Applied Soft Computing Journal 202 (2026) 115590

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Explainable ensemble learning framework for anxiety and stress detection in pediatric sleep EEG

Maria Bashir [a], Shaker El-Sappagh [b,1], Waleed Nazih [c], Abolghasem Sadeghi-Niaraki [d], Tamer Abuhmed [b,*]

[a] Department of Electrical and Computer Engineering, College of Information and Communication Engineering, Sungkyunkwan University, Suwon, 16419, South Korea
[b] Department of Computer Science and Engineering, College of Computing and Informatics, Sungkyunkwan University, Suwon, 16419, South Korea
[c] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-kharj, Saudi Arabia
[d] Department of Computer Science and Convergence Engineering for Intelligent Drone, XR Research Center, Sejong University, Seoul, South Korea

HIGHLIGHTS

- Introduces an explainable ensemble-learning framework for pediatric anxiety–stress screening using routine sleep EEG recordings.
- Uses a clinically verified cohort of 532 children from NCH Sleep DataBank with six EEG channels and patient-level validation.
- Shows that Approximate Entropy and Standard Deviation provide a compact, effective feature representation for sleep EEG classification.
- Achieves 86.30% accuracy and 89.20% AUC with a stacking ensemble, outperforming individual classifiers and dynamic ensemble baselines.
- Combines SHAP, LIME, and NeuroXAI to reveal frontal markers for anxiety and frontal-central EEG patterns associated with stress.

ARTICLE INFO

Keywords:
Anxiety
Stress
Pediatric sleep EEG
Ensemble machine learning
Explainable AI

ABSTRACT

Anxiety and stress are common pediatric mental health conditions that often emerge during childhood and adolescence and can significantly affect long-term well-being. However, differentiating between these conditions remains challenging because their symptoms often overlap, and clinical assessment can be subjective. This study proposes an explainable ensemble learning framework to support the objective screening of pediatric anxiety and stress using sleep Electroencephalogram (EEG) signals. A clinically verified cohort of 532 children, including 247 with anxiety and 285 with stress, was selected from the Nationwide Children's Hospital Sleep DataBank. Six scalp EEG channels, F3, F4, C3, C4, O1, and O2, were analyzed using a standardized two-hour sleep window after signal filtering, artifact rejection, and annotation alignment. Approximate Entropy and Standard Deviation features were extracted from 30s epochs and used to train classical classifiers, static ensemble models, and dynamic ensemble selection methods optimized through Bayesian search. The proposed models were evaluated using patient-level stratified cross-validation to prevent data leakage and were further assessed on an independent external cohort from the Boston Children's Hospital Sleep Corpus to examine cross cohort generalizability. Explainability was integrated using SHAP, LIME, and NeuroXAI to provide global, local, and neurophysiologically informed interpretations of model behavior. Compared with prior EEG studies, which are often limited by small cohorts, adult participants, or short task-based recordings, the proposed framework provides a lightweight, interpretable, and clinically relevant approach based on routine pediatric sleep EEG data. This work highlights the potential of explainable sleep EEG-based machine learning to support scalable, noninvasive, and transparent screening assistance for pediatric anxiety and stress.

* Corresponding author.
Email addresses: mariabashir@g.skku.edu (M. Bashir), shaker.elsappagh@gu.edu.eg (S. El-Sappagh), w.nazeeh@psau.edu.sa (W. Nazih), a.sadeghi@sejong.ac.kr (A. Sadeghi-Niaraki), tamer@skku.edu (T. Abuhmed).
[1] Present address: Faculty of Computer Science and Engineering, Galala University, Suez, 435611, Egypt; Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Benha, 13518, Egypt.

https://doi.org/10.1016/j.asoc.2026.115590
Received 1 October 2025; Received in revised form 1 May 2026; Accepted 18 May 2026
Available online 3 June 2026

1. Introduction

Mental health disorders are a significant public health concern that adversely affect individuals across all age groups. Those diagnosed with severe mental illnesses face a markedly increased risk of premature mortality, with life expectancy reduced by about 30 years in low-income countries and around 20 years in high-income nations [1]. The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), underscores the profound societal and personal burden of mental disorders, defining over 45 distinct conditions [2]. Although mental health disorders are highly prevalent, studies indicate that only around one-third of affected individuals actually receive appropriate clinical care [3]. This persistent gap highlights the urgent need for early and reliable screening and decision-support strategies. Anxiety and stress are among the most prevalent pediatric mental health conditions, typically emerging during childhood and adolescence. Together, they affect nearly one in three individuals by late adolescence [4]. They are associated with substantial functional impairment, poor academic performance, and an increased risk of later depression and psychiatric disorders. Differentiating anxiety from stress remains clinically challenging because both conditions share overlapping physiological and behavioral symptoms, while current assessments often depend on subjective questionnaires, interviews, and behavioral interpretation. Stress is highly prevalent in children with chronic conditions, such as inflammatory bowel disease, where it frequently co-occurs with anxiety and worsens clinical outcomes [5]. Although anxiety tends to be more persistent and anticipatory, stress is more situational and reactive; their overlap makes reliable differentiation difficult [4]. This overlap highlights the urgent need for objective biomarkers that can support early detection and separation of pediatric anxiety and stress. Several approaches have been proposed to address the diagnosis of mental health disorders, including neurophysiological measurements [6], neurological evaluations [7], and various neuroimaging techniques [8]. These approaches fall broadly into functional and structural categories: functional techniques such as electroencephalography (EEG) [9], magnetoencephalography (MEG) [10], functional magnetic resonance imaging (fMRI) [11], and positron emission tomography (PET) [12] capture real-time brain activity, while structural methods such as structural MRI (sMRI) [13], diffusion tensor imaging (DTI) [14], and computed tomography (CT) [15] provide anatomical information.

Among these, EEG has emerged as a practical tool due to its portability, non-invasive nature, cost-effectiveness, and high temporal resolution. It has been extensively applied in mental health research, supporting diagnoses across a range of conditions, including anxiety, depression, stress, schizophrenia, epilepsy, Parkinson's disease, and Alzheimer's disease [16–21]. By recording electrical fluctuations across scalp electrodes, EEG offers insights into the large-scale activity of neural populations [22,23]. Unlike task-based paradigms that demand active participation, sleep EEG provides a stable and clinically practical window for assessment, already integrated into pediatric hospitals through routine polysomnography. Research has shown that sleep-derived EEG signals capture developmental markers that are predictive of brain maturation and long-term outcomes. For instance, Ventura et al. [24] showed that infant EEG can reflect neurodevelopment as early as four months, and [25] reported similar findings across the first six months of life. These studies confirm that pediatric EEG carries valuable information about brain function, but prior research has largely emphasized developmental trajectories rather than mental health. Existing pediatric EEG work has primarily focused on epilepsy, attention-deficit/hyperactivity disorder, and developmental monitoring, while large-scale sleep EEG studies targeting pediatric anxiety and stress remain largely underexplored [26,27]. Machine learning (ML) has been increasingly applied to EEG for emotion recognition and disorder classification, with techniques such as support vector machines, random forests, and convolutional neural networks reporting promising results in adult stress and anxiety detection [28–30]. Multimodal approaches have also gained traction, combining EEG with cardiac signals [31], video [32,33], voice [34], and text [35,36], opening new possibilities for comprehensive and personalized assessments. Ensemble methods [37,38], which combine multiple classifiers, have demonstrated improved robustness and generalization compared with single models, making them particularly suited for heterogeneous clinical data. Although machine learning has shown promise for EEG-based mental health assessment, most existing studies have been conducted in small adult cohorts, relied on short task-based recordings, and provided limited model interpretability. In parallel, pediatric EEG research has mainly focused on developmental monitoring, epilepsy, and attention-deficit/hyperactivity disorder, with limited attention to anxiety and stress. Therefore, large-scale pediatric sleep EEG studies that combine ensemble learning with explainable artificial intelligence for anxiety and stress differentiation remain largely absent. This gap motivates the proposed explainable ensemble learning framework.

The main contributions of this study are summarized as follows:

- We present a large-scale pediatric sleep EEG study for differentiating between anxiety and stress using a clinically verified cohort of 532 children from the NCH Sleep DataBank, including 247 anxiety cases and 285 stress cases.
- We develop an end-to-end machine learning pipeline that integrates EEG preprocessing, sleep-stage annotation alignment, artifact rejection, standardized epoching, feature extraction, model training, hyperparameter optimization, and explainability analysis.
- We systematically evaluate multiple feature configurations, signal durations, EEG channel combinations, filter settings, classical classifiers, static ensemble models, and dynamic ensemble selection methods to identify the most effective configuration for anxiety and stress classification.
- We show that the combination of Approximate Entropy and Standard Deviation extracted from six EEG channels over a standardized two-hour sleep window provides the strongest feature representation among the tested configurations.
- We perform patient-level stratified cross-validation to ensure that patients do not overlap across training, validation, and testing partitions, reducing the risk of patient-level data leakage and supporting a more reliable internal evaluation.
- We conduct external validation using an independent cohort from the Boston Children's Hospital Sleep Corpus to assess the cross-cohort generalizability of the proposed anxiety and stress classification framework.
- We provide additional sensitivity analyses, including EEG segment-duration analysis, channel-configuration analysis, age-stratified evaluation, and multiclass diagnostic extension, to examine the robustness and clinical relevance of the proposed framework.
- We integrate SHAP, LIME, and NeuroXAI to provide complementary global, local, and channel-level explanations, revealing that frontal activity contributes strongly to anxiety classification, while combined frontal-central activity is more informative for stress.
- We demonstrate the potential of lightweight and interpretable machine learning applied to routine pediatric sleep EEG as a scalable, non-invasive, and transparent screening support framework for pediatric mental health assessment.

To the best of our knowledge, this is the first large-scale explainable pediatric sleep EEG framework for differentiating between anxiety and stress using routine clinical recordings. Unlike traditional assessment approaches that rely heavily on subjective self-reports, interviews, and behavioral interpretation, the proposed method provides physiological evidence that can support clinical screening and decision-making. By integrating ensemble learning with SHAP, LIME, and NeuroXAI, the framework offers both predictive performance and transparent feature- and channel-level explanations. This combination of performance and interpretability supports the development of scalable, non-invasive, and trustworthy screening support tools for pediatric mental health assessment.

The remainder of this paper is organized as follows. Section 2 surveys existing research on EEG-based mental health detection, with an emphasis on anxiety and stress. Section 3 presents the proposed framework, including the dataset, EEG signal preprocessing steps, feature extraction methods, patient-level partitioning, classification models and their optimization, and explainability techniques used in this study. Section 4 provides details of the experimental design. Section 5 presents and analyzes the experimental findings, including the comparative evaluation of classifiers, ensemble models, and dynamic selection methods, external validation to assess cross-cohort generalizability, and interpretability results obtained using SHAP, LIME, and NeuroXAI. This section also reports sensitivity analyses, including a multiclass diagnostic extension and age-stratified evaluation. Section 6 outlines the current limitations and directions for future research. Finally, Section 7 summarizes the key contributions and findings of this work.

2. Related work

Research on pediatric EEG has predominantly focused on neurological disorders such as epilepsy, ADHD, and developmental outcomes. These studies highlight the utility of EEG in pediatric clinical contexts, demonstrating its value for diagnostic and developmental monitoring. For example, Ventura et al. [24] showed that infant sleep EEG features such as delta power and spindles at four months predict developmental outcomes at 18 months, while another study [25] reported that longitudinal changes in slow-wave, theta, and sigma activity between three and six months were linked to later motor and social outcomes. Beyond developmental outcomes, EEG has been extensively applied to epilepsy detection [26,27,39] and ADHD diagnosis [40]. However, these applications have largely emphasized neurological and developmental trajectories rather than mental health. Despite the high prevalence of pediatric anxiety and stress, EEG-based approaches targeting these conditions remain scarce. To situate our contribution, we therefore review the broader body of related studies on anxiety and stress detection, the majority of which have been conducted in adults.

2.1. Anxiety detection and classification models

EEG has been widely explored for anxiety detection, often combined with machine learning to enhance classification accuracy. Aldayel and Al-Nafjan [41] compared several feature extraction and classification methods, reporting that Random Forest achieved the best accuracy (87.5%) for distinguishing anxious from non-anxious participants. However, their dataset was limited to ten EEG channels and did not examine the contribution of individual electrodes. Shen et al. [42] proposed a multidimensional EEG analysis framework, achieving 97.83% accuracy using an SVM model, yet their dataset included fewer than 100 participants, raising concerns about representativeness. Sakib et al. [43] studied only 40 subjects and showed that artifact removal improved detection, particularly in frontal regions, but channel-level analyses were not pursued. Li et al. [44] applied SVM and KNN to classify four anxiety states, but achieved an accuracy of only 62.56% on a sample of 14 individuals. Luo et al. [45] employed ensemble methods to classify generalized anxiety disorder, reporting 98.1% accuracy with 119 participants, though imbalanced classes may have affected outcomes. Other approaches, such as multimodal fusion, have also been explored: One study [46] integrated EEG and ECG with Gradient Tree Boosting, achieving 97.3% accuracy, while [47] reported 96.1% accuracy using Naïve Bayes on a pilot dataset of only 16 subjects. Collectively, these studies report high accuracies but rely on very small samples, limiting generalizability and raising risks of overfitting. Moreover, few works provide insights into the relative contributions of individual EEG channels, which may reveal physiologically meaningful patterns.

2.2. Stress detection and classification models

Parallel efforts have examined EEG-based stress detection. Jebelli et al. [48] extracted statistical and spectral features, achieving up to 80.32% accuracy using SVM, though their study involved only seven participants and was vulnerable to noise and motion artifacts. Umar Saeed et al. [49] used a single-channel EEG to classify stress versus non-stress states, reporting 78.5% accuracy using SVM. While simpler and portable, this approach was limited in spatial resolution. Minguillon et al. [50] combined EEG with ECG, EMG, and GSR to classify stress into three levels, achieving 86% accuracy, though results were highly dependent on rigorous preprocessing. So et al. [51] focused on mental workload detection using a single frontal channel, reaching 75% accuracy using SVM, suggesting frontal activity as an important marker but again limited by sample size. Rajendran et al. [52] evaluated multiple classifiers, including Naïve Bayes, KNN, SVM, and ensembles, and found SVM achieved the best accuracy (89.74%) on eight-channel EEG, though still constrained by sample size. More recently, Ince et al. [53] proposed a CubicPattKNN framework, reporting 96.29% accuracy using 14-channel EEG data, but their dataset remained relatively modest.

2.3. Summary and research gap

Existing EEG-based anxiety and stress detection studies demonstrate the potential of neurophysiological signals for mental health assessment. However, most studies have been conducted in small adult cohorts, often using task-evoked or short-duration EEG recordings, which limits their generalizability to routine pediatric clinical settings. In addition, reported accuracies are sometimes high despite limited sample sizes, raising concerns about overfitting and reproducibility. Most prior studies also provide limited interpretability, making it difficult to understand which EEG channels or signal characteristics contribute to model decisions. In contrast, pediatric EEG research has primarily focused on developmental monitoring, epilepsy, and attention-deficit/hyperactivity disorder, with limited investigation into anxiety and stress. Large-scale pediatric sleep EEG studies that jointly differentiate between anxiety and stress while integrating ensemble learning and explainable artificial intelligence remain scarce. This gap motivates the proposed framework, which uses routine pediatric sleep EEG recordings, compares multiple feature sets and model configurations, and incorporates SHAP, LIME, and NeuroXAI to provide transparent and clinically meaningful explanations.

3. Proposed framework

Fig. 1 provides a simplified overview of the proposed framework for EEG-based screening support of pediatric anxiety and stress using routine sleep EEG data. It summarizes the main pipeline stages, including dataset preparation, EEG preprocessing, data splitting and scaling, feature extraction, model training and optimization, and explainability analysis. Fig. 2 presents the detailed workflow of the framework, showing the specific processing steps applied to the NCH dataset for primary evaluation and the BCH dataset for external validation. The following subsections describe each component of the pipeline in detail, including signal preprocessing, sleep-stage alignment, feature extraction, patient-level data partitioning, model optimization, and explainability analysis.

3.1. Dataset preparation

In this study, EEG recordings were obtained from the Nationwide Children's Hospital Sleep DataBank (NCHSDB) [54], one of the most extensive clinical resources available for pediatric sleep research. The dataset includes 3984 polysomnography (PSG) recordings collected from 3673 unique pediatric patients over a two-year period, between December 2017 and December 2019, in real-world clinical settings. In addition to physiological recordings, the dataset provides demographic and diagnostic information extracted from electronic health records (EHRs), including diagnostic codes mapped according to DSM and ICD terminology. Together, these components provide a comprehensive foundation for exploring EEG-based markers in pediatric

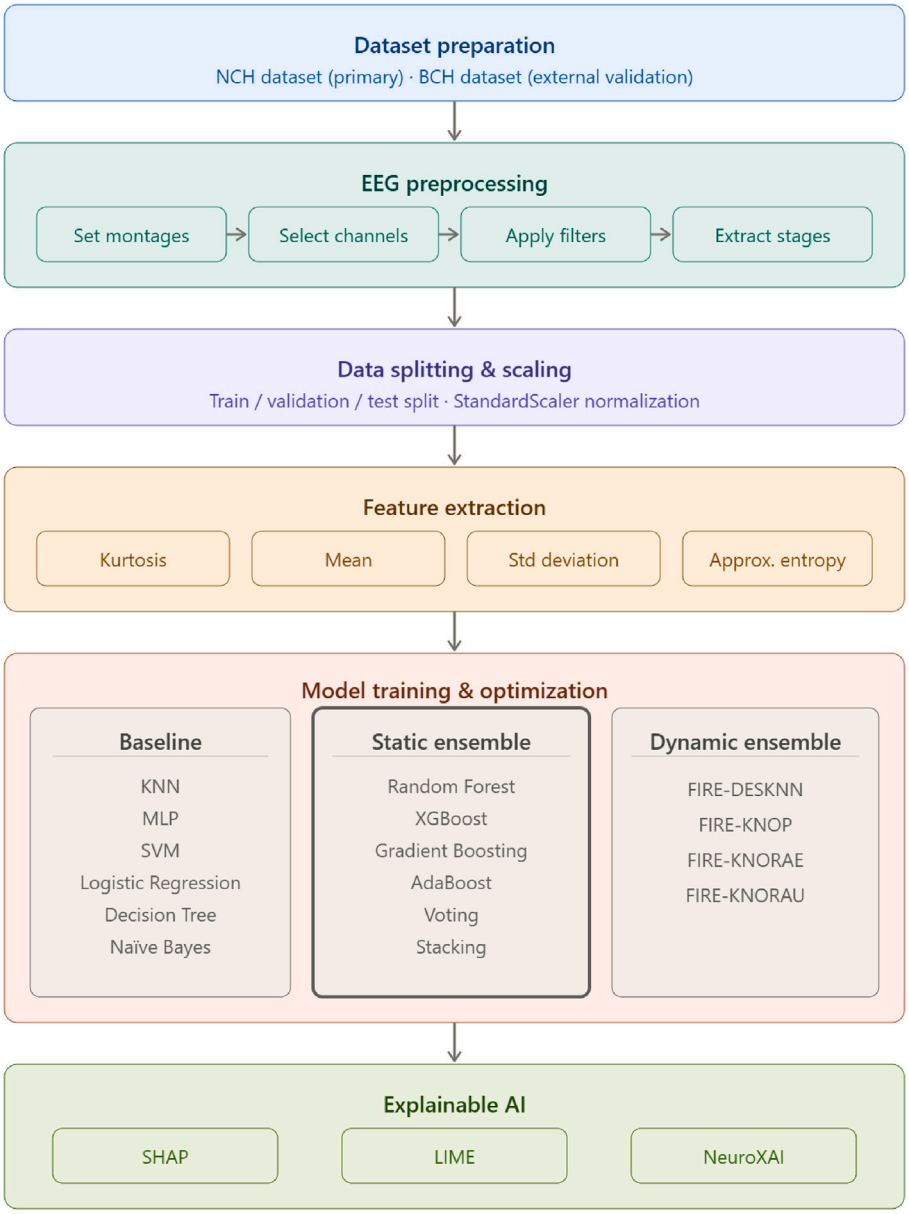


Fig. 1. Simplified overview of the proposed explainable ensemble learning framework for pediatric anxiety and stress screening using sleep EEG. The diagram summarizes the main stages, including dataset preparation, EEG preprocessing, data splitting and scaling, feature extraction, model training and optimization, and explainability analysis.

