and more efficient for implementation
5	Xu et al. (2022)	Gradient Boosting Decision Tree	EEG	This study incorporates dimensionality reduction, resulting in improved clinical interpretability
6	Kunekar et al. (2024)	ML & Deep Learning Review	Various datasets	This study focuses more on clinically interpretable latent factor modeling
7	This study	EFA + Decision Tree	Public clinical (non-EEG) dataset	Offers a balance between clinical interpretability and predictive performance with a transparent model structure

Table 8 shows that most recent studies in epilepsy prediction are dominated by deep learning and complex machine learning approaches that primarily focus on improving predictive accuracy. Although these methods are capable of achieving high performance, their main limitation lies in low interpretability, which restricts their direct applicability in clinical settings. In contrast, this study proposes a simpler yet more

interpretable approach by integrating Exploratory Factor Analysis (EFA) and Decision Tree. This hybrid framework enables dimensionality reduction while preserving clinical meaning through the construction of interpretable latent factors. In addition, the use of Decision Tree provides transparency in the decision-making process, which is highly important in medical applications.

However, compared to ensemble and deep learning-based methods, this study still exhibits limitations in terms of precision. This indicates a trade-off between interpretability and predictive performance, where simpler and more interpretable models tend to be easier to understand but require further optimization to improve positive class prediction accuracy.

The main contribution of this study is the proposal of a hybrid framework that combines Exploratory Factor Analysis (EFA) and Decision Tree for epilepsy onset prediction, integrating dimensionality reduction and interpretable classification within a unified workflow. The proposed hybrid framework offers a meaningful balance between predictive performance and interpretability. Unlike purely data-driven deep learning approaches, this method emphasizes clinically meaningful feature transformation, making it more suitable for real-world medical decision support systems. Furthermore, by incorporating latent factor modeling prior to classification, the proposed approach reduces feature redundancy while preserving essential clinical information. This design enhances model transparency without significantly compromising predictive capability, thereby supporting its applicability in early epilepsy screening and assisting clinicians in understanding the underlying risk structure of the data.

This study transforms high-dimensional clinical variables into latent factors that represent meaningful clinical risk structures, thereby improving interpretability compared to conventional feature-based models. The Decision Tree model is then applied to the factor-level representation, enabling transparent, rule-based decision-making that is suitable for clinical interpretation. This study provides a clinically oriented alternative to black-box deep learning models, making it appropriate for early-stage epilepsy screening and clinical decision support systems. Future work may focus on improving precision through ensemble learning techniques or hybrid optimization strategies while maintaining the interpretability advantages of the current framework. In addition, further validation using larger, multi-center, or real-time datasets could enhance the generalizability and clinical applicability of the proposed model.

4. Conclusions

This study presents a hybrid approach combining Factor Analysis and a Decision Tree for early screening of epilepsy onset. Factor Analysis reduced 15 observed variables into five latent factors, which were used as inputs to the Decision Tree. The model identified FAC2_1 (Clinical Neurological Abnormalities) as the root node, emphasizing its key role in classification. Evaluation results show an accuracy of 88.14%, sensitivity of 80.77%, and specificity of 88.73%, demonstrating reliable detection of epilepsy cases while minimizing misclassification of non-seizure instances. Although precision is relatively low

(36.52%), the higher sensitivity makes the model suitable for early-stage screening where minimizing missed detections is critical. The F1-score of 50.30% reflects a moderate balance between precision and recall. Compared to PCA-based or ensemble methods, this framework offers superior interpretability. Factor Analysis provides clinically meaningful insights, while the Decision Tree generates transparent classification rules, enhancing model explainability for medical decision support.

Future work includes improving precision and robustness through ensemble learning, cost-sensitive classification, or hybrid models that integrate interpretable and deep learning techniques. Additionally, validating the model on multi-center or real-time EEG datasets will assess generalizability and clinical applicability. Incorporating temporal dynamics or multimodal patient data could further strengthen predictive performance and support the development of practical, real-time seizure monitoring systems.

Acknowledgements

The authors would like to express their gratitude to Hang Tuah University, Pekanbaru, for their ongoing support and research facilities, which have significantly contributed to the successful completion of this research.

References

- [1] E. Beghi, "The Epidemiology of Epilepsy," *Neuroepidemiology*, vol. 54, no. 2, pp. 185–191, Mar. 2020, <https://doi.org/10.1159/000503831>
- [2] "Epilepsy: a public health imperative," Jun. 2019. <https://www.who.int/publications/i/item/epilepsy-a-public-health-imperative> (accessed Aug. 26, 2025).
- [3] V. L. Feigin et al., "Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021," *Lancet Neurol.*, vol. 23, no. 10, pp. 973–1003, Oct. 2024, doi:[10.1016/S1474-4422\(24\)00369-7](https://doi.org/10.1016/S1474-4422(24)00369-7).
- [4] E. Boran et al., "High-density ECoG improves the detection of high frequency oscillations that predict seizure outcome," *Clin. Neurophysiol.*, vol. 130, no. 10, pp. 1882–1888, Oct. 2019, doi: [10.1016/j.clinph.2019.07.008](https://doi.org/10.1016/j.clinph.2019.07.008).
- [5] S. B. Dumanis, J. A. French, C. Bernard, G. A. Worrell, and B. E. Fureman, "Seizure forecasting from idea to reality. Outcomes of the my seizure gauge epilepsy innovation institute workshop," *eNeuro*, vol. 4, no. 6, Nov. 2017, <https://doi.org/10.1523/ENEURO.0349-17.2017>
- [6] X. Li, Y. Qiao, L. Duan, and J. Miao, "Epilepsy EEG signals classification based on sparse principal component logistic regression model," *Comput. Methods Biomech. Biomed. Engin.*, vol. 28, no. 8, pp. 1295–1303, 2025, <https://doi.org/10.1080/10255842.2024.2321991>
- [7] U. R. Acharya, S. Vinitha Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated EEG analysis of epilepsy: A review," *Knowledge-Based Syst.*, vol. 45, pp. 147–165, Jun. 2013, <https://doi.org/10.1016/j.knsys.2013.02.01>
- [8] M. C. Guerrero, J. S. Parada, and H. E. Espitia, "Principal Components Analysis of EEG Signals for Epileptic Patient Identification," *Comput. 2021, Vol. 9, Page 133*, vol. 9, no. 12, p. 133, Dec. 2021, doi: [10.3390/computation9120133](https://doi.org/10.3390/computation9120133)
- [9] X. Zhang, X. Zhang, Q. Huang, and F. Chen, "A review of epilepsy detection and prediction methods based on EEG signal processing and deep learning," *Front. Neurosci.*, vol. 18, p. 1468967, Nov. 2024, doi: [10.3389/fnins.2024.1468967](https://doi.org/10.3389/fnins.2024.1468967).