populations. The sleep data component of the NCHSDB contains annotated PSG recordings provided in European Data Format (EDF), accompanied by corresponding annotation files in Tab-Separated Value (TSV) format. Each PSG file includes multiple physiological signals, including EEG, Electromyography (EMG), Electrooculography (EOG), Electrocardiography (ECG), and respiratory channels. The total number of channels varies across recordings, most commonly including 25, 26, or 29 channels, depending on the clinical protocol. Among these, seven channels correspond to EEG signals, while the remaining channels capture chin muscle tone, eye movements, cardiac activity, and respiratory parameters such as respiratory rate and respiratory effort. A full breakdown of the included channels is provided in Table 1. The recording durations vary substantially, ranging from a few minutes to more than 16 hours, with an average duration of approximately 10.3 hours. Most recordings were sampled at 256 Hz, although a smaller number were recorded at higher sampling rates, including 400 Hz and 512 Hz. The corresponding TSV annotation files provide detailed sleep-stage labels and event markers, making the dataset suitable for both clinical analysis and sleep-related machine learning research. The health data component provides extensive demographic and clinical information extracted from EHRs. It includes patient demographics, sleep study information, medication records, diagnostic codes, vital signs, procedural histories, and healthcare encounters. With more than 5.6 million entries, these data entries are systematically linked to the corresponding PSG recordings using unique patient identifiers, enabling integrated patient-level analysis across physiological and clinical dimensions. The dataset contains more than 5 million annotated events across 3984 PSG files. Each annotation file specifies the onset time, duration, and event label, including sleep stages and physiological events. Approximately 79.48% of all annotations correspond to sleep staging, providing a strong foundation for automated sleep-stage-aware analysis. Sleep stages are categorized as Wakefulness (W), Rapid Eye Movement (REM), labeled as R, and

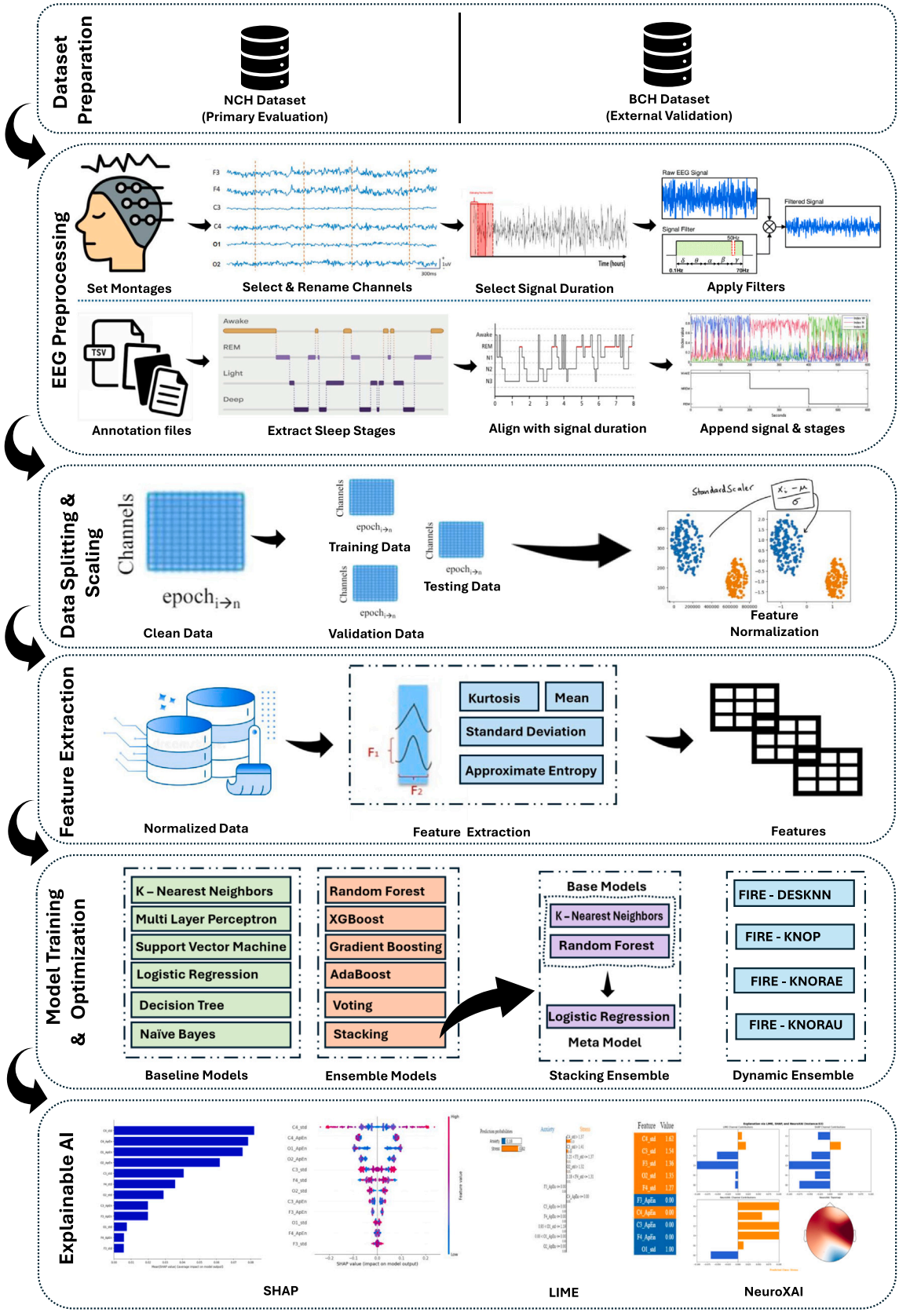


Fig. 2. Detailed workflow of the proposed explainable ensemble learning framework for pediatric sleep EEG analysis and screening support. The pipeline includes NCH data preparation for primary evaluation, BCH data for external validation, EEG preprocessing, sleep-stage alignment, patient-level data splitting and scaling, feature extraction, classifier training and optimization, ensemble learning, dynamic ensemble selection, and explainability using SHAP, LIME, and NeuroXAI.

Table 1
List of all EEG channels and a selection of commonly included non-EEG channels found in the EDF files of the NCHSDB [54]. Brief descriptions are provided for the non-EEG channels. This study utilizes the first six EEG channels listed. CO_2 denotes carbon dioxide.

Channel Name	Description
EEG F3-M2	EEG
EEG F4-M1	EEG
EEG C3-M2	EEG
EEG C4-M1	EEG
EEG O1-M2	EEG
EEG O2-M1	EEG
EEG CZ-O1	EEG
SNORE	Measure of snoring or air vibrations
Resp Flow	Airflow channel
OSAT	Oxygen saturation
ECG LA-RA	Electrical activity of the heart
CAPNO	End-tidal CO_2 waveform
EOG LOC-M2	Electrical signals from eye movements
EMG CHIN1-CHIN2	Electrical activity of muscles

Non-Rapid Eye Movement (NREM), which is further subdivided into N1, N2, and N3. In the dataset, these stages are labeled as W, N1, N2, N3, and R.

The NCHSDB includes 3984 pediatric clinical sleep studies from 3673 patients. Among these patients, 3400 underwent a single PSG session, 238 had two sessions, and 35 had more than two sessions, with a maximum of five sessions per patient. To reduce redundancy and minimize temporal variability, only one PSG recording per patient was retained. Patients with multiple available PSG sessions were excluded from the final cohort. For the primary binary classification task, only patients with an exclusive diagnosis of either anxiety or stress were included. Patients diagnosed with both anxiety and stress were excluded from the primary analysis to maintain clear class separation. Patients with other psychiatric or neurological disorders, such as attention-deficit/hyperactivity disorder, epilepsy, or developmental delay, were also excluded to reduce diagnostic confounding. Fig. 3 summarizes the cohort construction process.

The dataset comprised 532 pediatric patients aged 1 to 17 years, consistent with the pediatric scope of the NCHSDB. The final dataset used in this study is summarized in Table 2. Diagnostic labels were assigned using the final diagnosis and the corresponding ICD codes available in the patient-level EHR files. For example, anxiety-related diagnostic labels included codes such as F41.1, corresponding to generalized anxiety disorder, whereas stress-related diagnoses included codes such as F43.0, corresponding to acute stress reaction. Patients with only anxiety-related diagnostic codes were assigned to the anxiety class, whereas patients with only stress-related diagnostic codes were assigned to the stress class. Patients with both anxiety-related and stress-related diagnostic codes were identified as comorbid cases and excluded from the primary binary classification analysis. These comorbid cases were retained only for an additional sensitivity experiment evaluating the effect of including a third diagnostic group. All diagnostic labels used in this study were obtained from hospital EHRs and reflect clinical judgments made by pediatric specialists. For very young children, including those under three years of age, diagnoses were not based solely on self-report. Instead, they were derived from structured pediatric psychiatric evaluations involving clinical observation, caregiver interviews, and standardized diagnostic frameworks adapted from DSM-5/ICD-10 criteria. This procedure supports diagnostic validity across the full pediatric cohort.

Table 2
Specifications of the pediatric sleep EEG cohort used in this study.

Number of Patients	532
Age range	1–17 years
Sampling frequency	256 Hz
EEG channels	F3, F4, C3, C4, O1, and O2
Epoch length	30 s
EEG duration retained per subject	2 hours
Total epochs	12,714
Sleep stages included	W, N1, N2, N3, and R

The overall gender distribution of the final cohort was nearly balanced, with 265 females and 267 males. The dataset included 247 patients diagnosed with anxiety and 285 patients diagnosed with stress. Fig. 4 presents the age distribution of the anxiety and stress groups. The anxiety group was generally older, with a mean age of 11.69 ± 4.01 years, whereas the stress group was younger, with a mean age of 5.44 ± 4.46 years. Further stratification by gender showed that females with anxiety had a mean age of 12.30 ± 4.10 years, compared with 10.90 ± 3.80

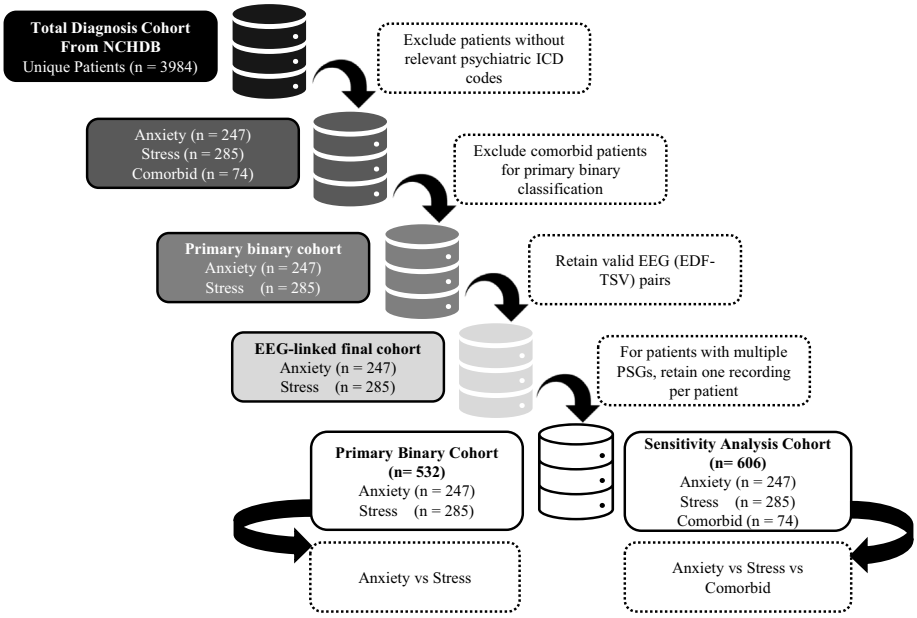


Fig. 3. Cohort construction process for NCHSDB used in this study.

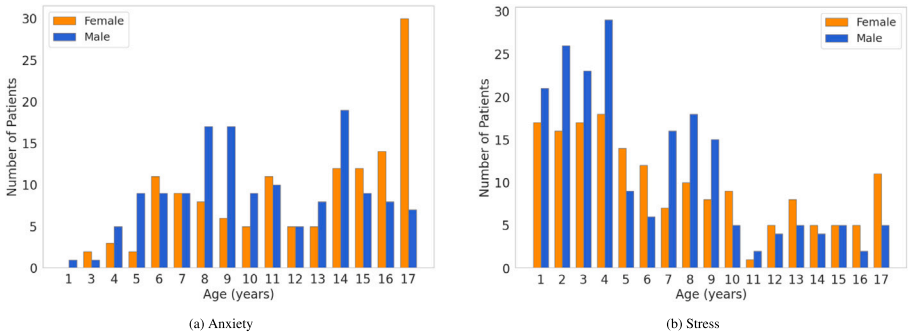


Fig. 4. Age and gender distributions of pediatric patients with anxiety and stress. The figure illustrates the cohort composition used for classification.

years for males with anxiety. Among patients with stress, female patients had a mean age of 5.90 ± 4.60 years, while male patients had a mean age of 4.95 ± 4.20 years. These findings indicate meaningful age and gender variations within the clinical presentation of anxiety and stress in this pediatric cohort. Although Fig. 4 illustrates the age and gender distributions of the final cohort. The final dataset used in this study is summarized in Table 2.

3.2. EEG preprocessing

Raw EEG recordings are commonly affected by artifacts originating from muscle activity, eye blinks, body movements, electrode-related noise, and environmental interference. Such artifacts can distort EEG signal characteristics and may lead to unreliable feature extraction and classification results. Therefore, the initial preprocessing stage focused on improving the quality, consistency, and interpretability of the raw EEG recordings before feature extraction and model training. This stage included channel standardization, montage assignment, resampling, signal filtering, sleep-stage annotation alignment, epoch segmentation, and artifact rejection. These procedures were guided by established EEG preprocessing practices and domain-expert recommendations [55,56].

Each recording originally contained seven EEG channels. However, six channels were retained for analysis: two frontal channels (F3 and F4), two central channels (C3 and C4), and two occipital channels (O1 and O2). The Cz electrode was excluded because it was not consistently available across all patient recordings. This selection ensured a uniform channel configuration across the final cohort. All channel labels were aligned with the standard 10–20 electrode placement system [57], which supports compatibility with established neuroscience literature and facilitates anatomical interpretation of electrode positions. The spatial configuration of the selected electrodes is shown in Fig. 5. A predefined montage configuration was applied to standardize electrode locations, support anatomical localization, and facilitate interpretation of spatial EEG patterns [58]. The montage assignment and signal processing procedures were implemented using the MNE-Python package [59], a widely used open-source library for EEG and MEG analysis [60].

To reduce variability in sampling rates and computational complexity, all EEG signals were resampled to 256 Hz. This sampling rate preserved the EEG frequency content relevant to the selected analysis while maintaining a manageable computational cost [61]. Recordings originally sampled at higher frequencies, such as 400 Hz or 512 Hz, were downsampled to 256 Hz to ensure consistency across patients. The dataset includes paired EEG signal files in EDF format and corresponding annotation files in TSV format containing sleep-stage labels. To identify a stable and informative analysis window, EEG segments ranging from one to four hours after excluding the initial hour were systematically evaluated. The first hour was excluded because it provided insufficient sleep-stage coverage for many recordings. Among the evaluated durations, the two-hour window provided the best trade-off

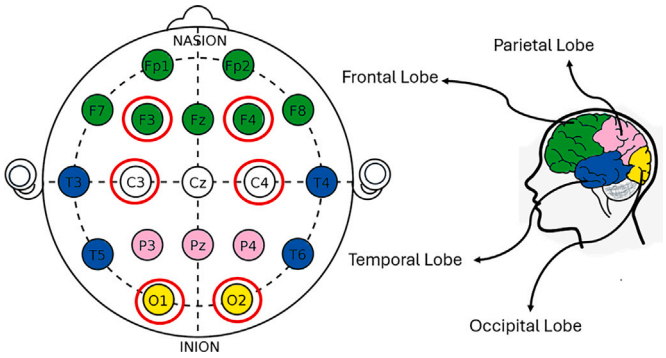


Fig. 5. Channel locations of the EEG signals showing frontal, parietal, temporal, and occipital lobes. Channels in our dataset are highlighted in red. Channel nAmes are according to the international 10–20 system.

between sleep-stage representation, signal stability, and classification performance, as supported by the results in Table 6.

To enhance signal quality while preserving relevant EEG activity, band-pass and notch filtering were applied before epoch segmentation and feature extraction. A band-pass filter was used to retain the frequency range relevant to EEG-based classification while reducing low-frequency drift and high-frequency noise. Because the recordings were collected in the United States, a 60 Hz notch filter was applied to suppress power-line interference. This filtering configuration was applied consistently across all recordings.

The annotation files include several event types, including sleep stages (W, N1, N2, N3, and R), body movements, position changes, and light-related events [54]. In this study, only sleep-stage annotations were retained for alignment with EEG epochs. Other event labels, including movement, position, and light-related events, were excluded from the classification pipeline. The retained sleep-stage labels were aligned with the corresponding EEG segments to generate synchronized sleep EEG epochs with verified sleep-stage information.

After filtering, each recording was segmented into non-overlapping 30-second epochs. Sleep-stage alignment was performed using the absolute recording timeline. For each epoch, the sleep-stage label was assigned by matching the epoch start time to the corresponding annotation interval. Epochs without a valid sleep-stage match were discarded to ensure that all retained EEG segments corresponded to verified annotated sleep intervals within the selected analysis window. As a result, the retained EEG epochs were sleep-stage-aware rather than arbitrary signal segments, preserving their temporal relationship with the underlying sleep architecture. In this study, sleep-stage annotations were used only for temporal alignment and validation of EEG epochs during preprocessing, whereas the classification objective remained the binary differentiation between anxiety and stress. To reduce artifact contamination,

Table 3
Summary of EEG feature extraction methods evaluated in this study.

Feature	Category	Mathematical Formulation	Interpretation
Power Spectral Density using Welch's method	Frequency-domain	$S(f) = \frac{1}{MV}\left\|\sum_{k=0}^{M-1} x_j(k)v(k)e^{-j2\pi f}\right\|^2$ $P_{\text{avg}}(f) = \frac{1}{N}\sum_{m=0}^{N-1} S_m(f)$	Estimates the distribution of EEG signal power across frequency components and supports analysis of rhythmic sleep EEG activity.
Approximate Entropy	Non-linear complexity	$\text{ApEn}(m,r) = \phi_m(r) - \phi_{m+1}(r)$ $\phi_m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \log C_i^m(r)$	Quantifies signal irregularity, complexity, and unpredictability within each EEG epoch.
Mean	Time-domain statistical	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$	Measures the central tendency of EEG amplitude values within an epoch.
Standard Deviation	Time-domain statistical	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$	Measures amplitude variability and dispersion of EEG samples within each epoch.
Kurtosis	Time-domain statistical	$K = \frac{N\sum_{i=1}^{N}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^2}$	Describes the distributional shape of EEG amplitude values and the prominence of extreme values.

we employed a FASTER-based artifact rejection procedure inspired by the Fully Automated Statistical Thresholding for EEG Artifact Rejection (FASTER) algorithm [62]. FASTER is an automated and unsupervised artifact rejection method that detects outliers using statistical parameters computed across different aspects of the EEG signal, including channels, epochs, independent components, and channel-epoch segments. The original method combines statistical thresholding with independent component analysis, contaminated component removal, epoch rejection, and interpolation of noisy channels when applicable. In the present study, artifact rejection was performed after sleep-stage alignment and before feature extraction. Because this study used a compact six-channel EEG montage, the FASTER-based procedure was applied conservatively at the epoch level to identify and remove contaminated EEG epochs. This strategy avoided unstable independent component analysis and extensive channel interpolation in low-density EEG data while preserving subject-level grouping required for patient-wise cross-validation.

3.3. Feature extraction

Feature extraction is an essential step in EEG-based analysis because raw EEG signals are high-dimensional, non-stationary, and highly variable across patients and sleep stages. In automated EEG analysis, feature-based representations help reduce signal complexity while preserving physiologically meaningful information relevant to downstream classification. Since anxiety and stress may influence sleep EEG dynamics through changes in signal variability, spectral composition, and temporal irregularity, different categories of features were considered in this study. EEG features are commonly grouped into linear and non-linear methods. Linear features include time-domain and frequency-domain measures, which describe amplitude variation, central tendency, distributional properties, and spectral power characteristics. Non-linear features capture irregularity, complexity, and unpredictability in EEG dynamics, which are especially important because EEG signals are generated by complex neurophysiological processes and often exhibit non-stationary behavior.

In this study, feature extraction was performed on non-overlapping 30-second EEG epochs obtained from six standardized EEG channels: F3, F4, C3, C4, O1, and O2. The extracted features included Power Spectral Density (PSD) using Welch's method, Approximate Entropy (ApEn), Standard Deviation (Std), Mean, and Kurtosis. These features were selected to represent complementary EEG characteristics, including spectral power, non-linear signal irregularity, amplitude variability, central tendency, and distributional shape. The extracted features were subsequently used as input to ML models for differentiating between anxiety and stress from pediatric sleep EEG recordings.

Time-domain features provide a direct description of EEG amplitude characteristics over time. These features are useful for quantifying signal variation, average amplitude behavior, and distributional changes within each epoch. In this study, Mean, Standard Deviation, and Kurtosis were evaluated as time-domain statistical features. Among these, Standard Deviation was retained in the best-performing feature configuration because it captures amplitude variability within each EEG epoch. Frequency-domain features describe how the EEG signal power is distributed across different frequency components. This is particularly relevant in sleep EEG analysis, where rhythmic activity varies across conventional frequency bands, including delta, theta, alpha, sigma, and beta bands. PSD features were estimated using Welch's method to evaluate spectral power characteristics. Although PSD was included in the feature comparison experiment, its classification performance was lower than the entropy and variability-based feature configuration. Therefore, PSD was retained for experimental comparison but was not selected in the final best-performing feature set. Non-linear features are important for characterizing the irregularity and complexity of EEG signals. Since EEG activity reflects complex neurophysiological processes, non-linear analysis can provide additional information beyond conventional time-domain and frequency-domain measures. In this study, Approximate Entropy was extracted as a non-linear complexity feature to quantify the unpredictability and irregularity of EEG patterns within each epoch. Approximate Entropy was selected in the final feature configuration because it provided complementary information to Standard Deviation. Specifically, Approximate Entropy captures non-linear signal complexity, whereas Standard Deviation captures amplitude variability.

Table 3 summarizes the feature categories, mathematical formulations, and interpretations of the extracted EEG features. For each 30-second epoch, features were extracted independently from each of the six EEG channels. Therefore, the resulting feature representation preserved channel-specific information while converting each EEG epoch into a structured numerical feature vector suitable for machine learning classification. After feature extraction, all features were standardized within each data partitioning fold. To prevent information leakage, the scaler was fitted only on the training subset and then applied to the validation and test subsets using the same scaling parameters. The experimental comparison showed that the combination of Approximate Entropy and Standard Deviation achieved the best classification performance. This feature configuration was therefore selected as the final representation for the proposed anxiety and stress classification framework. Approximate Entropy contributed non-linear complexity information, while Standard Deviation captured amplitude variability, making the two features complementary for pediatric sleep EEG classification.

3.4. Data splitting and scaling

To ensure a robust and unbiased evaluation, all experiments were conducted using patient-level data partitioning. This strategy ensured that epochs from the same patient did not appear across training, validation, and test subsets, thereby preventing patient-level information leakage. Specifically, a stratified 10-fold cross-validation strategy was applied at the patient level to preserve the anxiety and stress class distribution across folds. In each fold, patients were divided into mutually exclusive training, validation, and test subsets. The training subset was used to fit the classifiers, the validation subset was used for model

selection and hyperparameter optimization, and the test subset was used only for final performance evaluation. After patient-level partitioning, feature normalization was applied using the Standard Scaler [63]. To avoid data leakage, the scaler was fitted only on the training subset within each fold and then applied to the corresponding validation and test subsets using the same scaling parameters. This procedure ensured that no information from the validation or test patients was used during feature normalization. Each feature was standardized using $z = \frac{x-\mu_{\text{train}}}{\sigma_{\text{train}}}$, where μ_{train} and σ_{train} denote the mean and standard deviation computed from the training set only.

3.5. Model training and optimization

After feature extraction, patient-level partitioning, and feature normalization, several classical machine learning models, static ensemble models, and dynamic ensemble selection models were trained to classify anxiety and stress states from pediatric sleep EEG features. The classical machine learning models included K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), XGBoost, Gradient Boosting, and AdaBoost. These algorithms have been widely used in EEG-based mental health and biomedical classification studies [64]. To improve predictive stability and reduce model-specific variability, static ensemble learning strategies were also evaluated, including Voting and Stacking classifiers [65]. For the Voting ensemble, a soft voting strategy was used to aggregate probability-level predictions from the optimized base classifiers. In the Stacking ensemble, optimized base classifiers were first trained independently, and their outputs were then used to train a meta-learner. This allowed the ensemble to combine complementary decision patterns from multiple classifiers and capture more complex relationships in the EEG feature space. In addition to static ensembles, Dynamic Ensemble Selection (DES) methods were incorporated to account for local variability in EEG patterns across patients and epochs. The evaluated DES variants included FIRE-DESKNN, FIRE-KNOP, FIRE-KNORAE, and FIRE-KNORAU. These methods dynamically select the most competent classifier or subset of classifiers for each test instance based on the local decision region. The DES pool was constructed using the optimized baseline classifiers, and Dynamic Frienemy Pruning (DFP) was enabled to improve local classifier selection. Bayesian optimization was used for systematic hyperparameter tuning [38,66]. This probabilistic search strategy enabled efficient exploration of the hyperparameter space for each classifier. For ensemble models, the base classifiers were individually optimized before being integrated into the Voting or Stacking framework. Similarly, for DES models, parameters controlling the dynamic selection process were tuned to improve adaptive and reliable classification performance. This multi-model training strategy enabled a comprehensive comparison between individual classifiers, static ensembles, and dynamically adaptive ensemble methods. The final model performance metrics and comparative results are reported in Section 5.

3.6. Explainable artificial intelligence

The integration of Explainable Artificial Intelligence (XAI) into EEG-based mental health analysis plays a crucial role in improving model transparency, particularly for clinical interpretation and potential decision-support applications [21,38,67]. Since EEG signals are spatially distributed across the scalp and reflect complex neurophysiological activity, it is important to identify which features and channels contribute most strongly to the model predictions. To address this, we integrated three complementary XAI methods into the proposed framework: SHapley Additive Explanations (SHAP) [68], Local Interpretable Model-Agnostic Explanations (LIME) [69], and NeuroXAI [70]. SHAP and LIME were used to explain predictions in the engineered feature space, whereas NeuroXAI was used to provide channel-wise interpretations closer to the spatial structure of EEG signals. To ensure an unbiased explanation analysis, XAI samples were selected only from the held-out test subsets of the patient-level cross-validation folds. No training samples were used for explanation generation. Within each fold, a balanced subset of correctly classified samples from the anxiety and stress classes was selected from the test set so that both diagnostic classes were represented. This strategy ensured that the explanations reflected model behavior on unseen patients rather than memorized training patterns. SHAP and LIME were applied to the extracted EEG feature vectors, whereas NeuroXAI was used to generate channel-level explanations. For patient-level interpretation, epoch-level explanations were aggregated across retained epochs to obtain stable subject-level feature and channel importance profiles.

3.6.1. SHAP

SHAP builds on principles from cooperative game theory to allocate feature-level contribution scores to a model prediction. Consider a model $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ and a specific input $\mathbf{x}_0$. The SHAP value ϕ_j for feature j captures its average contribution across all possible combinations of features:

$$\phi_j(\mathbf{x}_0, \mathcal{F}) = \sum_{S \subseteq D \setminus \{j\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[\mathcal{F}_{S\cup\{j\}}(\mathbf{x}_0) - \mathcal{F}_S(\mathbf{x}_0)\right] \tag{1}$$

where $D = \{1, 2, \ldots, d\}$ represents the full set of features, and $\mathcal{F}_S(\mathbf{x}_0)$ is the expected output of the model when only the subset S is considered. SHAP satisfies the additive completeness property, meaning that the model output can be decomposed as:

$$\mathcal{F}(\mathbf{x}_0) = \phi_0 + \sum_{j=1}^{d} \phi_j \tag{2}$$

where ϕ_0 denotes the baseline prediction, corresponding to the expected model output over the background samples. In this study, SHAP values were computed for the extracted EEG features and then aggregated according to EEG channel and feature type. This allowed us to identify which channel-specific feature contributions had the greatest influence on the final classification, including whether their effects supported the anxiety or stress prediction.

3.6.2. LIME

LIME provides local explanations by approximating the behavior of a complex model around a specific input using an interpretable surrogate model. For a given input $\mathbf{x}_0$, LIME generates perturbed samples $\mathbf{x}_i$ and assigns each perturbation a proximity weight based on its similarity to the original input. A local surrogate model $\mathcal{G}$ is then trained to approximate the predictive behavior of the black-box model $\mathcal{F}$:

$$\mathcal{E}(\mathbf{x}_0) = \arg\min_{\mathcal{G}\in\mathcal{H}} \sum_{i=1}^{N} w(\mathbf{x}_i) \cdot \left(\mathcal{F}(\mathbf{x}_i) - \mathcal{G}(\mathbf{m}_i)\right)^2 + R(\mathcal{G}) \tag{3}$$

where $\mathbf{m}_i \in \{0, 1\}^d$ is a binary interpretable mask, $w(\mathbf{x}_i)$ is the proximity weight, and $R(\mathcal{G})$ is a regularization term that controls the complexity of the surrogate model. The proximity weight is commonly defined as:

$$w(\mathbf{x}_i) = \exp\left(-\frac{\delta(\mathbf{x}_0, \mathbf{x}_i)^2}{\sigma^2}\right) \tag{4}$$

The surrogate model can be expressed as:

$$\mathcal{G}(\mathbf{m}_i) = \theta_0 + \sum_{j=1}^{d} \theta_j \cdot \mathbf{m}_i^j \tag{5}$$

where θ_j represents the local importance coefficient of feature j. In this study, LIME was applied to the engineered EEG feature vectors. The resulting local feature importance scores were aggregated to provide region-wise and feature-wise relevance patterns for anxiety and stress classification.

3.6.3. NeuroXAI

To account for the structured spatial nature of EEG data, we implemented NeuroXAI, a method designed for spatially grounded attribution in neural signals. Unlike conventional feature-based explanation methods, NeuroXAI focuses on channel-wise relevance by perturbing EEG inputs through channel masking. Given an input $\mathbf{x}_0$, NeuroXAI generates perturbed samples using binary masks $\mathbf{m}_i \in \{0,1\}^d$, where each mask represents the presence or absence of specific EEG channels. A surrogate model $\mathcal{G}$ is then trained to approximate the black-box model $\mathcal{F}$:

$$\mathcal{E}(\mathbf{x}_0) = \arg\min_{\mathcal{G}\in\mathcal{H}} \sum_{i=1}^{N} w(\mathbf{x}_i) \cdot \left(\mathcal{F}(\mathbf{x}_i) - \mathcal{G}(\mathbf{m}_i)\right)^2 + R(\mathcal{G}) \tag{6}$$

The proximity weight is determined based on spatial similarity between the original and perturbed samples:

$$w(\mathbf{x}_i) = \exp\left(-\frac{\delta(\mathbf{x}_0, \mathbf{x}_i)^2}{\sigma^2}\right) \tag{7}$$

with cosine distance defined as:

$$\delta(\mathbf{x}_0, \mathbf{x}_i) = 1 - \frac{\phi(\mathbf{x}_0) \cdot \phi(\mathbf{x}_i)}{\|\phi(\mathbf{x}_0)\| \cdot \|\phi(\mathbf{x}_i)\|} \tag{8}$$

where $\phi(\cdot)$ maps the EEG input into a vectorized representation. Unlike LIME, NeuroXAI employs a non-linear feature selection mechanism using the Hilbert-Schmidt Independence Criterion Lasso (HSIC-Lasso). Given an output kernel matrix $\mathbf{Y}$ and individual channel kernels $\mathbf{K}_k$, the channel importance weights β_k are computed by solving:

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{2} \left\| \mathbf{Y} - \sum_{k=1}^{d} \beta_k \mathbf{K}_k \right\|_{\mathrm{F}}^2 + \lambda \|\beta\|_1 \quad \text{s.t. } \beta_k \geq 0 \tag{9}$$

In this study, NeuroXAI was applied at both the instance and patient levels to generate channel importance profiles. Epoch-level channel attributions were aggregated across retained epochs to obtain stable patient-level explanations. These outputs were visualized using bar charts and topographic maps to reveal the spatial contributions of EEG regions to anxiety and stress classification.

4. Experimental setup

This section describes the experimental protocols and implementation details used to train, evaluate, and interpret the proposed EEG-based anxiety and stress classification framework. It includes EEG segment-duration analysis, feature extraction strategies, channel-configuration analysis, filtering settings, patient-level cross-validation, model training and optimization, external validation, performance evaluation, and explainability analysis using SHAP, LIME, and NeuroXAI. These components were designed to ensure a consistent, reproducible, and leakage-free evaluation across baseline machine learning classifiers, static ensemble models, and dynamic ensemble selection methods. The overall experimental flow is summarized in Fig. 6.

4.1. Patient-level stratified cross-validation

All experiments were conducted using a stratified 10-fold cross-validation strategy applied at the patient level. Patients were partitioned into ten non-overlapping folds while preserving the proportion of anxiety and stress cases across folds. In each cross-validation iteration, one fold was used as the test subset, one fold was used as the validation subset, and the remaining eight folds were used for model training. The validation subset was used for hyperparameter optimization and model selection, whereas the test subset was kept unseen during training and optimization and was used only for final performance evaluation.

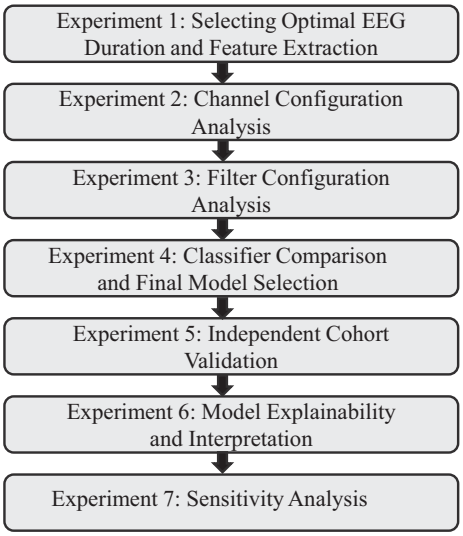


Fig. 6. Experimental workflow of the proposed study.

The same patient-level partitioning strategy was applied across all experiments to ensure consistency and prevent patient-level information leakage, such that epochs from the same patient never appeared in more than one subset within a given fold.

4.2. Training and hyperparameter optimization

All baseline classifiers, static ensemble models, and dynamic ensemble selection methods were trained for binary anxiety versus stress classification using the extracted EEG feature vectors. The input features were obtained from the selected two-hour EEG segments using non-overlapping 30-second epochs and the final six-channel configuration consisting of F3, F4, C3, C4, O1, and O2. The final selected feature configuration consisted of Approximate Entropy and Standard Deviation, as described in Section 3.3. Within each patient-level cross-validation fold, feature normalization was performed by fitting the Standard Scaler only on the training partition and then applying the same transformation to the validation and test partitions. This prevented information leakage from unseen patients during preprocessing. Bayesian optimization was used to tune the main hyperparameters of the evaluated classifiers and ensemble models. The search space included model-specific parameters such as the number of estimators, maximum tree depth, learning rate, regularization strength, number of neighbors, kernel parameters, and hidden-layer configuration, depending on the classifier. The validation fold was used to guide hyperparameter selection and model selection, whereas the held-out test fold was used only for final performance evaluation. For static ensemble models, the base classifiers were optimized independently before being integrated into the Voting and Stacking ensembles. In the Stacking ensemble, optimized base classifiers generated intermediate predictions, which were passed to a logistic regression meta-classifier for the final decision. For dynamic ensemble selection methods, the optimized classifier pool was used, and selection-related parameters such as neighborhood size and pruning configuration were tuned using the validation fold. The same training, validation, and testing protocol was maintained across all classifiers and ensemble models to ensure a fair and consistent comparison. The final optimized configurations are summarized in Table 4. Since optimization was performed independently within each cross-validation fold, the reported values represent the most frequently selected configurations across folds, whereas continuous parameters that varied across folds are reported using their median values.

Table 4
Final optimal hyper-parameters selected by Bayesian optimization under stratified 10-fold cross-validation for the anxiety vs. stress classification.

Model	Optimal Hyper-parameters
K-Nearest Neighbors	n_neighbors: 5; weights: distance; metric: euclidean; leaf_size: 30
Support Vector Machine	kernel: rbf; C: 10.0; gamma: scale; tol: 10^{-3}
Logistic Regression	C: 1.0; penalty: l2; solver: lbfgs; max_iter: 1000
Decision Tree	criterion: gini; max_depth: 5; min_samples_split: 5; min_samples_leaf: 2
Naive Bayes	var_smoothing: 10^{-9}
Random Forest	n_estimators: 200; criterion: gini; max_depth: 10; min_samples_split: 2; min_samples_leaf: 1; max_features: sqrt
XGBoost	n_estimators: 200; max_depth: 3; learning_rate: 0.05; subsample: 0.8; colsample_bytree: 0.8; min_child_weight: 1; gamma: 0; reg_alpha: 0.1; reg_lambda: 1.0
Gradient Boosting	n_estimators: 200; learning_rate: 0.1; max_depth: 3; subsample: 0.8; min_samples_split: 2; min_samples_leaf: 1
AdaBoost	n_estimators: 100; learning_rate: 0.1; estimator: DecisionTreeClassifier (max_depth = 1)
Multi-Layer Perceptron	hidden_layer_sizes: (100); activation: relu; solver: adam; alpha: 0.001; learning_rate_init: 0.001; learning_rate: constant; max_iter: 1000
Voting Ensemble	voting: soft; estimators: optimized KNN, Random Forest, XGBoost, SVM
Stacking Ensemble	estimators: optimized KNN, Random Forest; final_estimator: LogisticRegression (C = 1.0); cv: 5
FIRE-DESKNN	pool_classifiers: optimized baseline classifiers; k: 5; safe_k: 5; DFP: True; method: DESKNN
FIRE-KNOP	pool_classifiers: optimized baseline classifiers; k: 5; safe_k: 5; DFP: True; method: KNOP
FIRE-KNORAE	pool_classifiers: optimized baseline classifiers; k: 5; safe_k: 5; DFP: True; method: KNORAE
FIRE-KNORAU	pool_classifiers: optimized baseline classifiers; k: 5; safe_k: 5; DFP: True; method: KNORAU

Table 5
Evaluation metrics used for anxiety and stress classification.

Metric	Description	Formula
Accuracy	Measures the overall proportion of correctly classified samples among all evaluated samples.	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	Measures the proportion of samples predicted as positive that are actually positive, reflecting the reliability of positive predictions.	$\text{Precision} = \frac{TP}{TP + FP}$
Recall	Measures the proportion of actual positive samples that are correctly identified by the model.	$\text{Recall} = \frac{TP}{TP + FN}$
F1-score	Computes the harmonic mean of precision and recall, providing a balanced measure when both false positives and false negatives are important.	$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
AUC	Represents the area under the receiver operating characteristic curve and measures the model's ability to distinguish between anxiety and stress across different classification thresholds.	$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$

4.3. Evaluation metrics

Model performance was evaluated using standard binary classification metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). These metrics were computed on the held-out test fold in each patient-level cross-validation iteration. Since each EEG recording was divided into multiple non-overlapping 30-second epochs, epoch-level predictions were aggregated at the patient level to obtain the final patient-wise prediction. Specifically, patient-level predictions were obtained by aggregating the predicted probabilities across all retained epochs from the same patient, followed by the assignment of the final class label. This ensured that the reported performance reflected patient-level classification rather than isolated epoch-level decisions. For each experiment, results are reported as the mean and standard deviation across the ten cross-validation folds. In addition, 95% confidence intervals were computed to estimate performance variability. The same evaluation protocol was applied uniformly across baseline machine learning classifiers, static ensemble models, dynamic ensemble selection methods, and external validation experiments. The definitions and mathematical formulations of the evaluation metrics used in this study are summarized in Table 5.

4.4. Explainability evaluation

To interpret the predictions of the best-performing model, explainability analysis was conducted using SHAP, LIME, and NeuroXAI. Explanation samples were selected only from the held-out test subsets of the patient-level cross-validation folds. No training samples were used for explanation generation. Within each fold, a balanced subset of correctly classified samples from the anxiety and stress classes was selected from the test set to ensure that both diagnostic classes were represented. SHAP and LIME were applied to the extracted EEG feature vectors, whereas NeuroXAI was used to generate channel-level explanations. Epoch-level explanations were further aggregated at the patient level to obtain stable patient-level feature and channel importance profiles.