- [10] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, pp. 1200–1205, Jul. 2015, doi: [10.1109/MIPRO.2015.7160458](https://doi.org/10.1109/MIPRO.2015.7160458).
- [11] M. Simul, H. Talukder, and R. Bin Sulaiman, "Comparative Analysis of Epileptic Seizure Prediction: Exploring Diverse Pre-Processing Techniques and Machine Learning Models," Aug. 2023, Accessed: Jan. 10, 2026. [Online]. Available: <https://doi.org/10.48550/arXiv.2308.05176>
- [12] L. Bai, G. Litscher, and X. Li, "Epileptic Seizure Detection Using Machine Learning: A Systematic Review and Meta-Analysis," *Brain Sci.*, vol. 15, no. 6, Jun. 2025, <https://doi.org/10.3390/brainsci15060634>
- [13] K. Tsiouris, V. C. Pezoulas, M. Zervakis, S. Konitsiotis, D. D. Koutsouris, and D. I. Fotiadis, "A Long Short-Term Memory deep learning network for the prediction of epileptic seizures using EEG signals," *Comput. Biol. Med.*, vol. 99, pp. 24–37, Aug. 2018, doi: [10.1016/j.compbiomed.2018.05.019](https://doi.org/10.1016/j.compbiomed.2018.05.019).
- [14] Y. Berrich and Z. Guennoun, "EEG-based epilepsy detection using CNN-SVM and DNN-SVM with feature dimensionality reduction by PCA," *Sci. Reports* 2025 151, vol. 15, no. 1, pp. 14313–, Apr. 2025, doi: [10.1038/s41598-025-95831-z](https://doi.org/10.1038/s41598-025-95831-z)
- [15] S. Wong et al., "EEG datasets for seizure detection and prediction- A review," *Epilepsia open*, vol. 8, no. 2, pp. 252–267, Jun. 2023, doi: [10.1002/epi4.12704](https://doi.org/10.1002/epi4.12704)
- [16] M. Farooq, A. Zulfikar, S. R.- Diagnostics, and undefined2023, "Epileptic seizure detection using machine learning: taxonomy, opportunities, and challenges," Accessed: May 06, 2026. [Online], <https://doi.org/10.3390/diagnostics13061058>
- [17] T. Mar Ingolfsson et al., "Minimizing artifact-induced false-alarms for seizure detection in wearable EEG devices with gradient-boosted tree classifiers," *nature.comTM Ingolfsson, S Benatti, X Wang, A Bernini, P Ducouret, P Ryvlin, S Beniczky, L BeniniScientific reports*, 2024•nature.com, vol. 14, p. 2980, 2024, doi: [10.1038/s41598-024-52551-0](https://doi.org/10.1038/s41598-024-52551-0)
- [18] E. D. Pontes, M. Pinto, F. Lopes, and C. Teixeira, "Concept-drifts adaptation for machine learning EEG epilepsy seizure prediction," *Sci. Rep.*, vol. 14, no. 1, pp. 1–13, Dec. 2024, doi: [10.1038/s41598-024-57744-1](https://doi.org/10.1038/s41598-024-57744-1);SUBJMETA.
- [19] X. Zhang, X. Zhang, Q. Huang, and F. Chen, "A review of epilepsy detection and prediction methods based on EEG signal processing and deep learning," *Front. Neurosci.*, vol. 18, 2024, doi: [10.3389/fnins.2024.1468967](https://doi.org/10.3389/fnins.2024.1468967).
- [20] H. Hairani, A. Anggrawan, A. Ningsih, K. Marzuki, L. G. R. Putra, and V. C. Rebecca, "Anemia Classification using Random Forest, SVM, and Naive Bayes Methods," pp. 1–5, Nov. 2025, doi: [10.1109/ICEEIE66203.2025.11252384](https://doi.org/10.1109/ICEEIE66203.2025.11252384).
- [21] A. Anuragi, D. S. Sisodia, and R. B. Pachori, "Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets: an empirical review," *SpringerA Anuragi, DS Sisodia, RB PachoriArtificial Intell. Rev.* 2024•Springer, vol. 57, no. 3, p. 75, Mar. 2024, <https://doi.org/10.1007/s10462-024-10711-8>
- [22] J. Caterson, A. Lewin, and E. Williamson, "The application of explainable artificial intelligence (XAI) in electronic health record research: A scoping review," *journals.sagepub.comJ Cater. A Lewin, E WilliamsonDigital Heal.* 2024•journals.sagepub.com, vol. 10, Jan. 2024, doi: [10.1177/20552076241272657](https://doi.org/10.1177/20552076241272657).