5. Results and discussion

5.1. Experiment 1: selecting the optimal EEG duration and feature extraction method

Experiment 1 was conducted to identify the optimal EEG signal duration and the most effective feature extraction strategy for anxiety and stress classification. Random Forest (RF) was used as the baseline classifier because it is robust, handles high-dimensional feature spaces well, and is commonly used in EEG-based classification studies. As described in Section 3.1, the average EEG recording duration per patient was approximately 10 hours. To reduce computational complexity while preserving clinically meaningful sleep EEG information, the first hour of each recording was excluded because it often contained limited sleep-stage annotations or was dominated by wakefulness. After excluding the initial hour, four EEG durations were evaluated:

one-hour, two-hour, three-hour, and four-hour. Several feature extraction strategies were evaluated across these EEG durations, including frequency-domain, time-domain, and non-linear features. The tested features included Power Spectral Density (PSD), Kurtosis, Mean, Standard Deviation (Std), Approximate Entropy (ApEn), and different combinations of these features. For ApEn, two embedding dimensions, $m = 2$ and $m = 3$, were tested using the Chebyshev distance metric. The results of all feature-duration settings are summarized in Table 6. Table 6 shows that simple individual features such as PSD, Mean, and Kurtosis produced relatively weak classification performance, with accuracy near chance level, around 48–52%. This indicates that these features alone were not sufficient to capture anxiety and stress-related EEG patterns. In contrast, Std-based and ApEn-based settings achieved stronger performance, suggesting that amplitude variability and non-linear signal irregularity were more informative for this task. Notably, the embedding dimension $m = 2$ consistently performed better than $m = 3$ across ApEn-based feature combinations. The final selected setting was the 2-hour EEG segment using ApEn and Std with embedding dimension $m = 2$ and Chebyshev distance. This setting achieved an accuracy of 82.53 ± 1.50% [81.46–83.60], precision of 81.52 ± 1.66% [80.33–82.71], recall of 83.28 ± 1.48% [82.22–84.34], F1-score of 82.97 ± 1.59% [81.83–84.11], and AUC of 84.59 ± 1.42% [83.57–85.61]. Although some larger feature combinations achieved slightly higher values for individual metrics, the ApEn and Std setting over the 2-hour segment provided the best overall balance across all five evaluation metrics while remaining compact and physiologically interpretable. This trend is also shown in Fig. 7, which compares the AUC performance of the ApEn and Std ($m = 2$) setting across the four evaluated durations. The AUC increased from 82.80% at 1h to the highest value of 84.59% at 2 hours, then decreased to 80.20% at 3 hours and 78.50% at 4 hours. The error bars, based on the standard deviation across 10-fold cross-validation, show that the two-hour setting was consistently strong across patient-level folds. The larger variance at longer durations, especially at 4 hours ($\sigma = 1.99$ compared with $\sigma = 1.42$ at 2 hours), suggests that longer windows introduced more signal heterogeneity and noise, reducing model stability. The duration comparison further showed that performance improved from 1h to 2 hours for the strongest feature configurations. For example, the selected ApEn and Std ($m = 2$) setting improved from 80.20% accuracy at 1h to 82.53% at 2 hours, indicating that the 2-hour window captured more stable and informative sleep EEG patterns. However, extending the duration to 3 or 4 hours reduced performance, with accuracy dropping to 77.80% and 75.40%, respectively. This decline may be related to greater signal variability, additional noisy epochs, and more heterogeneous sleep-stage composition over longer recordings. Therefore, the 2-hour EEG segment with ApEn and Std was selected as the final feature-duration setting for the subsequent experiments.

5.2. Experiment 2: channel configuration analysis

This experiment evaluates the contribution of individual EEG channels and their combinations for anxiety and stress classification. The analysis used the optimal two-hour EEG segment and the ApEn and Std feature configuration identified in Experiment 1, Section 5.1. To ensure a fair comparison across channel settings, the same Random Forest classifier and patient-level stratified 10-fold cross-validation protocol were applied to all configurations. The evaluated channels covered three scalp regions: frontal (F3, F4), central (C3, C4), and occipital (O1, O2). Each channel was first tested individually and then progressively combined to examine whether spatially distributed EEG information improves classification performance. Table 7 summarizes the results. Among the single-channel settings, performance remained close to chance level for all channels, indicating that anxiety and stress related EEG patterns are not confined to one scalp region. C3 achieved the strongest individual performance, with an accuracy of 59.39 ± 3.61% [57.15–61.63] and an AUC of 63.50 ± 3.11% [61.57–65.43], followed by F4, with an accuracy of 57.09 ± 2.16% and an AUC of 61.03 ± 2.16%. F3 showed the weakest discriminative ability, with an accuracy of 54.59 ± 3.19% and an F1-score of 53.01 ± 3.76%. These results suggest that frontal, central, or occipital activity alone is not sufficient to capture the full EEG signature of pediatric anxiety and stress during sleep. Combining channels led to clear performance improvements. The C3–C4 pair produced the first major gain, reaching an accuracy of 73.55 ± 3.50% [71.38–75.72] and an AUC of 78.54 ± 3.44% [76.41–80.67]. This indicates that bilateral central information provides important discriminative value. Adding O2 further improved performance, with an accuracy of 78.50 ± 2.15% [77.17–79.83] and an AUC of 81.50 ± 1.84% [80.36–82.64], suggesting that occipital activity contributes complementary information beyond the central channels. The four-channel combination (C3, C4, O2, F4) improved precision to 81.50 ± 1.96% [80.29–82.71] and recall to 80.29 ± 2.14% [78.96–81.62], with an AUC of 82.24 ± 2.21% [80.87–83.61]. Adding O1 to form the five-channel set slightly reduced accuracy to 79.68 ± 3.26%, but increased recall to 83.47 ± 1.13% [82.77–84.17] and AUC to 83.98 ± 2.88% [82.19–85.77]. This suggests that additional occipital information improved sensitivity, although with a small trade-off in overall accuracy. The best overall performance was obtained using all six channels (C3, C4, O2, F4, O1, F3). This configuration achieved an accuracy of 81.46 ± 1.40% [80.59–82.33], precision of 83.64 ± 1.32% [82.82–84.46], recall of 81.10 ± 1.30% [80.29–81.91], F1-score of 79.68 ± 2.80% [77.94–81.42], and AUC of 84.77 ± 1.37% [83.92–85.62]. The steady AUC increase across the channel combinations (78.54 → 81.50 → 82.24 → 83.98 → 84.77) shows that frontal, central, and occipital channels provide complementary information rather than redundant signals. The largest improvement occurred when moving from single-channel to paired-channel analysis, followed by smaller gains as additional channels were added. These findings show that pediatric anxiety and stress classification during sleep benefits from multi-channel EEG coverage. Central channels provided the strongest initial contribution, but the full six-channel montage produced the most reliable representation by combining information from frontal, central, and occipital regions.

5.3. Experiment 3: filter configuration analysis

This experiment identified the most suitable frequency filtering strategy for the selected two-hour EEG segments, ApEn and Std features, and six-channel montage (C3, C4, O2, F4, O1, F3) determined in Experiments 1 and 2. For consistency, the same patient-level stratified 10-fold cross-validation protocol with Random Forest was used. Several bandpass ranges were evaluated, including 1–40 Hz, 1–45 Hz, 1–50 Hz, 1–60 Hz, 1–70 Hz, and 1–80 Hz. The aim was to identify the frequency range that best preserves anxiety and stress related EEG information while reducing noise. A standalone 60 Hz notch filter was not tested because notch filtering alone cannot remove broadband artifacts such as DC drift, muscle activity, or electrode movement. Its role is mainly to reduce line noise after selecting an appropriate bandpass range. Table 8 presents the results. The 1–40 Hz and 1–45 Hz settings showed comparable but lower performance, with AUC values of 81.56% and 80.96%, respectively. This suggests that frequency components above 40 Hz may contain useful information for anxiety and stress classification. Performance improved at 1–50 Hz, reaching an AUC of 84.20 ± 1.90% [82.84–85.56], and reached its best bandpass-only result at 1–60 Hz. This setting achieved an accuracy of 82.50 ± 1.13% [81.69–83.31], precision of 82.30 ± 1.42% [81.28–83.32], recall of 82.00 ± 1.56% [80.88–83.12], F1-score of 81.50 ± 2.40% [79.78–83.22], and AUC of 84.50 ± 1.12% [83.70–85.30]. The 1–60 Hz range covers the main clinically relevant EEG bands from delta to gamma, while limiting the contribution of high frequency muscle artifacts. Extending the upper cutoff to 70 Hz and 80 Hz reduced performance, with AUC values of 83.60% and 83.20%, respectively. This decrease suggests that wider frequency ranges may introduce additional high-frequency noise, especially muscle-related artifacts, which can affect the stability of non-linear features such as ApEn. The overall AUC pattern (81.56 → 80.96 → 84.20 → 84.50 →

Table 6
Performance of different feature extraction methods across EEG signal durations.

									ApEn(emb = '2', metric = 'chebyshev')					ApEn(emb = '3', metric = 'chebyshev')				
		PSD	Kurtosis	Mean	Std	Mean Std	Kurtosis Std	Kurtosis Mean Std	ApEn	ApEn Mean	ApEn Std	ApEn Mean Std	ApEn Mean Std Kurtosis	ApEn	ApEn Mean	ApEn Std	ApEn Mean Std	ApEn Mean Std Kurtosis
1 hour	mAcc±std	48.50 ±3.83	49.20 ±3.76	48.80 ±3.72	74.20 ±2.22	73.80 ±2.34	75.10 ±2.40	74.80 ±2.34	72.40 ±2.52	75.80 ±2.31	80.20 ±1.94	82.40 ±2.04	81.80 ±2.02	70.20 ±2.39	73.40 ±2.44	78.50 ±1.98	80.60 ±1.96	80.10 ±2.00
	[CI](%)	[46.13-50.87]	[46.87-51.53]	[46.49-51.11]	[72.82-75.58]	[72.35-75.25]	[73.61-76.59]	[73.35-76.25]	[70.84-73.96]	[74.37-77.23]	[79.00-81.40]	[81.14-83.66]	[80.55-83.05]	[68.72-71.68]	[71.89-74.91]	[77.27-79.73]	[79.39-81.81]	[78.86-81.34]
	mPre±std	47.80 ±4.50	46.50 ±4.31	48.20 ±4.50	74.80 ±2.60	71.20 ±2.55	75.40 ±2.57	71.20 ±2.56	71.80 ±2.73	70.40 ±2.79	79.50 ±2.12	77.20 ±2.26	76.10 ±2.06	69.80 ±2.72	68.80 ±2.68	77.80 ±2.09	76.20 ±2.26	74.80 ±2.22
	[CI](%)	[45.01-50.59]	[43.83-49.17]	[45.41-50.99]	[73.19-76.41]	[69.62-72.78]	[73.81-76.99]	[69.61-72.79]	[70.11-73.49]	[68.67-72.13]	[78.19-80.81]	[75.80-78.60]	[74.82-77.38]	[68.11-71.49]	[67.14-70.46]	[76.50-79.10]	[74.80-77.60]	[73.42-76.18]
	mRe±std	49.20 ±4.13	51.80 ±4.17	49.50 ±4.16	72.50 ±2.43	72.40 ±2.36	74.20 ±2.43	76.80 ±2.30	73.20 ±2.43	73.20 ±2.41	81.20 ±2.00	80.30 ±2.12	81.50 ±1.90	71.50 ±2.65	71.50 ±2.64	79.50 ±1.84	79.10 ±1.86	80.20 ±1.94
	[CI](%)	[46.64-51.76]	[49.22-54.38]	[46.92-52.08]	[70.99-74.01]	[70.94-73.86]	[72.69-75.71]	[75.37-78.23]	[71.69-74.71]	[71.71-74.69]	[79.96-82.44]	[78.99-81.61]	[80.32-82.68]	[69.86-73.14]	[69.86-73.14]	[78.36-80.64]	[77.95-80.25]	[79.00-81.40]
	mF1±std	48.40 ±4.20	48.90 ±4.20	48.60 ±4.09	73.60 ±2.47	71.70 ±2.45	74.80 ±2.52	73.80 ±2.63	72.40 ±2.67	71.80 ±2.74	80.30 ±2.06	78.70 ±2.17	78.50 ±1.91	70.60 ±2.58	70.00 ±2.49	78.60 ±2.14	77.50 ±2.00	77.30 ±1.92
	[CI](%)	[45.80-51.00]	[46.30-51.50]	[46.06-51.14]	[72.07-75.13]	[70.18-73.22]	[73.24-76.36]	[72.17-75.43]	[70.75-74.05]	[70.10-73.50]	[79.02-81.58]	[77.36-80.04]	[77.32-79.68]	[69.00-72.20]	[68.46-71.54]	[77.27-79.93]	[76.26-78.74]	[76.11-78.49]
	mAUC±std	50.80 ±3.18	51.40 ±3.35	51.20 ±3.19	76.80 ±1.90	77.20 ±1.95	79.20 ±1.83	78.20 ±2.10	76.20 ±2.04	78.20 ±1.99	82.80 ±1.52	85.20 ±1.64	86.20 ±1.56	74.20 ±2.00	76.20 ±2.21	81.20 ±1.67	83.80 ±1.50	84.80 ±1.55
	[CI](%)	[48.83-52.77]	[49.32-53.48]	[49.22-53.18]	[75.62-77.98]	[75.99-78.41]	[78.07-80.33]	[76.90-79.50]	[74.94-77.46]	[76.97-79.43]	[81.86-83.74]	[84.18-86.22]	[85.23-87.17]	[72.96-75.44]	[74.83-77.57]	[80.16-82.24]	[82.87-84.73]	[83.84-85.76]
2 hours	mAcc±std	50.20 ±3.63	51.80 ±3.57	50.50 ±3.68	71.50 ±2.06	74.20 ±2.12	72.80 ±2.14	76.20 ±2.31	74.80 ±2.40	76.40 ±2.29	**82.53 ±1.50**	81.20 ±1.75	80.60 ±1.94	72.50 ±2.17	74.80 ±2.43	80.20 ±1.73	79.80 ±1.77	79.20 ±1.91
	[CI](%)	[47.95-52.45]	[49.59-54.01]	[48.22-52.78]	[70.22-72.78]	[72.89-75.51]	[71.47-74.13]	[74.77-77.63]	[73.31-76.29]	[74.98-77.82]	**[81.46-83.60]**	[80.12-82.28]	[79.40-81.80]	[71.16-73.84]	[73.29-76.31]	[79.13-81.27]	[78.70-80.90]	[78.02-80.38]
	mPre±std	49.50 ±4.30	48.20 ±4.30	49.80 ±4.24	72.60 ±2.57	73.80 ±2.50	72.50 ±2.43	76.40 ±2.62	73.60 ±2.78	76.20 ±2.75	**81.52 ±1.66**	80.80 ±2.08	80.40 ±2.00	71.40 ±2.69	74.50 ±2.58	79.40 ±2.13	79.20 ±2.12	79.00 ±2.02
	[CI](%)	[46.83-52.17]	[45.53-50.87]	[47.17-52.43]	[71.01-74.19]	[72.25-75.35]	[70.99-74.01]	[74.78-78.02]	[71.88-75.32]	[74.50-77.90]	**[80.33-82.71]**	[79.51-82.09]	[79.16-81.64]	[69.73-73.07]	[72.90-76.10]	[78.08-80.72]	[77.89-80.51]	[77.75-80.25]
	mRe±std	51.00 ±3.91	53.50 ±3.94	51.20 ±3.84	70.80 ±2.25	74.50 ±2.39	74.80 ±2.44	75.30 ±2.22	75.20 ±2.37	77.30 ±2.53	**83.28 ±1.48**	81.20 ±1.96	81.00 ±1.89	73.20 ±2.43	75.60 ±2.32	81.20 ±1.97	80.50 ±1.94	80.30 ±1.80
	[CI](%)	[48.58-53.42]	[51.06-55.94]	[48.82-53.58]	[69.41-72.19]	[73.02-75.98]	[73.29-76.31]	[73.92-76.68]	[73.73-76.67]	[75.73-78.87]	**[82.22-84.34]**	[79.99-82.41]	[79.83-82.17]	[71.69-74.71]	[74.16-77.04]	[79.98-82.42]	[79.30-81.70]	[79.18-81.42]
	mF1±std	50.10 ±4.09	50.60 ±3.99	50.30 ±3.83	71.60 ±2.38	74.10 ±2.29	73.60 ±2.47	75.80 ±2.49	74.30 ±2.38	76.60 ±2.35	**82.97 ±1.59**	80.90 ±1.91	80.60 ±1.89	72.20 ±2.36	74.90 ±2.49	80.20 ±1.87	79.80 ±1.96	79.50 ±1.81
	[CI](%)	[47.56-52.64]	[48.13-53.07]	[47.93-52.67]	[70.12-73.08]	[72.68-75.52]	[72.07-75.13]	[74.26-77.34]	[72.82-75.78]	[75.14-78.06]	**[81.83-84.11]**	[79.72-82.08]	[79.43-81.77]	[70.74-73.66]	[73.36-76.44]	[79.04-81.36]	[78.59-81.01]	[78.38-80.62]
	mAUC±std	52.40 ±3.12	54.20 ±3.21	52.80 ±3.01	74.20 ±1.99	78.10 ±1.74	74.80 ±1.91	80.10 ±1.88	77.50 ±1.88	79.50 ±1.97	**84.59 ±1.42**	84.10 ±1.68	84.20 ±1.47	75.80 ±1.90	77.80 ±2.11	82.80 ±1.60	83.20 ±1.41	84.60 ±1.56
	[CI](%)	[50.47-54.33]	[52.21-56.19]	[50.93-54.67]	[72.97-75.43]	[77.02-79.18]	[73.62-75.98]	[78.93-81.27]	[76.33-78.67]	[78.28-80.72]	**[83.57-85.61]**	[83.06-85.14]	[83.29-85.11]	[74.62-76.98]	[76.49-79.11]	[81.81-83.79]	[82.33-84.07]	[83.63-85.57]
3 hours	mAcc±std	49.10 ±3.81	48.50 ±3.66	49.30 ±3.71	67.80 ±2.70	71.50 ±2.12	69.50 ±2.61	73.40 ±2.29	68.20 ±2.64	72.50 ±2.37	77.80 ±1.76	78.30 ±1.89	77.20 ±1.76	66.80 ±2.87	70.50 ±2.42	75.60 ±2.26	76.50 ±2.26	75.80 ±2.07
	[CI](%)	[46.74-51.46]	[46.23-50.77]	[47.00-51.60]	[66.13-69.47]	[70.19-72.81]	[67.88-71.12]	[71.98-74.82]	[66.56-69.84]	[71.03-73.97]	[76.71-78.89]	[77.13-79.47]	[76.11-78.29]	[65.02-68.58]	[69.00-72.00]	[74.20-77.00]	[75.10-77.90]	[74.52-77.08]
	mPre±std	48.30 ±4.17	45.80 ±4.37	48.10 ±4.20	67.50 ±3.14	70.80 ±2.70	69.80 ±3.12	73.10 ±2.46	67.50 ±3.08	72.30 ±2.77	77.20 ±2.06	77.80 ±2.14	76.80 ±2.26	66.20 ±3.20	70.20 ±2.64	75.10 ±2.59	76.00 ±2.43	75.20 ±2.42
	[CI](%)	[45.72-50.88]	[43.09-48.51]	[45.50-50.70]	[65.55-69.45]	[69.13-72.47]	[67.87-71.73]	[71.58-74.62]	[65.59-69.41]	[70.58-74.02]	[75.92-78.48]	[76.47-79.13]	[75.40-78.20]	[64.22-68.18]	[68.56-71.84]	[73.49-76.71]	[74.49-77.51]	[73.70-76.70]
	mRe±std	50.10 ±3.94	50.20 ±3.91	50.00 ±4.07	66.90 ±2.80	71.20 ±2.48	68.50 ±2.90	72.80 ±2.25	69.10 ±2.97	74.10 ±2.52	78.50 ±2.01	78.50 ±1.81	77.50 ±1.90	67.50 ±2.95	72.30 ±2.33	76.30 ±2.36	76.80 ±2.38	76.10 ±2.34
	[CI](%)	[47.66-52.54]	[47.78-52.62]	[47.48-52.52]	[65.16-68.64]	[69.66-72.74]	[66.70-70.30]	[71.41-74.19]	[67.26-70.94]	[72.54-75.66]	[77.25-79.75]	[77.38-79.62]	[76.32-78.68]	[65.67-69.33]	[70.86-73.74]	[74.84-77.76]	[75.32-78.28]	[74.65-77.55]
	mF1±std	49.15 ±3.97	47.80 ±4.18	48.90 ±4.17	67.20 ±2.87	70.90 ±2.46	69.10 ±2.98	72.90 ±2.54	68.20 ±2.86	73.10 ±2.46	77.80 ±2.03	78.10 ±1.95	77.10 ±2.11	66.80 ±3.01	71.20 ±2.59	75.70 ±2.26	76.30 ±2.44	75.60 ±2.25
	[CI](%)	[46.69-51.61]	[45.21-50.39]	[46.32-51.48]	[65.42-68.98]	[69.38-72.42]	[67.25-70.95]	[71.33-74.47]	[66.43-69.97]	[71.58-74.62]	[76.54-79.06]	[76.89-79.31]	[75.79-78.41]	[64.93-68.67]	[69.59-72.81]	[74.30-77.10]	[74.79-77.81]	[74.21-76.99]
	mAUC±std	51.30 ±3.23	50.80 ±3.08	51.50 ±3.12	70.50 ±2.18	75.30 ±1.86	71.20 ±2.34	76.50 ±1.94	71.80 ±2.27	76.20 ±2.07	80.20 ±1.67	81.50 ±1.49	81.80 ±1.60	70.50 ±2.28	74.20 ±1.91	78.50 ±1.67	80.10 ±1.92	81.20 ±1.71
	[CI](%)	[49.30-53.30]	[48.89-52.71]	[49.57-53.43]	[69.15-71.85]	[74.15-76.45]	[69.75-72.65]	[75.30-77.70]	[70.39-73.21]	[74.92-77.48]	[79.16-81.24]	[80.58-82.42]	[80.81-82.79]	[69.09-71.91]	[73.02-75.38]	[77.46-79.54]	[78.91-81.29]	[80.14-82.26]
4 hours	mAcc±std	47.80 ±3.99	47.30 ±4.01	47.90 ±4.05	66.40 ±2.90	70.10 ±2.38	70.30 ±2.49	71.50 ±2.24	65.50 ±2.81	70.20 ±2.50	75.40 ±2.12	76.50 ±2.40	75.80 ±2.24	64.20 ±3.01	68.20 ±2.96	73.20 ±2.33	74.20 ±2.21	74.50 ±2.12
	[CI](%)	[45.33-50.27]	[44.81-49.79]	[45.39-50.41]	[64.60-68.20]	[68.62-71.58]	[68.76-71.84]	[70.11-72.89]	[63.76-67.24]	[68.65-71.75]	[74.09-76.71]	[75.01-77.99]	[74.41-77.19]	[62.33-66.07]	[66.37-70.03]	[71.76-74.64]	[72.83-75.57]	[73.19-75.81]
	mPre±std	46.90 ±4.50	44.90 ±4.50	46.70 ±4.50	66.80 ±3.32	69.50 ±2.82	69.20 ±2.68	71.80 ±2.72	65.20 ±3.25	70.80 ±2.95	74.80 ±2.61	76.20 ±2.60	75.30 ±2.47	63.80 ±3.36	68.50 ±3.46	72.80 ±2.61	74.30 ±2.67	73.80 ±2.70
	[CI](%)	[44.11-49.69]	[42.11-47.69]	[43.91-49.49]	[64.74-68.86]	[67.75-71.25]	[67.54-70.86]	[70.11-73.49]	[63.19-67.21]	[68.97-72.63]	[73.18-76.42]	[74.59-77.81]	[73.77-76.83]	[61.72-65.88]	[66.36-70.64]	[71.18-74.42]	[72.65-75.95]	[72.13-75.47]
	mRe±std	48.50 ±4.23	49.10 ±4.07	48.30 ±4.23	65.20 ±3.06	69.80 ±2.54	68.80 ±2.46	70.50 ±2.51	66.80 ±3.19	71.50 ±2.53	76.10 ±2.38	76.80 ±2.32	76.20 ±2.23	65.10 ±3.00	69.80 ±3.14	74.10 ±2.40	74.50 ±2.22	74.80 ±2.43
	[CI](%)	[45.88-51.12]	[46.58-51.62]	[45.68-50.92]	[63.30-67.10]	[68.23-71.37]	[67.28-70.32]	[68.94-72.06]	[64.82-68.78]	[69.93-73.07]	[74.62-77.58]	[75.36-78.24]	[74.82-77.58]	[63.24-66.96]	[67.85-71.75]	[72.61-75.59]	[73.12-75.88]	[73.29-76.31]
	mF1±std	47.60 ±4.31	46.80 ±4.30	47.40 ±4.25	65.90 ±3.16	69.60 ±2.63	68.90 ±2.44	71.10 ±2.50	65.90 ±3.14	71.10 ±2.73	75.40 ±2.48	76.40 ±2.53	75.70 ±2.58	64.30 ±3.08	69.10 ±3.12	73.40 ±2.41	74.30 ±2.32	74.20 ±2.43
	[CI](%)	[44.93-50.27]	[44.13-49.47]	[44.77-50.03]	[63.94-67.86]	[67.97-71.23]	[67.39-70.41]	[69.55-72.65]	[63.95-67.85]	[69.41-72.79]	[73.86-76.94]	[74.83-77.97]	[74.10-77.30]	[62.39-66.21]	[67.17-71.03]	[71.91-74.89]	[72.86-75.74]	[72.69-75.71]
	mAUC±std	49.90 ±3.26	49.50 ±3.49	50.10 ±3.32	69.10 ±2.39	73.80 ±2.11	73.10 ±2.04	74.20 ±2.05	69.20 ±2.39	74.10 ±2.00	78.50 ±1.99	79.80 ±1.85	80.20 ±2.02	68.10 ±2.43	72.30 ±2.64	76.20 ±1.84	78.20 ±1.78	79.50 ±1.80
	[CI](%)	[47.88-51.92]	[47.34-51.66]	[48.04-52.16]	[67.62-70.58]	[72.49-75.11]	[71.84-74.36]	[72.93-75.47]	[67.72-70.68]	[72.86-75.34]	[77.27-79.73]	[78.65-80.95]	[78.95-81.45]	[66.59-69.61]	[70.66-73.94]	[75.06-77.34]	[77.10-79.30]	[78.38-80.62]

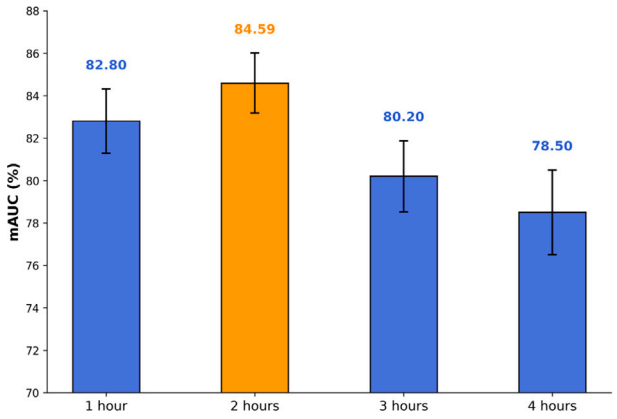


Fig. 7. mAUC performance of the ApEn + Std feature configuration (embedding dimension $m = 2$, Chebyshev distance) across EEG signal durations.

83.60 → 83.20) supports 1–60 Hz as the most effective bandpass range. The final preprocessing pipeline combined the 1–60 Hz bandpass filter with a 60 Hz notch filter to reduce power-line interference. This configuration achieved the best overall performance, with an accuracy of 84.50 ± 2.10% [83.00–86.00], precision of 85.20 ± 1.80% [83.91–86.49], recall of 84.80 ± 1.90% [83.44–86.16], F1-score of 83.10 ± 1.70% [81.88–84.32], and AUC of 86.80 ± 1.50% [85.73–87.87]. The notch filter improved AUC by 2.30 percentage points compared with 1–60 Hz bandpass filtering alone, indicating that 60 Hz line noise can affect ApEn-based signal irregularity estimates. Therefore, the 1–60 Hz bandpass filter combined with 60 Hz notch filtering was selected as the final preprocessing setting for the remaining experiments.

5.4. Experiment 4: classifier comparison and final model selection

This experiment compared sixteen machine learning configurations, including individual classifiers, ensemble methods, and FIRE-based hybrid selectors, to identify the best model for anxiety and stress classification. All preprocessing choices finalized in Experiments 1–3 were kept unchanged, including the two-hour EEG segments, ApEn and Std features, six channels (C3, C4, O2, F4, O1, F3), and 1–60 Hz bandpass filtering with a 60 Hz notch filter. Random Forest was used as the baseline model in Experiments 1–3 because of its stability and interpretability during feature, channel, and filter selection. In this experiment, we examined whether other classifiers or ensemble strategies could further improve performance using the finalized preprocessing pipeline. Bayesian-optimized hyperparameters (Table 4) and patient-level stratified 10-fold cross-validation were applied consistently across all models. Table 9 presents the results. Among the individual classifiers, K-Nearest Neighbors achieved the strongest performance, with an accuracy of 84.92 ± 1.10% and an AUC of 87.80 ± 1.05%. It was followed closely by Random Forest, with an AUC of 87.25 ± 1.15%, and FIRE-KNOP, with an AUC of 87.65 ± 1.10%. Overall, tree-based ensembles and FIRE hybrid selectors performed better than linear models such as Logistic Regression and SVM, as well as Naive Bayes. This suggests that the EEG patterns related to anxiety and stress are likely non-linear and involve interactions that are better captured by neighborhood-based or partition-based learning methods.

To examine whether the observed AUC differences were statistically reliable, pairwise comparisons were performed against the best-performing model using the Wilcoxon signed-rank test on fold-wise patient-level AUC values. All individual classifiers showed significant differences from the Stacking ensemble ($p < 0.001$). The Voting classifier, which was the closest competitor with an AUC of 88.50 ± 0.85%, also performed significantly lower than Stacking ($p = 0.011$). This indicates that even the 0.70-point AUC difference between Voting and Stacking was unlikely to be caused by random fold-to-fold variation.

The Stacking ensemble achieved the best overall performance, with an accuracy of 86.30 ± 0.75% [85.84–86.76], precision of 86.15 ± 0.78% [85.67–86.63], recall of 86.05 ± 0.80% [85.55–86.55], F1-score of 86.10 ± 0.77% [85.62–86.58], and AUC of 89.20 ± 0.65% [88.80–89.60]. The lower standard deviations observed for ensemble models, compared with individual classifiers, suggest that combining models improves prediction stability by reducing the effect of individual model bias and variance. The use of the non-parametric Wilcoxon test further supports this finding, as it does not require normally distributed fold-level AUC values and is suitable for the limited number of folds in 10-fold cross-validation. Based on these results, the Stacking ensemble was selected as the final

Table 7
Classification performance using individual EEG channels and their combinations.

Channel(s)	mAcc±std [CI](%)	mPre±std [CI](%)	mRe±std [CI](%)	mF1±std [CI](%)	mAUC±std [CI](%)
F3	54.59±3.19 [52.61-56.57]	51.32±4.19 [48.72-53.92]	55.45±2.84 [53.69-57.21]	53.01±3.76 [50.68-55.34]	56.67±2.87 [54.89-58.45]
F4	**57.09±2.16 [55.75-58.43]**	**55.28±2.56 [53.69-56.87]**	**56.94±2.74 [55.24-58.64]**	**55.79±3.87 [53.39-58.19]**	**61.03±2.16 [59.69-62.37]**
C3	**59.39±3.61 [57.15-61.63]**	**55.48±3.95 [53.03-57.93]**	**56.81±3.66 [54.54-59.08]**	**55.20±3.15 [53.25-57.15]**	**63.50±3.11 [61.57-65.43]**
C4	55.62±3.75 [53.30-57.94]	54.02±3.49 [51.86-56.18]	58.81±2.53 [57.24-60.38]	54.98±3.16 [53.02-56.94]	59.37±4.03 [56.87-61.87]
O1	55.79±2.44 [54.28-57.30]	52.86±2.52 [51.30-54.42]	58.49±2.68 [56.83-60.15]	53.33±3.34 [51.26-55.40]	58.07±2.13 [56.75-59.39]
O2	54.83±2.91 [53.03-56.63]	52.14±4.10 [49.60-54.68]	54.92±4.61 [52.06-57.78]	55.98±2.06 [54.70-57.26]	57.78±2.67 [56.13-59.43]
C3, C4	73.55±3.50 [71.38-75.72]	72.97±1.59 [71.98-73.96]	72.74±3.08 [70.83-74.65]	72.85±2.26 [71.45-74.25]	78.54±3.44 [76.41-80.67]
C3, C4, O2	78.50±2.15 [77.17-79.83]	76.20±2.34 [74.75-77.65]	76.80±2.11 [75.49-78.11]	75.40±2.27 [73.99-76.81]	81.50±1.84 [80.36-82.64]
C3, C4, O2, F4	80.24±2.37 [78.77-81.71]	81.50±1.96 [80.29-82.71]	80.29±2.14 [78.96-81.62]	78.01±2.56 [76.42-79.60]	82.24±2.21 [80.87-83.61]
C3, C4, O2, F4, O1	79.68±3.26 [77.66-81.70]	82.65±1.09 [81.97-83.33]	83.47±1.13 [82.77-84.17]	80.43±2.14 [79.10-81.76]	83.98±2.88 [82.19-85.77]
C3, C4, O2, F4, O1, F3	**81.46±1.40 [80.59-82.33]**	**83.64±1.32 [82.82-84.46]**	**81.10±1.30 [80.29-81.91]**	**79.68±2.80 [77.94-81.42]**	**84.77±1.37 [83.92-85.62]**

Table 8
Performance comparison using different filter frequencies.

Filter	Frequency (Hz)	mAcc±std [CI](%)	mPre±std [CI](%)	mRe±std [CI](%)	mF1±std [CI](%)	mAUC±std [CI](%)
Bandpass	1 - 40	79.48±2.05 [78.01-80.95]	78.80±2.85 [76.76-80.84]	80.34±1.63 [79.17-81.51]	77.81±2.51 [76.01-79.61]	81.56±1.84 [80.24-82.88]
	1 - 45	78.68±1.94 [77.29-80.07]	78.05±1.52 [76.96-79.14]	81.38±1.51 [80.30-82.46]	76.62±2.17 [75.07-78.17]	80.96±1.70 [79.74-82.18]
	1 - 50	82.23±2.21 [80.65-83.81]	80.33±2.46 [78.57-82.09]	79.65±3.16 [77.39-81.91]	77.71±2.65 [75.81-79.61]	84.20±1.90 [82.84-85.56]
	1 - 60	**82.50±1.13 [81.69-83.31]**	**82.30±1.42 [81.28-83.32]**	**82.00±1.56 [80.88-83.12]**	**81.50±2.40 [79.78-83.22]**	**84.50±1.12 [83.70-85.30]**
	1 - 70	81.50±1.25 [80.61-82.39]	80.80±1.96 [79.40-82.20]	81.20±2.26 [79.58-82.82]	80.50±1.05 [79.75-81.25]	83.60±1.15 [82.78-84.42]
	1 - 80	80.80±2.31 [79.15-82.45]	80.20±1.59 [79.06-81.34]	82.50±1.40 [81.50-83.50]	79.80±1.39 [78.81-80.79]	83.20±2.48 [81.43-84.97]
Bandpass + Notch	**1 - 60, 60**	**84.50±2.10 [83.00-86.00]**	**85.20±1.80 [83.91-86.49]**	**84.80±1.90 [83.44-86.16]**	**83.10±1.70 [81.88-84.32]**	**86.80±1.50 [85.73-87.87]**

Table 9
Model performance with Bayesian optimization under 10-fold patient-level stratified cross-validation for anxiety versus stress classification. Results are reported as mean ± standard deviation with 95% confidence intervals across cross-validation folds. The *p*-values compare each model against the Stacking ensemble using the Wilcoxon signed-rank test on fold-wise patient-level AUC values.

Model	mAcc±std [CI](%)	mPre±std [CI](%)	mRe±std [CI](%)	mF1±std [CI](%)	mAUC±std [CI](%)	*p*-value
K-Nearest Neighbors	84.92±1.10 [84.24-85.60]	84.75±1.12 [84.06-85.44]	84.95±1.15 [84.24-85.66]	84.85±1.13 [84.15-85.55]	87.80±1.05 [87.15-88.45]	<0.001
MultiLayer Perceptron	81.40±1.80 [80.28-82.52]	81.20±1.85 [80.05-82.35]	81.00±1.82 [79.87-82.13]	81.10±1.83 [79.97-82.23]	84.20±1.75 [83.12-85.28]	<0.001
Decision Tree	77.50±2.40 [76.01-78.99]	77.20±2.45 [75.68-78.72]	77.40±2.42 [75.90-78.90]	77.30±2.43 [75.79-78.81]	80.50±2.30 [79.07-81.93]	<0.001
Support Vector Machine	76.80±2.55 [75.22-78.38]	76.50±2.60 [74.89-78.11]	76.20±2.58 [74.60-77.80]	76.35±2.59 [74.74-77.96]	79.80±2.45 [78.28-81.32]	<0.001
Logistic Regression	68.20±2.50 [66.65-69.75]	67.80±2.55 [66.22-69.38]	68.50±2.52 [66.94-70.06]	68.10±2.53 [66.53-69.67]	71.50±2.40 [70.01-72.99]	<0.001
Naive Bayes	56.50±4.20 [53.90-59.10]	58.20±4.25 [55.57-60.83]	54.80±4.30 [52.13-57.47]	56.40±4.26 [53.76-59.04]	60.20±4.40 [57.47-62.93]	<0.001
Random Forest	84.50±1.20 [83.76-85.24]	84.30±1.22 [83.54-85.06]	84.40±1.25 [83.63-85.17]	84.35±1.23 [83.59-85.11]	87.25±1.15 [86.54-87.96]	<0.001
XGBoost	84.35±1.18 [83.62-85.08]	84.20±1.21 [83.45-84.95]	84.30±1.22 [83.54-85.06]	84.25±1.20 [83.51-84.99]	86.95±1.15 [86.24-87.66]	<0.001
Gradient Boosting	79.80±2.15 [78.47-81.13]	79.40±2.20 [78.04-80.76]	79.60±2.18 [78.25-80.95]	79.50±2.19 [78.14-80.86]	82.60±2.05 [81.33-83.87]	<0.001
AdaBoost	71.50±2.80 [69.76-73.24]	71.20±2.85 [69.43-72.97]	71.00±2.82 [69.25-72.75]	71.10±2.83 [69.35-72.85]	74.80±2.70 [73.13-76.47]	<0.001
Voting	85.90±0.95 [85.31-86.49]	85.65±0.98 [85.04-86.26]	85.80±1.00 [85.18-86.42]	85.72±0.97 [85.12-86.32]	88.50±0.85 [87.97-89.03]	0.011
Stacking	**86.30±0.75 [85.84-86.76]**	**86.15±0.78 [85.67-86.63]**	**86.05±0.80 [85.55-86.55]**	**86.10±0.77 [85.62-86.58]**	**89.20±0.65 [88.80-89.60]**	—
FIRE-DESKNN	84.25±1.25 [83.48-85.02]	84.40±1.28 [83.61-85.19]	83.90±1.30 [83.09-84.71]	84.15±1.27 [83.36-84.94]	86.70±1.26 [85.92-87.48]	<0.001
FIRE-KNOP	84.85±1.12 [84.16-85.54]	84.60±1.15 [83.89-85.31]	84.70±1.18 [83.97-85.43]	84.65±1.16 [83.93-85.37]	87.65±1.10 [86.97-88.33]	<0.001
FIRE-KNORAE	84.45±1.22 [83.69-85.21]	84.55±1.25 [83.78-85.32]	84.20±1.24 [83.43-84.97]	84.38±1.23 [83.62-85.14]	87.10±1.18 [86.37-87.83]	<0.001
FIRE-KNORAU	84.60±1.18 [83.87-85.33]	84.45±1.20 [83.71-85.19]	84.55±1.22 [83.79-85.31]	84.50±1.21 [83.75-85.25]	87.40±1.15 [86.69-88.11]	<0.001

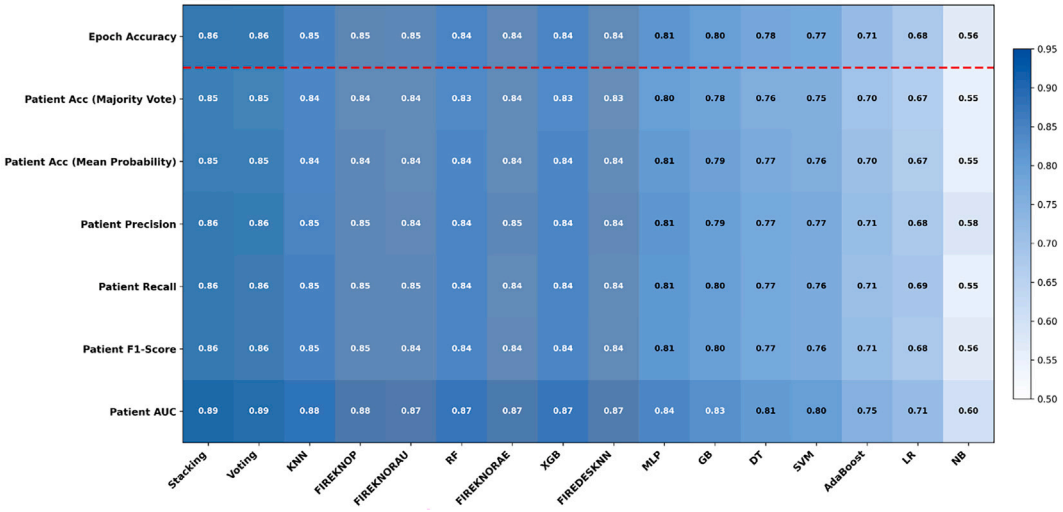


Fig. 8. Heatmap of model performance across baseline, static ensemble, and dynamic ensemble classifiers.

classification model for the proposed anxiety and stress detection framework. To complement the tabular results, Fig. 8 shows a heatmap of the sixteen models across seven evaluation metrics, with columns sorted by patient-level AUC. The color pattern shows a clear performance hierarchy. Stacking, Voting, KNN, FIRE-KNOP, FIRE-KNORAU, and Random Forest appear in the strongest performance region, while AdaBoost, Logistic Regression, and Naive Bayes show weaker discriminative ability. The Patient AUC row shows the widest range, from 0.60 to 0.89, indicating that AUC provides the clearest separation among the evaluated models. The close agreement between epoch-level accuracy and patient-level accuracy for the top-performing models also suggests that the Stacking ensemble remains stable across different aggregation levels. In contrast, lower-ranked models showed larger differences between epoch-level and patient-level predictions. This visualization supports the statistical findings in Table 9, confirming that ensemble methods form the strongest performance group. To evaluate the effect of Bayesian optimization, all sixteen models were also tested using default scikit-learn hyperparameters, as shown in Table 10. Without Bayesian optimization, performance decreased across all models, although the size of the decrease varied by model type. Models with more hyperparameters showed the largest drops. AdaBoost decreased by 3.80 AUC points, SVM by 3.30 points, and MLP by 2.90 points, indicating that their default settings were not optimal for EEG-derived ApEn and Std features. In contrast, ensemble and neighborhood-based models showed smaller but still meaningful reductions, including Stacking (−1.40), KNN (−1.40), and Random Forest (−1.35). Importantly, the overall model ranking remained almost unchanged, with Stacking, Voting, and KNN occupying the top three positions in both the default and Bayesian-optimized settings. This shows that the performance hierarchy was mainly driven by model suitability rather than by optimization artifacts. Therefore, Bayesian optimization was retained as an important part of the training pipeline, providing an average AUC improvement of 2.1 percentage points while preserving the same model selection outcome.