- [23] N. Aziz, A. Manzoor, M. Qureshi, and W. Rashwan, "Explainable AI in healthcare: systematic review of clinical decision support systems," *medrxiv.orgNA Aziz, A Manzoor, MDM Qureshi, MA Qureshi, W RashwanmedRxiv.* 2024•medrxiv.org, 2024, doi: [10.1101/2024.08.10.24311735](https://doi.org/10.1101/2024.08.10.24311735)
- [24] N. Freyer, D. Groß, and M. Lipprandt, "The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons," *SpringerN Freyer, D Groß, M LipprandtBMC Med. Ethics*, 2024•Springer, vol. 25, no. 1, Dec. 2024, doi: [10.1186/s12910-024-01103-2](https://doi.org/10.1186/s12910-024-01103-2).
- [25] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review," *SpringerM Frasca, D La Torre, G Pravettoni, I CuticaDiscover Artif. Intell.* 2024•Springer, vol. 4, no. 1, p. 15, Dec. 2024, doi: [10.1007/s44163-024-00114-7](https://doi.org/10.1007/s44163-024-00114-7)
- [26] H. F. Kaiser, "An Index of Factorial Simplicity," *Psychometrika*, vol. 39, no. 1, pp. 31–36, Mar. 1974, <https://doi.org/10.1007/BF02291575>
- [27] J. F. Hair, "Multivariate Data Analysis - Joseph F. Hair, William C. Black, Barry J. Babin, Rolph E. Anderson - Google Buku," 2019. https://books.google.co.id/books/about/Multivariate_Data_Analysis.html?id=0R9ZswEACAAJ&redir_esc=y (accessed Aug. 26, 2025).
- [28] R. Ottman, "Genetic epidemiology of epilepsy," *Epidemiol. Rev.*, vol. 19, no. 1, pp. 120–128, 1997, doi: [10.1093/oxfordjournals.epirev.a017934](https://doi.org/10.1093/oxfordjournals.epirev.a017934)
- [29] A. K. Ngugi, C. Bottomley, I. Kleinschmidt, J. W. Sander, and C. R. Newton, "Estimation of the burden of active and lifetime epilepsy: a meta-analytic approach," *Epilepsia*, vol. 51, no. 5, pp. 883–890, May 2010, doi: [10.1111/j.1528-1167.2009.02481](https://doi.org/10.1111/j.1528-1167.2009.02481).
- [30] V. L. Feigin et al., "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, Jan. 2022, doi: [10.1177/17474930211065917](https://doi.org/10.1177/17474930211065917)
- [31] "A Clinical Guide to Epileptic Syndromes and Their Treatment, 2nd ed," *AJNR Am. J. Neuroradiol.*, vol. 29, no. 9, p. e86, Jun. 2008, doi: [10.3174/ajnr.A1231](https://doi.org/10.3174/ajnr.A1231).
- [32] F. M. C. Besag, "Epilepsy in patients with autism: links, risks and treatment challenges," *Neuropsychiatr. Dis. Treat.*, vol. 14, pp. 1–10, Dec. 2017, doi: [10.2147/NDT.S120509](https://doi.org/10.2147/NDT.S120509).
- [33] K. Morales et al., "Stratifying cognitive and behavioral comorbidities in children with new-onset seizures – The influence of sociodemographic disadvantage," *Epilepsy Behav.*, vol. 165, Apr. 2025, doi: [10.1016/j.yebeh.2025.110267](https://doi.org/10.1016/j.yebeh.2025.110267).
- [34] J. E. Lawn et al., "Every Newborn: Progress, priorities, and potential beyond Survival," *Lancet*, vol. 384, no. 9938, pp. 189–205, Jul. 2014, doi: [10.1016/S0140-6736\(14\)60496-7](https://doi.org/10.1016/S0140-6736(14)60496-7).
- [35] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African J. Biomed. Res.*, pp. 4023–4031, Nov. 2024, <https://doi.org/10.53555/AJBR.v27i4S.4345>.
- [36] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: [10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239).
- [37] W. Rahayu et al., "Synthetic Minority Oversampling Technique (SMOTE) for Boosting the Accuracy of C4.5 Algorithm Model," *J. Artif. Intell. Eng. Appl.*, vol. 3, no. 3, pp. 624–630, Jun. 2024, doi: <https://doi.org/10.59934/jaiea.v3i3.469>
- [38] Y. Han and I. Joe, "Enhancing Machine Learning Models Through PCA, SMOTE-ENN, and Stochastic Weighted Averaging," *Appl. Sci.* 2024, Vol. 14, Page 9772, vol. 14, no. 21, p. 9772, Oct. 2024. <https://doi.org/10.3390/app14219772>