5.5. Experiment 5: independent cohort validation

To evaluate the stability of the proposed framework and its ability to generalize beyond the training dataset, we performed external validation using the Boston Children's Hospital (BCH) Sleep Corpus [71]. This is an independent pediatric polysomnography dataset that includes the same six EEG channels used in this study: F3, F4, C3, C4, O1, and O2, along with expert sleep-stage annotations and de-identified clinical diagnosis information. These characteristics make it suitable for assessing cross-dataset generalization. Patient selection followed the same diagnostic criteria used for the NCHSDB cohort. Anxiety-only and stress-only cases were identified from the de-identified diagnosis files and restricted

Table 10
Model performance without Bayesian optimization under 10-fold patient-level stratified cross-validation for anxiety vs. stress classification.

Model	mAcc±std [CI](%)	mPre±std [CI](%)	mRe±std [CI](%)	mF1±std [CI](%)	mAUC±std [CI](%)
K-Nearest Neighbors	83.70±1.30 [82.89-84.51]	83.55±1.32 [82.73-84.37]	83.55±1.35 [82.71-84.39]	83.55±1.33 [82.72-84.38]	86.40±1.25 [85.63-87.17]
MultiLayer Perceptron	78.50±2.00 [77.26-79.74]	78.30±2.05 [77.03-79.57]	78.10±2.02 [76.85-79.35]	78.20±2.03 [76.94-79.46]	81.30±1.95 [80.09-82.51]
Decision Tree	74.80±2.70 [73.13-76.47]	74.60±2.75 [72.90-76.30]	74.70±2.72 [73.02-76.38]	74.65±2.73 [72.96-76.34]	78.10±2.65 [76.46-79.74]
Support Vector Machine	73.50±2.85 [71.74-75.26]	73.20±2.90 [71.40-75.00]	72.90±2.88 [71.12-74.68]	73.05±2.88 [71.27-74.83]	76.50±2.80 [74.76-78.24]
Logistic Regression	65.50±3.40 [63.39-67.61]	65.20±3.45 [63.06-67.34]	65.80±3.42 [63.68-67.92]	65.50±3.43 [63.37-67.63]	68.80±3.35 [66.72-70.88]
Naive Bayes	56.00±4.60 [53.15-58.85]	57.80±4.65 [54.92-60.68]	54.30±4.55 [51.48-57.12]	56.00±4.58 [53.16-58.84]	59.80±4.50 [57.01-62.59]
Random Forest	83.30±1.30 [82.49-84.11]	83.00±1.32 [82.18-83.82]	83.20±1.35 [82.36-84.04]	83.10±1.33 [82.28-83.92]	85.90±1.25 [85.13-86.67]
XGBoost	81.85±1.50 [80.92-82.78]	81.70±1.52 [80.76-82.64]	81.85±1.55 [80.89-82.81]	81.78±1.53 [80.83-82.73]	84.40±1.45 [83.50-85.30]
Gradient Boosting	77.20±2.40 [75.71-78.69]	76.90±2.45 [75.38-78.42]	77.10±2.42 [75.60-78.60]	77.00±2.43 [75.49-78.51]	80.00±2.35 [78.54-81.46]
AdaBoost	68.00±3.20 [66.02-69.98]	67.70±3.25 [65.69-69.71]	67.50±3.22 [65.51-69.49]	67.60±3.23 [65.60-69.60]	71.00±3.15 [69.05-72.95]
Voting	84.40±1.20 [83.66-85.14]	84.15±1.22 [83.39-84.91]	84.35±1.25 [83.58-85.12]	84.25±1.23 [83.49-85.01]	87.20±1.15 [86.49-87.91]
Stacking	84.80±1.10 [84.12-85.48]	84.65±1.12 [83.96-85.34]	84.55±1.15 [83.84-85.26]	84.60±1.13 [83.90-85.30]	87.80±1.05 [87.15-88.45]
FIRE-DESKNN	82.10±1.50 [81.17-83.03]	82.25±1.52 [81.31-83.19]	81.80±1.55 [80.84-82.76]	82.02±1.53 [81.07-82.97]	84.70±1.45 [83.80-85.60]
FIRE-KNOP	83.20±1.45 [82.30-84.10]	82.65±1.48 [81.73-83.57]	82.75±1.50 [81.82-83.68]	82.70±1.48 [81.78-83.62]	85.60±1.40 [84.73-86.47]
FIRE-KNORAE	82.95±1.48 [82.03-83.87]	83.05±1.50 [82.12-83.98]	82.70±1.52 [81.76-83.64]	82.88±1.50 [81.95-83.81]	85.50±1.42 [84.62-86.38]
FIRE-KNORAU	83.35±1.30 [82.54-84.16]	83.20±1.32 [82.38-84.02]	83.30±1.35 [82.46-84.14]	83.25±1.33 [82.43-84.07]	86.00±1.25 [85.23-86.77]

Table 11
External validation results comparing NCH-trained Stacking ensemble performance on NCH cross-validation folds versus independent BCH cohort. Results reported as mean ± standard deviation with 95% confidence intervals.

Model	NCH → NCH					NCH → BCH				
	mAcc ±std [CI]	mPre ±std [CI]	mRe ±std [CI]	mF1 ±std [CI]	mAUC ±std [CI]	mAcc ±std [CI]	mPre ±std [CI]	mRe ±std [CI]	mF1 ±std [CI]	mAUC ±std [CI]
Stacking	86.30±0.75 [85.76-86.84]	86.15±0.78 [85.59-86.71]	86.05±0.80 [85.48-86.62]	86.10±0.77 [85.55-86.65]	89.20±0.65 [88.74-89.66]	80.50±2.50 [78.71-82.29]	79.20±3.00 [77.05-81.35]	79.80±2.80 [77.80-81.80]	79.50±2.90 [77.43-81.57]	84.50±2.50 [82.71-86.29]

to the pediatric age range used in the main experiments. After applying these criteria, a near-balanced patient-level subset of 54 patients was obtained, including 30 anxiety-only and 24 stress-only patients. The same preprocessing pipeline used in Experiments 1–3 was then applied without modification. This included two-hour EEG segments after excluding the first hour, ApEn and Std feature extraction, the six-channel montage, 1–60 Hz bandpass filtering, and 60 Hz notch filtering. Importantly, the Stacking ensemble trained on the full NCHSDB cohort was applied directly to the BCH data, with no retraining, hyperparameter adjustment, or threshold calibration. Therefore, the reported results reflect external generalization rather than dataset-specific tuning.

Table 11 presents the comparative results. On the NCHSDB cohort, the Stacking ensemble achieved an accuracy of 86.30 ± 0.75%, precision of 86.15 ± 0.78%, recall of 86.05 ± 0.80%, F1-score of 86.10 ± 0.77%, and AUC of 89.20 ± 0.65%, consistent with the 10-fold cross-validation results reported in Experiment 4. On the independent BCH cohort, performance decreased, as expected, due to cross-hospital domain differences. The model achieved an accuracy of 80.50 ± 2.50%, precision of 79.20 ± 3.00%, recall of 79.80 ± 2.80%, F1-score of 79.50 ± 2.90%, and AUC of 84.50 ± 2.50%. The AUC decreased by 4.70 percentage points, showing a moderate but acceptable reduction when moving from the original dataset to an independent hospital cohort. The larger standard deviations for BCH also indicate greater variability on unseen data, which is expected in external validation. Notably, the AUC on BCH remained higher than the accuracy (84.50% vs. 80.50%). This suggests that the model preserved useful discriminative ability, even though the decision threshold may not have been fully calibrated for the new dataset. In other words, the model could still rank anxiety and stress cases meaningfully, although exact classification became more challenging under cross-dataset conditions. These findings show that the proposed anxiety and stress detection framework can generalize to independent pediatric sleep EEG data, with the expected performance reduction caused by differences between hospital datasets. This performance trend is shown in Fig. 9.

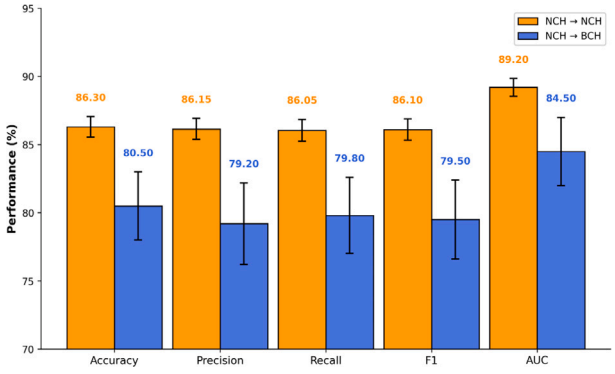


Fig. 9. External validation comparison between NCH cross-validation (Orange) and independent BCH cohort (blue) for the Stacking ensemble across five evaluation metrics.

5.6. Experiment 6: model explainability and interpretation

To enhance the interpretability of our machine learning framework, we employed multiple XAI techniques, namely SHAP, LIME, and NeuroXAI. These methods were selected to provide complementary perspectives on model behavior: SHAP and LIME offer feature-level explanations for global and instance-specific insights, while NeuroXAI enables channel-level, spatially grounded interpretation specifically tailored for multi-channel EEG data. The application of these XAI methods serves multiple objectives. First, they provide transparency into the decision-making process of black-box models by identifying the most influential features and EEG channels contributing to each prediction. Second, they support the reduction of input dimensionality by highlighting features or channels with minimal importance, thereby facilitating more efficient model training and streamlined preprocessing. Finally,

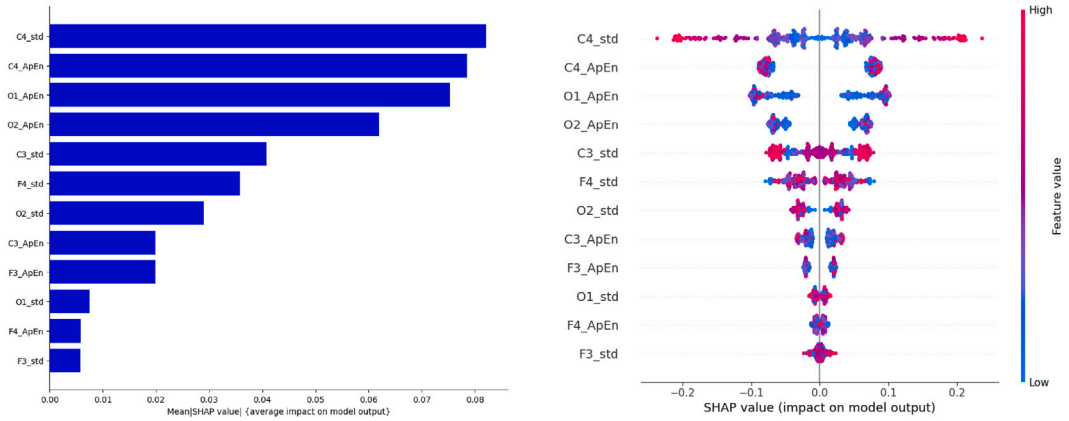


Fig. 10. SHAP plots for feature importance in the model anxiety and stress detection. (a) SHAP summary bar plot of feature importance in the model. (b) SHAP beeswarm plot of feature importance in the model.

these insights contribute to a deeper understanding of the neurophysiological markers associated with anxiety and stress, offering potential utility in both clinical diagnostics and wearable EEG system design [21]. Among the evaluated models, the RF classifier was selected for XAI interpretation due to its strong performance and compatibility with feature-level explanation techniques. While ensemble methods such as stacking achieved slightly higher accuracy, their layered complexity makes direct interpretability more challenging. In contrast, RF offers a balance between model fidelity and interpretability, making it suitable for post-hoc explanation. In the following subsections, we present and compare the results of SHAP, LIME, and NeuroXAI. Each method is applied to the set of high-confidence predictions to analyze consistency across techniques and to extract spatial and functional insights. By triangulating results from these three complementary methods, we aim to uncover robust patterns of channel-level importance that differentiate anxiety from stress.

5.6.1. SHAP-based global feature explanation

Fig. 10 illustrates the global interpretability results obtained through SHAP analysis applied to the RF classifier. The summary bar chart organizes EEG-derived features according to their average influence on the model's predictions. The top-ranked contributors included measures of signal variability and entropy, particularly from the central channel C4 and the occipital channels O1 and O2. This pattern suggests that both signal complexity and dynamic fluctuations in these regions significantly impact the model's ability to differentiate between anxiety and stress. In addition, the global SHAP beeswarm plot offers insight into the direction and strength of each feature's impact. For instance, elevated entropy and variability in signals from the central region (notably C4) were predominantly associated with anxiety classifications. Conversely, lower values for these features appeared more frequently in stress predictions. These findings suggest that temporal irregularity and information richness in central brain activity may serve as key indicators of anxiety-related neural dynamics. These interpretability results are consistent with prior findings in neuroscience and affective computing [41,43,72–74], which highlight the relevance of central and occipital brain regions in the assessment of emotional and cognitive states. The alignment between our model's explanations and established domain knowledge enhances the credibility and clinical interpretability of the classification outcomes.

5.6.2. LIME-based local prediction explanation

To deepen our understanding of the model's internal decision-making, we applied local interpretability techniques to two individual EEG instances using both LIME and SHAP. These methods provide transparent, per-instance explanations of how individual features contributed to the final classification outcome. For instance 1, shown in Figs. 11(a) and 11(b), the model confidently predicted the class anxiety. Both LIME and SHAP identified signal variability in the central region (C4) as the most influential factor in the prediction. Additional contributors included variability in the occipital (O2) and frontal (F3) regions, whereas features such as occipital variability (O1) and entropy in the central region (C4) were weakly associated with the opposite class, stress. Notably, most entropy-based features contributed minimally in this case, suggesting that the prediction was primarily driven by temporal variability rather than signal complexity. In instance 2, illustrated in Figs. 11(c) and 11(d), the model classified the signal as stress with high confidence. Both LIME and SHAP again agreed that the decision was predominantly influenced by high variability in the central (C4 and C3) and frontal (F3) regions, with the central channel (C4) showing the strongest impact. Minor counter-effects were observed from features such as central entropy (C4) and occipital variability (O1), which weakly favored the alternative class, anxiety. These local explanations confirm a consistent pattern across both instances, showing that variability features from central and frontal channels were the most influential, while entropy-based features played a smaller role. This highlights that temporal variability in the EEG signal, rather than its complexity, was more critical in distinguishing anxiety from stress on a per-instance basis. Interestingly, many of the same regions, particularly central (C3, C4) and frontal (F3), emerged as important in both cases. However, their influence varied in direction and magnitude depending on the instance, demonstrating the model's ability to adaptively interpret subtle shifts in signal dynamics. For example, while increased variability in the central and occipital regions contributed to the anxiety classification in instance 1, those same signal characteristics supported a stress prediction in instance 2 due to differing contextual thresholds. From a neurophysiological standpoint, the recurrence of central and frontal regions is consistent with existing literature [41,43,72–74]. Both anxiety and stress are associated with activity in overlapping cortical areas, particularly those monitored by electrodes in the central (C3, C4) and frontal (F3, F4) regions. The distinction between these mental states thus lies not in which brain regions are involved, but in how the signal patterns differ in variability and timing, which this model appears capable of capturing through nuanced, context-aware prediction logic.

5.6.3. NeuroXAI-based channel-level explanation

While SHAP and LIME provide valuable insights into model behavior through global and local feature contributions, they are not

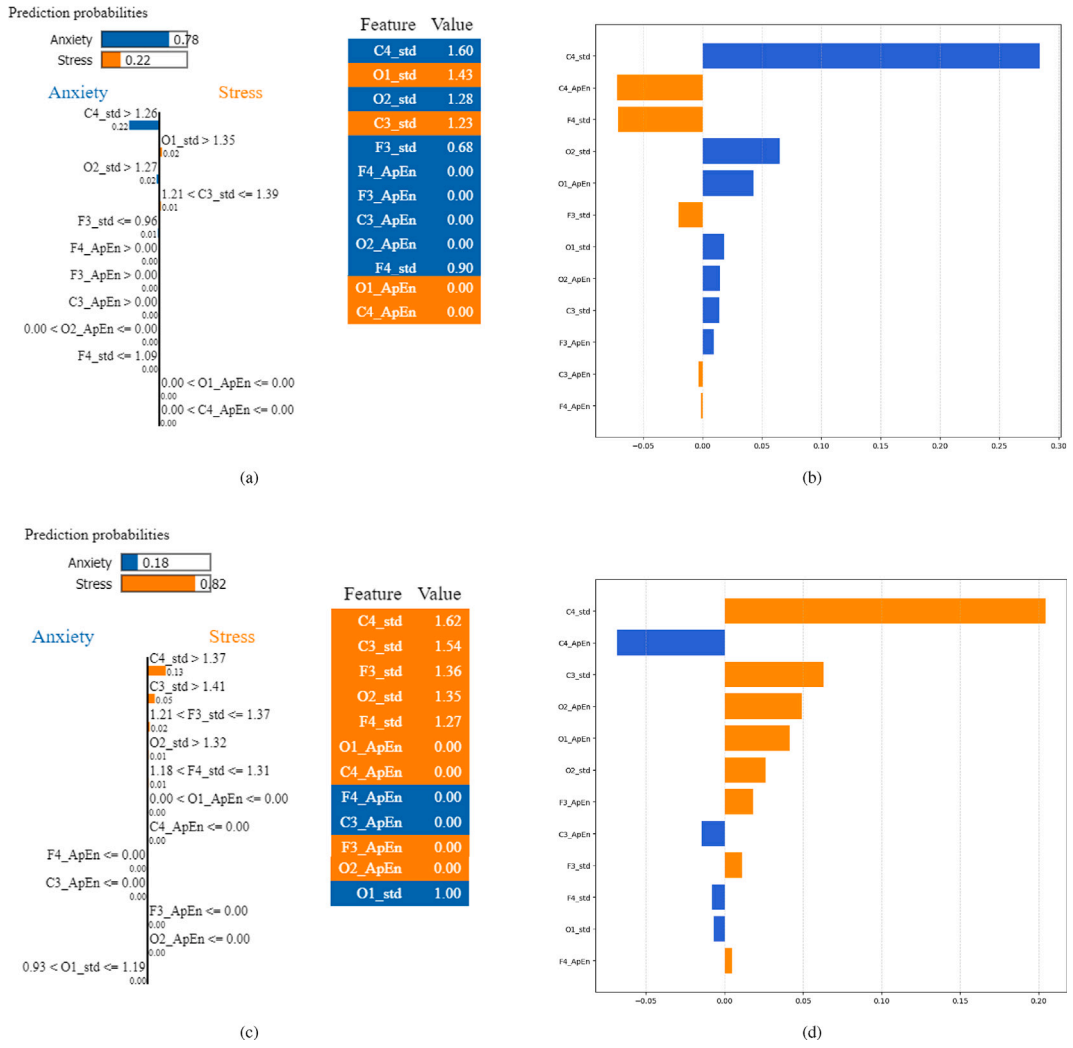


Fig. 11. Visual explanations of the model's predictions for anxiety and stress classification. (a) LIME explanation for a sample predicted as anxiety, showing which features influenced the decision. (b) SHAP bar plot for the same sample, highlighting each feature's impact. (c) LIME explanation for a sample predicted as stress. (d) SHAP bar plot showing how features contributed to the stress prediction.

inherently designed to account for the spatiotemporal structure of EEG signals. To address this limitation, we incorporated NeuroXAI [70], a domain-specific explainability framework tailored for neural data. NeuroXAI employs perturbation-based strategies that systematically assess the importance of each EEG channel by evaluating its influence on the classifier's output. In this work, NeuroXAI was applied in a post-hoc setting to analyze each EEG instance, where all six electrodes (F3, F4, C3, C4, O1, O2) were individually perturbed. The resulting changes in prediction probability were used to compute channel importance scores, which were then visualized using horizontal bar plots and topographic scalp maps (topomaps). These visualizations facilitate intuitive spatial interpretation of the model's decision-making process. In this section, we illustrate NeuroXAI outputs alongside SHAP and LIME explanations for six representative instances, three predicted as anxiety (A1, A2, A3) and three as stress (S1, S2, S3), each with prediction confidence exceeding 75%. Since EEG interpretation is typically conducted at the channel level, SHAP and LIME feature importances were aggregated accordingly. For a given EEG channel C with associated features $\{x_{C1}, x_{C2}, \ldots, x_{Ck}\}$, the total channel relevance is computed as:

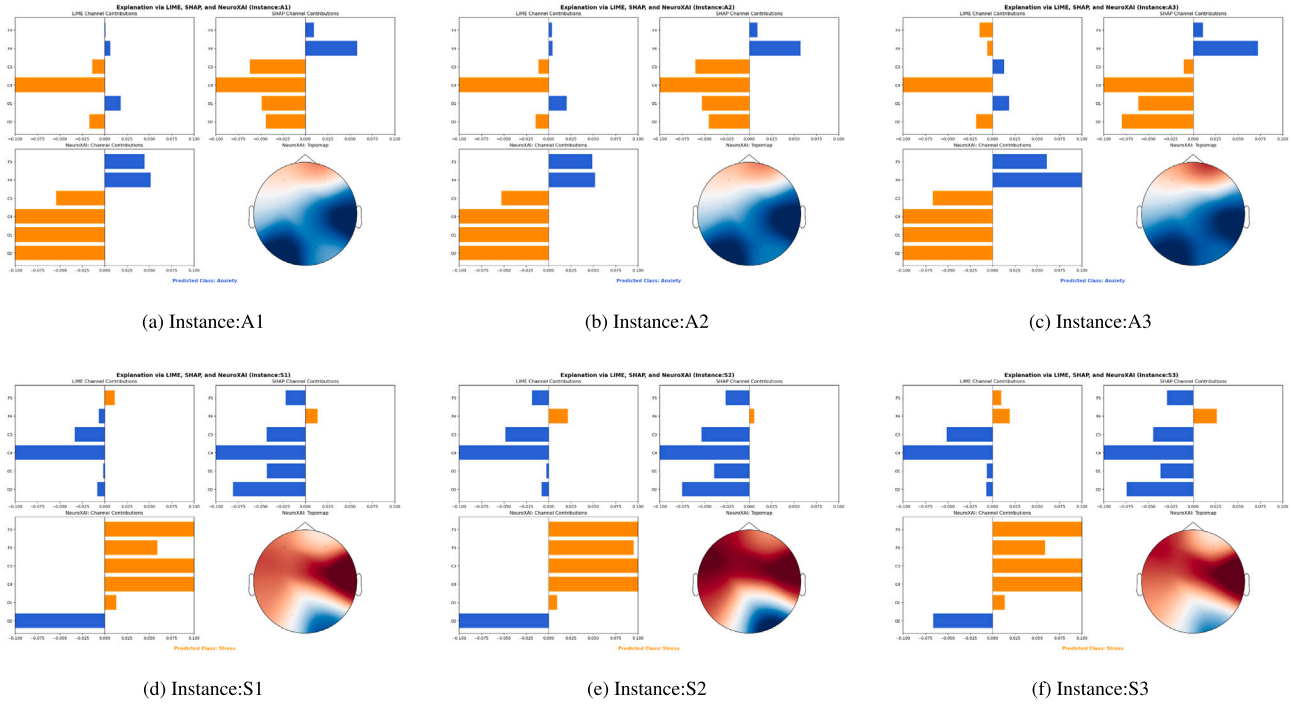
$$\Phi_C = \sum_{j=1}^{k} \phi_{Cj}, \tag{10}$$

Table 12
Channel-Level positive contributions Across explainability Methods. This table summarizes the EEG channels contributing positively to anxiety or stress predictions across LIME, SHAP, and NeuroXAI. A1—A3 correspond to anxiety instances, while S1—S3 represent stress.

Instance	Predicted Class	LIME	SHAP	NeuroXAI
A1	Anxiety	F3, F4, O1	F3, F4	F3, F4
A2	Anxiety	F3, F4, O1	F3, F4	F3, F4
A3	Anxiety	C3, O1	F3, F4	F3, F4
S1	Stress	F3	F4	F3, F4, C3, C4, O1
S2	Stress	F4	F4	F3, F4, C3, C4, O1
S3	Stress	F3, F4	F4	F3, F4, C3, C4, O1

where ϕ_{Cj} denotes the importance score of the j-th feature from channel C. This aggregation aligns feature-level XAI results with NeuroXAI's spatial analysis for consistent interpretation and enables fair comparison across the three interpretability methods, as summarized in Table 12 and visualized in Fig. 12.

As detailed in Table 12, a clear pattern of spatial convergence emerges as follows: For anxiety predictions, NeuroXAI consistently emphasized frontal regions (F3, F4) across all three instances. SHAP showed perfect agreement with NeuroXAI, assigning strong importance to both



(a) Instance:A1 (b) Instance:A2 (c) Instance:A3

(d) Instance:S1 (e) Instance:S2 (f) Instance:S3

Fig. 12. Visual comparison of channel-level attributions for anxiety and stress predictions using LIME, SHAP, and NeuroXAI. Instances A1–A3 represent anxiety predictions, and S1–S3 represent stress. Each figure illustrates channel-wise importance using LIME and SHAP (bar plots), alongside NeuroXAI visualizations (bar plot and topomap). Blue bars denote contributions toward anxiety, while orange bars indicate support for stress. Red and blue colors in topomaps represent high and low channel contributions, respectively. These visual patterns align with the aggregated results presented in Table 7.

frontal regions (F3, F4) in every case. LIME similarly identified frontal regions (F3, F4) and the occipital region (O1) in A1 and A2, but shifts focus to the central region (C3) and the occipital region (O1) in A3, diverging from SHAP and NeuroXAI. This divergence illustrates the inherently local nature of LIME, while the alignment of SHAP and NeuroXAI suggests these frontal channels are reliable predictors of anxiety within the EEG model's internal logic. For stress predictions, NeuroXAI consistently identifies frontal (F3, F4) and central (C3, C4) regions as the most influential across all three samples. Additionally, channel O1 is also involved in every case, as shown in the NeuroXAI outputs. LIME provided more focused attributions of the frontal region, highlighting channels, such as, F3 in S1, F4 in S2, and both F3 and F4 in S3. SHAP uniformly emphasizes F4 in all stress cases, suggesting a consistent frontal contribution. Despite variations in specificity and granularity, the convergence across methods reinforces the diagnostic relevance of central-frontal regions for stress recognition, whereas NeuroXAI offers the most spatially comprehensive interpretation. Integrating NeuroXAI into our pipeline not only provides channel-level insights but also bridges the gap between black-box predictions and domain-relevant EEG physiology. This unified interpretability framework offers a comprehensive view of the model's reasoning and supports its clinical deployment in mental health diagnostics [21]. Our multimethod explainability analysis reveals that anxiety is primarily characterized by the dominance of frontal (F3, F4) channels. In contrast, frontal (F3, F4) and central EEG channels (C3, C4) are the most influential in classifying stress, consistently showing elevated contribution scores across NeuroXAI, SHAP, and LIME. While the three XAI methods emphasize different aspects, global patterns (SHAP), local prediction rationales (LIME), and spatial channel relevance (NeuroXAI), which complementary insights converged on the frontal and central regions, strengthening the reliability of interpretation. This spatial distinction suggests that while both emotional states engage frontal brain activity, stress involves broader sensorimotor regions, whereas anxiety remains more frontally concentrated. These findings align with existing neurophysiological evidence and reinforce the clinical interpretability of our EEG-based classifier.

5.7. Experiment 7: sensitivity analysis

To further assess the robustness of the proposed framework, we conducted two sensitivity analyses: a multiclass diagnostic evaluation and an age-based subgroup analysis. These experiments were designed to examine whether the final pipeline remained stable under a more complex class setting and across different pediatric age ranges.

5.7.1. Multiclass sensitivity analysis

In addition to the main binary anxiety versus stress classification task, we performed a multiclass sensitivity experiment with three groups: anxiety-only, stress-only, and comorbid anxiety + stress. The same preprocessing pipeline was used, including two-hour EEG segments, ApEn and Std features, six EEG channels, 1–60 Hz bandpass filtering with a 60 Hz notch filter, and patient-level stratified 10-fold cross-validation. This ensured that the multiclass results could be compared fairly with the main binary setting. Model performance was assessed using multiclass accuracy, macro-precision, macro-recall, macro-F1, and macro one-versus-rest AUC. Table 13 reports the results of this experiment. The Stacking ensemble achieved an accuracy of 76.84 ± 3.12% [74.61–79.07], macro-precision of 74.92 ± 3.45% [72.45–77.39], macro-recall of 73.68 ± 3.87% [70.91–76.45], macro-F1 of 73.95 ± 3.51% [71.44–76.46], and macro-AUC of 82.11 ± 2.74% [80.15–84.07]. Compared with the binary setting, where the model achieved an AUC of 89.20% and an accuracy of 86.30%, the multiclass task showed a reduction of 7.09 percentage points in AUC and 9.46 percentage points in accuracy. This decrease is expected because the comorbid class introduces an additional category that shares characteristics with both anxiety and stress, making the decision boundaries less

Table 13
Results of the multiclass sensitivity analysis. Metrics are reported as mean ± standard deviation with 95% confidence intervals.

Setting	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Macro-AUC
Anxiety / Stress / Comorbid	76.84 ± 3.12 [74.61-79.07]	74.92 ± 3.45 [72.45-77.39]	73.68 ± 3.87 [70.91-76.45]	73.95 ± 3.51 [71.44-76.46]	82.11 ± 2.74 [80.15-84.07]

Table 14
Results of the age-based sensitivity analysis for binary anxiety versus stress classification. Metrics are reported as mean ± standard deviation with 95% confidence intervals.

Age Group	Subjects	mAcc (%)	mPre (%)	mRec (%)	mF1 (%)	mAUC (%)
0–5 years	A: 28, S: 162	79.42 ± 4.86 [75.94-82.90]	68.35 ± 6.12 [63.97-72.73]	65.18 ± 6.45 [60.57-69.79]	66.21 ± 5.98 [61.93-70.49]	81.47 ± 3.91 [78.67-84.27]
6–10 years	A: 72, S: 70	85.63 ± 3.24 [83.31-87.95]	85.11 ± 3.56 [82.56-87.66]	84.92 ± 3.71 [82.27-87.57]	84.88 ± 3.43 [82.43-87.33]	88.76 ± 2.95 [86.65-90.87]
11–17 years	A: 120, S: 80	83.94 ± 3.78 [81.24-86.64]	81.72 ± 4.14 [78.76-84.68]	80.95 ± 4.32 [77.86-84.04]	81.08 ± 4.06 [78.18-83.98]	86.91 ± 3.18 [84.64-89.18]

A: Anxiety, S: Stress.

distinct. The larger standard deviations in the multiclass setting also suggest greater uncertainty when separating three clinically overlapping groups. Despite this reduction, the macro-AUC of 82.11% shows that the model still maintained meaningful discriminative ability in the three-class setting. The results suggest that ApEn and Std features capture useful EEG patterns beyond the main binary task. However, the binary anxiety versus stress configuration remained more stable and produced stronger overall performance. Therefore, the multiclass setting should be viewed as an exploratory extension, particularly useful when identifying comorbid cases is clinically important, while recognizing that the overlap between anxiety, stress, and comorbidity can limit class-specific precision and recall.

5.7.2. Age-based sensitivity analysis

To examine whether the proposed framework generalizes across pediatric developmental stages, we performed an age-stratified sensitivity analysis by dividing the NCHSDB cohort into three groups: early childhood (0–5 years), middle childhood (6–10 years), and adolescence (11–17 years). The same preprocessing pipeline, feature extraction strategy, and Stacking ensemble configuration were used for each group, allowing age-related differences in performance to be assessed without changes in the experimental setting. Table 14 presents the results. The 6–10 year group achieved the best overall performance, with an accuracy of 85.63 ± 3.24% and an AUC of 88.76 ± 2.95%, which was close to the full-cohort binary results. This group also had the most balanced class distribution, with 72 patients with anxiety and 70 patients with stress, which likely helped the model learn more stable decision boundaries. In contrast, the 0–5 year group showed lower performance, with an accuracy of 79.42 ± 4.86% and an AUC of 81.47 ± 3.91%. This reduction was mainly related to the strong class imbalance in this group, with 28 patients with anxiety and 162 patients with stress. As a result, anxiety detection was less reliable, as reflected by a lower precision of 68.35 ± 6.12% and a recall of 65.18 ± 6.45%. The 11–17 year group showed intermediate performance, with an AUC of 86.91 ± 3.18%, suggesting that model performance was influenced more by class balance than by age alone. These findings suggest that the framework performs reliably in school-age children and adolescents, while results in early childhood should be interpreted with caution. The weaker performance in the 0–5 year group appears to be mainly due to limited anxiety cases and greater variability in early childhood EEG patterns. Future work should include more balanced recruitment in younger children or consider cost-sensitive learning approaches to reduce majority-class bias.

5.8. Comparison with the literature studies

Although prior works have not explored anxiety and stress classification simultaneously using sleep EEG data, we report comparisons with studies addressing these conditions independently to provide a perspective on model performance. Table 15 provides a comprehensive overview of recent research efforts that have applied machine learning to EEG data for the detection of anxiety and stress. A key distinction among these studies is that nearly all were conducted on small adult cohorts, whereas the present work is uniquely centered on a large pediatric population, highlighting the feasibility of EEG-based biomarkers for childhood mental health. To date, no prior study has combined pediatric sleep EEG with explainability for anxiety and stress classification. Several studies relied on relatively small datasets, including [44] with 12 subjects, Aldayel and Al-Nafjan [41] with 23 subjects, and [48] with 7 subjects. While these studies report strong classification results, limited participant numbers can reduce the reliability of the models in broader applications, particularly due to risks of overfitting and lack of general variability in EEG signals. In contrast, the present study draws on a large-scale public pediatric EEG dataset containing 532 clinically diagnosed children [54], allowing for more statistically robust outcomes and greater potential for clinical generalization in early-life settings. Another key difference lies in the focus on EEG channels and brain regions. While not all studies specify which regions were most important, several did emphasize particular areas. For instance, [43,44], and [45] highlighted the importance of frontal channels, while [74] pointed to a broader set including frontal, central, and occipital regions. These findings align closely with findings from the current study, which show that the frontal (F3, F4) and central (C3, C4) regions are important for differentiating between anxiety and stress. Feature importance analysis in Section 5.6 using SHAP and LIME confirmed the significant role of channels such as C3 and C4 in distinguishing between anxiety and stress. This is supported by earlier research suggesting that central EEG activity often reflects somatic stress responses, while frontal signals are closely linked to emotional processing and cognitive control [43,72,74]. Although some studies, such as those by Luo et al. [45] and Shen et al. [42], report very high accuracy scores above 97%, these results were obtained from brief EEG recordings typically under 15 s and on small participant groups. Such conditions may not reflect real-world performance, especially in clinical settings where variability is high. By contrast, the model proposed in this study achieved an accuracy of 86.30% using 2-hour EEG recordings, a stacking ensemble approach, and interpretable statistical features. The integration of domain-relevant EEG regions, a robust feature extraction strategy, and ensemble modeling contributes to a method that is not only accurate but also scalable, non-invasive, and practical for future applications in pediatric mental health screening.

6. Study limitations and future work

Several limitations of this study should be acknowledged. First, the NCHSDB contains polysomnography recordings collected during overnight sleep monitoring. In pediatric populations, full PSG montages can cause physical discomfort, and involuntary movements during sleep may introduce motion and electrode artifacts. Although artifact rejection and notch filtering were applied, some residual noise may still have affected feature stability, especially for amplitude-sensitive features such as Standard Deviation. Future work should evaluate wake-state EEG datasets with fewer electrodes, such as 4–6 frontal or central

Table 15
Comparison with the recent literature studies that use ML techniques to detect anxiety and stress using EEG data.

Ref.	Diagnosis	Dataset	No. of Patient(s)	No. of Channel(s)	Significant Channel(s)	Signal Duration	Filter(s) (Hz)	Sampling Frequency (Hz)	Feature Extraction	ML Classifier	Acc(%)	Pre(%)	Re(%)	F1(%)	Explainability
Ours	Anxiety, Stress	NCHSDB [54], BCH [71]	532, 54	6	Frontal, Central	2 hr	Band pass (1-60)	256	ApEn, Std	Stacking Ensemble	86.30	86.15	86.05	86.10	SHAP, LIME, NeuroXAI
[41]	Anxiety	DASPS [75]	23	10	–	1 sec	Band pass (4-45)	128	DWT	RF	87.5	87.65	87.5	–	–
[42]	Anxiety	Collected Data	45	16	–	10 min	4th order Butterworth (4-30)	100	PSD	SVM	97.83	–	97.55	97.95	–
[44]	Anxiety	Collected Data	12	6	Frontal	15 min	Band pass, HW (0.5-50)	500	Hjorth, ApEn, Mean square	SVM	62.56	–	–	–	–
[45]	Anxiety	Collected Data	80	16	Frontal, Temporal	10 sec	Band pass (4-30)	125	PLI	CatBoost	98.1	–	–	–	–
[43]	Anxiety	Collected Data	40	8	Frontal	5 min	–	128	Hjorth	SVM	87.03	85.93	93.73	90.00	–
[52]	Stress	Collected Data	25	8	–	6 min	Band pass (0.1-40)	500	DWT	SVM	95.83	97.78	96.70	–	–
[48]	Stress	Collected Data	7	14	–	5 sec	Band pass (0.5-64)	64	Correlation & Wrapper	SVM	80.32	–	–	–	–
[49]	Stress	Collected Data	28	1	–	3 min	Band pass (1-50)	512	FFT	SVM	78.57	–	–	66.20	–
[50]	Stress	Collected Data	10	4	–	8 min	4th order Butterworth (1-48)	1000	Correlation Analysis	LDA	86.00	–	–	–	–
[53]	Stress	Collected Data	310	14	–	15 sec	–	128	CubicPat	tKNN	96.29	–	–	96.22	–
[74]	Stress	Collected Data	28	8	Frontal, Central, Occipital, Temporal	5 min	–	250	wavelet	SVM	87.5	–	81.25	–	–

channels, to determine whether the ApEn and Std features remain effective in less constrained recording settings. Second, the proposed framework relied on handcrafted features, ApEn, and Std, extracted from 30-second EEG epochs. This design supports compatibility with standard sleep-stage annotations and clinical workflows, but it does not allow the model to learn spatiotemporal representations directly from raw EEG signals. Future studies may explore deep learning architectures such as 1-D CNNs for spectral feature learning, LSTMs or Transformers for temporal dependency modeling, and graph neural networks for inter-channel spatial relationships. These models may capture transient sleep-related patterns, such as K-complexes and sleep spindles, that may not be fully represented by window-level entropy and amplitude features. However, their clinical value should be tested through strong cross-site validation. Third, the epoch duration was fixed at 30-second to remain consistent with standard sleep-scoring practice. This setting may miss clinically relevant EEG changes that occur at shorter or longer time scales. Multiscale epoch analysis using shorter and longer windows could help determine whether anxiety and stress signatures are linked to brief events or broader sleep-state transitions. Fourth, although external validation using the Boston Children's Hospital Sleep Corpus supports cross-site generalizability, inter-dataset differences remain important. Variations in recording hardware, sampling rates, electrode materials, and diagnostic coding practices between NCHSDB and BCH datasets may have contributed to the observed 4.7 percentage-point AUC reduction. Further validation using multi-center pediatric cohorts with broader demographic and geographic diversity is needed before clinical translation. Finally, although SHAP and LIME provided post-hoc feature attribution, a gap remains between algorithmic explanation and clinical interpretation. Future work should convert SHAP-based importance patterns into more clinically useful visual tools, such as scalp topographic heatmaps based on the standard 10–20 system, so that sleep technologists and pediatric neurologists can more easily review and interpret model findings.

7. Conclusion

This study investigated the use of pediatric sleep EEG for the interpretable classification of anxiety and stress. Using data from the NCHSDB, we developed a complete framework that included signal preprocessing, artifact rejection, annotation alignment, feature extraction, model training, and explainability analysis. A standardized two-hour sleep segment from 532 children was analyzed, making this study one of the first large-scale investigations of pediatric anxiety and stress using routine sleep EEG recordings. Several machine learning models were evaluated, including classical classifiers, dynamic ensemble selection methods, and static ensemble models. The results showed that ensemble-based learning can provide a useful direction for pediatric EEG-based mental health classification. In particular, the stacking ensemble showed strong overall performance, while dynamic ensemble selection methods also demonstrated the value of adaptive classifier use. These findings suggest that combining multiple classifiers can improve the reliability of EEG-based screening support compared with relying on a single model. To examine generalizability, an external validation experiment was conducted using an independent cohort from the Boston Children's Hospital Sleep Corpus. This evaluation helped assess whether the learned patterns were specific to the NCHSDB cohort or could transfer to another pediatric sleep EEG dataset. Although differences in cohort characteristics, recording conditions, and diagnostic distributions can affect cross-cohort performance, the external validation provides an important step toward evaluating the robustness and practical relevance of the proposed approach beyond a single dataset. To improve transparency and clinical interpretability, three complementary explainability methods were applied, namely SHAP, LIME, and NeuroXAI. SHAP and LIME provided feature-level explanations, while NeuroXAI supported channel-wise interpretation of the model decisions. The explanation results showed meaningful spatial patterns.

Stress-related predictions were mainly influenced by frontal and central EEG activity, particularly F3, F4, C3, and C4, whereas anxiety-related predictions were more strongly associated with frontal regions, especially F3 and F4. These findings are consistent with previous evidence linking frontal and central brain activity with emotional regulation and stress responses, while extending these observations to pediatric sleep EEG. Compared with earlier EEG studies on anxiety and stress, which often used small adult cohorts, task-based recordings, or only spectral features, this work shows that routine sleep EEG from a large pediatric population can provide useful and clinically meaningful information for distinguishing between these conditions. The differences from previous studies may be due to variations in age group, recording conditions, feature design, and evaluation protocol. This further highlights the value of pediatric sleep EEG as a distinct and important source of information for mental health research. In conclusion, this study presents an explainable ensemble learning framework for pediatric anxiety and stress screening using sleep EEG. By combining patient-level evaluation, external validation, optimized machine learning models, and multiple explainability methods, including SHAP, LIME, and NeuroXAI, the proposed framework provides both predictive performance and transparent interpretation. These findings offer a useful foundation for future screening support systems that are noninvasive, scalable, and better aligned with pediatric clinical practice.

CRediT authorship contribution statement

Maria Bashir: Writing – original draft, Methodology, Formal analysis. **Shaker El-Sappagh:** Writing – review & editing, Validation, Investigation. **Waleed Nazih:** Software, Resources. **Abolghasem Sadeghi-Niaraki:** Visualization, Conceptualization. **Tamer Abuhmed:** Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition.

Ethics approval

This study utilized de-identified data from the Nationwide Children's Hospital and Boston Children's Hospital Sleep Corpus, resources hosted on PhysioNet and Brain Data Science Platform. Ethical approval and informed consent were obtained by the original data providers. As this research involves secondary analysis of anonymized data, no additional ethical approval or informed consent was required.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. RS-2021-NR058558), and the (Institute for Information & Communications Technology Planning & Evaluation) (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program (IITP-2026-RS-2020-II201821), AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (No. 2022-0-00688).

Data availability

The primary EEG dataset used in this study was obtained from the Nationwide Children's Hospital Sleep DataBank, a publicly available pediatric polysomnography repository hosted on PhysioNet. The dataset comprises 3984 polysomnography recordings from 3673 pediatric patients, accompanied by clinical and demographic information extracted from electronic health records. It can be accessed at: https://physionet.org/content/nch-sleep/1.0.0/, subject to acceptance of the PhysioNet credentialed access agreement.

The external validation cohort used in this study was obtained from the Boston Children's Hospital Sleep Corpus, a large-scale collection of 15,695 annotated pediatric polysomnography recordings along with expert sleep-stage annotations and de-identified clinical diagnosis information. The Corpus is accessible via the Brain Data Science Platform at: https://bdsp.io/content/l8c86mgywuneuy74ae71/, subject to the applicable data access terms and conditions.

References

[1] W.H. Organization, et al., Mental Health and COVID-19: Early Evidence of the Pandemic's Impact: Scientific Brief, 2 March 2022, Tech. rep., World Health Organization, 2022.
[2] A.S. Cortés, Diagnostic and statistical manual of mental disorders, text revision (DSM-5-TR), in: American Psychiatric Association, vol. 19, Psicooncología, 2022, pp. 339, https://doi.org/10.1176/appi.books.9780890425787. (2).
[3] World Health Organization, Mental health atlas 2020, 2021. https://www.who.int/publications/i/item/9789240036703 (Accessed: 26 May 2025).
[4] J.R. Strawn, L. Lu, T.S. Peris, A. Levine, J.T. Walkup, Research review: pediatric anxiety disorders–what have we learnt in the last 10 years? J. Child Psychol. Psychiatry 62 (2) (2021) 114–139.
[5] O. Belei, D.-G. Basaca, L. Olariu, M. Pantea, D. Bozgan, A. Nanu, I. Sîrbu, O. Mărginean, I. Enătescu, The interaction between stress and inflammatory bowel disease in pediatric and adult patients, J. Clin. Med. 13 (5) (2024) 1361.
[6] N. Guo, J. Koerts, L. Tucha, I. Fetter, C. Biela, M. König, M. Bossert, C. Diener, S. Aschenbrenner, M. Weisbrod, et al., Stability of attention performance of adults with ADHD over time: evidence from repeated neuropsychological assessments in one-month intervals, Int. J. Environ. Res. Public Health 19 (22) (2022) 15234.
[7] A. Uchida, J.A. Pillai, R. Bermel, S.E. Jones, H. Fernandez, J.B. Leverenz, S.K. Srivastava, J.P. Ehlers, Correlation between brain volume and retinal photoreceptor outer segment volume in normal aging and neurodegenerative diseases, PLOS One 15 (9) (2020) e0237078.
[8] A. Shoeibi, M. Khodatars, M. Jafari, N. Ghassemi, P. Moridian, R. Alizadehsani, S.H. Ling, A. Khosravi, H. Alinejad-Rokny, H.-K. Lam, et al., Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: a review, Inf. Fusion 93 (2023) 85–117.
[9] S. Liu, M. Duan, Y. Sun, L. Wang, L. An, D. Ming, Neural responses to social decision-making in suicide attempters with mental disorders, BMC Psychiatry 23 (1) (2023) 19.
[10] H.-M. Wu, F.-J. Hsiao, R.-S. Chen, D.-E. Shan, W.-Y. Hsu, M.-C. Chiang, Y.-Y. Lin, Attenuated NoGo-related beta desynchronisation and synchronisation in Parkinson's disease revealed by magnetoencephalographic recording, Sci. Rep. 9 (1) (2019) 7235.
[11] J. Hong, J. Hwang, J.-H. Lee, General psychopathology factor (p-factor) prediction using resting-state functional connectivity and a scanner-generalization neural network, J. Psychiatr. Res. 158 (2023) 114–125.
[12] S. Cervenka, A. Frick, R. Bodén, M. Lubberink, Application of positron emission tomography in psychiatry—methodological developments and future directions, Transl. Psychiatry 12 (1) (2022) 248.
[13] J. Gan, W. Liu, J. Fan, J. Yi, C. Tan, X. Zhu, Correlates of poor insight: a comparative fMRI and sMRI study in obsessive-compulsive disorder and schizo-obsessive disorder, J. Affect. Disord. 321 (2023) 66–73.
[14] P.D. Metzak, M.K. Shakeel, X. Long, M. Lasby, R. Souza, S. Bray, B.I. Goldstein, G. MacQueen, J. Wang, S.H. Kennedy, et al., Brain connectomes in youth at risk for serious mental illness: an exploratory analysis, BMC Psychiatry 22 (1) (2022) 611.
[15] R. Acharya, S. Kafle, D.B. Shrestha, Y.R. Sedhai, M. Ghimire, K. Khanal, Q.B. Malla, U. Nepal, R. Shrestha, B. Giri, Use of computed tomography of the head in patients with acute atraumatic altered mental status: a systematic review and meta-analysis, JAMA Netw. Open 5 (11) (2022) e2242805.
[16] S.K. Khare, S. March, P.D. Barua, V.M. Gadre, U.R. Acharya, Application of data fusion for automated detection of children with developmental and mental disorders: a systematic review of the last decade, Inf. Fusion (2023) 101898.
[17] N.S. Amer, S.B. Belhaouari, EEG signal processing for medical diagnosis, healthcare, and monitoring: a comprehensive review, IEEE Access 11 (2023) 143116–143142, https://doi.org/10.1109/ACCESS.2023.3341419
[18] Z. Ma, A. Li, J. Tang, J. Zhang, Z. Yin, Multimodal emotion recognition by fusing complementary patterns from central to peripheral neurophysiological signals across feature domains, Eng. Appl. Artif. Intell. 143 (2025) 110004.
[19] B.F.O. Coelho, A.B.R. Massaranduba, C.A. dos Santos Souza, G.G. Viana, I. Brys, R.P. Ramos, Parkinson's disease effective biomarkers based on hjorth features improved by machine learning, Expert Syst. Appl. 212 (2023) 118772.
[20] W. Zeng, M. Zhang, L. Shan, Y. Chen, Z. Li, S. Du, Interpretable classification of epileptic EEG signals using ALIF decomposition and attention-augmented cascaded deep neural networks, Appl. Soft Comput. (2025) 113211.
[21] M. Bashir, S. El-Sappagh, S.S. Woo, D.I. Kim, T. Abuhmed, Toward trustworthy digital healthcare: a system-level convergence of IoMT, large language models, and explainable AI, Inf. Fusion (2026) 104507, https://doi.org/10.1016/j.inffus.2026.104507, https://www.sciencedirect.com/science/article/pii/S1566253526003866.
[22] E. Lionakis, K. Karampidis, G. Papadourakis, Current trends, challenges, and future research directions of hybrid and deep learning techniques for motor imagery brain–computer interface, Multimodal Technol. Interact. 7 (10) (2023) 95.

[23] V. Bairagi, S. Kulkarni, A novel method for stress measuring using EEG signals, in: Advances in Information and Communication Networks: Proceedings of the 2018 Future of Information and Communication Conference (FICC), Vol. 2, Springer, 2019, pp. 671–684.

[24] S. Ventura, S.R. Mathieson, J.M. O'Toole, V. Livingstone, D.M. Murray, G.B. Boylan, Infant sleep EEG features at 4 months as biomarkers of neurodevelopment at 18 months, Pediatr. Res. (2025) 1–12.

[25] M. Beaugrand, V. Jaramillo, C. Mühlematter, S.F. Schoch, V. Reicher, A. Markovic, S. Kurth, Tracing infant sleep neurophysiology longitudinally from 3 to 6 months: EEG insights into brain development, npj Biol. Timing Sleep 3 (1) (2026) 9.

[26] M.I.B. Ahmed, S. Alotaibi, S. Dash, M. Nabil, A.O. AlTurki, A review on machine learning approaches in identification of pediatric epilepsy, SN Comput. Sci. 3 (6) (2022) 437.

[27] S.M. Alotaibi, M.I. Basheer, M.A. Khan, et al., Ensemble machine learning based identification of pediatric epilepsy, Comput. Mater. Contin. 68 (1) (2021).

[28] C.M. Sharma, V.M. Chariar, Diagnosis of mental disorders using machine learning: literature review and bibliometric mapping from 2012 to 2023, Heliyon 10 (12) (2024) e32548, https://doi.org/10.1016/j.heliyon.2024.e32548

[29] Y. Wang, R. Elrawas, A.-D. Nguyen, M. Girard, P. Mozharovskyi, E. Tartaglione, V.-T. Nguyen, Unified pipeline for generalized mental state detection using EEG signals, Expert Syst. Appl. (2025) 129422.

[30] J. Liu, G. Wu, Q. Fu, Y. Luo, S. Yang, S. Qiu, Y. Cao, W. Li, EEG-based emotion detection using long short-term memory network and reinforcement learning for enhanced feature selection, Appl. Soft Comput. 182 (2025) 113512.

[31] J. Henry, H. Lloyd, M. Turner, C. Kendrick, On the robustness of machine learning models for stress and anxiety recognition from heart activity signals, IEEE Sens. J. 23 (13) (2023) 14428–14436.

[32] A. Singh, D. Kumar, Detection of stress, anxiety and depression (SAD) in video surveillance using ResNet-101, Microprocess. Microsyst. 95 (2022) 104681.

[33] M. Gavrilescu, N. Vizireanu, Predicting depression, anxiety, and stress levels from videos using the facial action coding system, Sensors 19 (17) (2019) 3693.

[34] M. Tasnim, R.D. Ramos, E. Stroulia, L.A. Trejo, A machine-learning model for detecting depression, anxiety, and stress from speech, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 7085–7089.

[35] A. Fatima, Y. Li, T.T. Hills, M. Stella, Dasentimental: detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning, Big Data Cogn. Comput. 5 (4) (2021) 77.

[36] A. Priya, S. Garg, N.P. Tigga, Predicting anxiety, depression and stress in modern life using machine learning algorithms, Procedia Comput. Sci. 167 (2020) 1258–1267.

[37] N. Rahim, S. El-Sappagh, H. Rizk, O.A. El-Serafy, T. Abuhmed, Information fusion-based Bayesian optimized heterogeneous deep ensemble model based on longitudinal neuroimaging data, Appl. Soft Comput. 162 (2024) 111749.

[38] M. Bashir, N. Rahim, S. El-Sappagh, O.A. El-Serafy, T. Abuhmed, Trustworthy Alzheimer's diagnosis: integrating robustness, fairness, and explainability in neuroimaging based deep ensemble framework, Eng. Appl. Artif. Intell. 172 (2026) 114291.

[39] S. Vieluf, T. Hasija, M. Kuschel, C. Reinsberger, T. Loddenkemper, Developing a deep canonical correlation-based technique for seizure prediction, Expert Syst. Appl. 234 (2023) 120986.

[40] R. Catherine Joy, S. Thomas George, A. Albert Rajan, M. Subathra, Detection of ADHD from EEG signals using different entropy measures and ANN, Clin. EEG Neurosci. 53 (1) (2022) 12–23.

[41] M. Aldayel, A. Al-Nafjan, A comprehensive exploration of machine learning techniques for EEG-based anxiety detection, Peerj Comput. Sci. 10 (2024) e1829.

[42] Z. Shen, G. Li, J. Fang, H. Zhong, J. Wang, Y. Sun, X. Shen, Aberrated multidimensional EEG characteristics in patients with generalized anxiety disorder: a machine-learning based analysis framework, Sensors 22 (14) (2022) 5420.

[43] N. Sakib, K. Islam, T. Faruk, Effect of artifact removal in machine learning based anxiety screening using EEG signal, in: 2025 4th International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), IEEE, 2025, pp. 498–503.

[44] Z. Li, X. Wu, X. Xu, H. Wang, Z. Guo, Z. Zhan, L. Yao, The recognition of multiple anxiety levels based on electroencephalograph, IEEE Trans. Affect. Comput. 13 (1) (2019) 519–529.

[45] X. Luo, B. Zhou, J. Fang, Y. Cherif-Riahi, G. Li, X. Shen, Integrating EEG and ensemble learning for accurate grading and quantification of generalized anxiety disorder: a novel diagnostic approach, Diagnostics 14 (11) (2024) 1122.

[46] A. Jain, R. Kumar, Machine learning based anxiety detection using physiological signals and context features, in: 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), IEEE, 2024, pp. 116–121.

[47] A. Setiawan, M.A.S. Noor, A. Trisanto, G.F. Yohanes, D.D. Rumani, B.M. Wibawa, N. Rohadi, A. Turnip, et al., Pilot anxiety detection through brain signal using naïve Bayes method, in: 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), IEEE, 2024, pp. 1–6.

[48] H. Jebelli, S. Hwang, S. Lee, EEG-based workers' stress recognition at construction sites, Autom. Constr. 93 (2018) 315–324.

[49] S.M. Umar Saeed, S.M. Anwar, M. Majid, M. Awais, M. Alnowami, Selection of neural oscillatory features for human stress classification with single channel EEG headset, Biomed Res. Int. 2018 (1) (2018) 1049257.

[50] J. Minguillon, E. Perez, M.A. Lopez-Gordo, F. Pelayo, M.J. Sanchez-Carrion, Portable system for real-time detection of stress level, Sensors 18 (8) (2018) 2504.

[51] W.K. So, S.W. Wong, J.N. Mak, R.H. Chan, An evaluation of mental workload with frontal EEG, PLOS One 12 (4) (2017) e0174949.

[52] V. Rajendran, S. Jayalalitha, K. Adalarasu, R. Mathi, Machine learning based human mental state classification using wavelet packet decomposition-an EEG study, Multimed. Tools Appl. (2024) 1–20.

[53] U. Ince, Y. Talu, A. Duz, S. Tas, D. Tanko, I. Tasci, S. Dogan, A. Hafeez Baig, E. Aydemir, T. Tuncer, CubicPat: investigations on the mental performance and stress detection using EEG signals, Diagnostics 15 (3) (2025) 363.

[54] H. Lee, B. Li, S. DeForte, M.L. Splaingard, Y. Huang, Y. Chi, S.L. Linwood, A large collection of real-world pediatric sleep studies, Sci. Data 9 (1) (2022) 421.

[55] L.D. Sharma, V.K. Bohat, M. Habib, A.-Z. Ala'M, H. Faris, I. Aljarah, Evolutionary inspired approach for mental stress detection using EEG signal, Expert Syst. Appl. 197 (2022) 116634.

[56] H. Lee, A. Saeed, Pediatric sleep scoring in-the-wild from millions of multi-channel EEG signals, arXiv preprint arXiv:2207.06921, 2022.

[57] R.W. Homan, J. Herman, P. Purdy, Cerebral location of international 10–20 system electrode placement, Electroencephalogr. Clin. Neurophysiol. 66 (4) (1987) 376–382.

[58] J.N. Acharya, V.J. Acharya, Overview of EEG montages and principles of localization, J. Clin. Neurophysiol. 36 (5) (2019) 325–329.

[59] L. Esch, C. Dinh, E. Larson, D. Engemann, M. Jas, S. Khan, A. Gramfort, M.S. Hämäläinen, MNE: software for acquiring, processing, and visualizing MEG/EEG data, in: Magnetoencephalography: From Signals to Dynamic Cortical Networks, Springer Nature, 2019, pp. 355–371.

[60] A. Gramfort, M. Luessi, E. Larson, D.A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al., MEG and EEG data analysis with MNE-Python, Front. Neurosci. 7 (2013) 70133.

[61] N.M. Shafiq Surameery, A. Alazzawi, A.T. Asaad, Intelligent approaches for Alzheimer's disease diagnosis from EEG signals: systematic review, Int. J. Comput. Digit. Syst. 16 (1) (2024) 1–36.

[62] H. Nolan, R. Whelan, R.B. Reilly, FASTER: fully automated statistical thresholding for EEG artifact rejection, J. Neurosci. Methods 192 (1) (2010) 152–162.

[63] M.M. Ahsan, M.P. Mahmud, P.K. Saha, K.D. Gupta, Z. Siddique, Effect of data scaling methods on machine learning algorithms and model performance, Technologies 9 (3) (2021) 52.

[64] R. Islam, A. Layek, Stackensemblemind: enhancing well-being through accurate identification of human mental states using stack-based ensemble machine learning, Inform. Med. Unlocked (2023) 101405.

[65] S.K. Satapathy, D. Loganathan, Prognosis of automated sleep staging based on two-layer ensemble learning stacking model using single-channel EEG signal, Soft Comput. 25 (24) (2021) 15445–15462.

[66] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 25 (2012).

[67] S.H. El-Sappagh, W. Nazih, M. Alharbi, T. Abuhmed, Responsible artificial intelligence for mental health disorders: current applications and future challenges, J. Disabil. Res. 4 (1) (2025) 20240101.

[68] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[69] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[70] C.-H. Lee, D. Ahn, H. Kim, E.J. Ha, J.-B. Kim, D.-J. Kim, NeuroXAI: adaptive, robust, explainable surrogate framework for determination of channel importance in EEG application, Expert Syst. Appl. 261 (2025) 125364.

[71] A. Tripathi, W. Ganglberger, H. Sun, C. Alcott, N. Turley, R. Fitzgerald, A. Mitra, S. Waters, A. Gupta, A. Gupta, et al., The Boston children's hospital sleep corpus: a collection of 15695 annotated pediatric polysomnograms, SLEEPJ (2025) zsaf273.

[72] A. Nassibi, C. Papavassiliou, S.F. Atashzar, Depression diagnosis using machine intelligence based on spatiospectrotemporal analysis of multi-channel EEG, Med. Biol. Eng. Comput. 60 (11) (2022) 3187–3202.

[73] J. Li, G. Feng, J. Lv, Y. Chen, R. Chen, F. Chen, S. Zhang, M.-I. Vai, S.-H. Pun, P.-U. Mak, A lightweight multi-mental disorders detection method using entropy-based matrix from single-channel EEG signals, Brain Sci. 14 (10) (2024) 987.

[74] A.J. Marthinsen, I. Galtung, A. Cheema, C. Sletten, I.M. Andreassen, Ø. Sletta, A. Soler, M.M. Molinas Cabrera, Psychological stress detection with optimally selected EEG channel using machine learning techniques, CEUR Workshop Proc. 3576 (2023).

[75] A. Baghdadi, Y. Aribi, R. Fourati, N. Halouani, P. Siarry, A.M. Alimi, Dasps: a database for anxious states based on a psychological stimulation, arXiv preprint arXiv:1901.02942, 2019.