Research Article
Journal of Intelligent Systems: Theory and Applications 9 (2026) 1901082
DOI: 10.38016/jista.1901082

EEG Tabanlı Yalan Algılamada Dalgacık Ayrıştırma ve Öğrenme Mimarileri: LieWaves Veri Kümesi Üzerinde DWT Özellik Temsillerinin LOSOCV Tabanlı Karşılaştırmalı Değerlendirmesi

Rabia Cansel Uğur[1], Uğur Yüzgeç[2,*]

[1,2] Department of Computer Engineering, Bilecik Şeyh Edebali University, Bilecik, Türkiye

r4bi4cnsl@gmail.com, ugur.yuzgec@bilecik.edu.tr

Öz
Bu çalışma, kamuya açık LieWaves veri setini kullanarak denek bağımsız EEG tabanlı yalan tespiti için Ayrık Dalgacık Dönüşümü (DWT) tabanlı özellik temsillerini ve makine/derin öğrenme sınıflandırıcılarını sistematik olarak değerlendirmektedir. EEG sinyalleri bant geçiren filtreleme ve Otomatik ve Ayarlanabilir Artefakt Giderme (ATAR) işlemlerinden geçirilmiş, Örtüşen Kayar Pencere (OSW) yöntemiyle bölütlenmiş ve DWT ile ayrıştırılmıştır. Dört DWT konfigürasyonu (db4-Level4, db4-Level5, db5-Level4 ve db5-Level5), dış Bir Denek Dışarıda Bırakma Çapraz Doğrulama (LOSOCV) çerçevesinde üç denekli, denek bazlı iç holdout doğrulaması kullanılarak karşılaştırılmış; db5-Level5 en yüksek ortalama iç doğrulama performansına ulaşmıştır. Sabit db5-Level5 temsili kullanılarak DWT tabanlı istatistiksel özellikler EEGNet'ten esinlenen kompakt konvolüsyonel model, CNN, LSTM, BiLSTM, GRU, Rastgele Orman, KNN ve SVM ile LOSOCV altında değerlendirilmiştir. EEGNet'ten esinlenen model 0,5779±0,0903 doğruluk ve 0,5295±0,1359 makro F1 puanı ile en yüksek ortalama performansı elde etmiştir. Bununla birlikte, tüm modellerin denek bağımsız performansı sınırlı kalmıştır. Sonuçlar, yüksek ölçüde örtüşen EEG bölütlerinde denek bağımsız doğrulamanın önemini göstermekte ve önerilen iş akışının kullanıma hazır pratik/adli bir sistemden ziyade keşif amaçlı karşılaştırmalı bir çerçeve olarak değerlendirilmesi gerektiğini ortaya koymaktadır.

Anahtar kelimeler: EEG, Yalan Tespiti, Makine Öğrenmesi, Derin Öğrenme, Ayrık Dalgacık Dönüşümü.

Wavelet Decomposition and Learning Architectures for EEG-Based Lie Detection: A LOSOCV-Based Comparative Evaluation of DWT Feature Representations on the LieWaves Dataset

Abstract
This study systematically evaluates discrete wavelet transform (DWT)-based EEG feature representations and machine/deep learning classifiers for subject-independent deception detection using the publicly available LieWaves dataset. EEG signals were bandpass filtered, processed with Automatic and Tunable Artifact Removal (ATAR), segmented with the Overlapping Sliding Window (OSW) method, and decomposed using DWT. Four configurations (db4-Level4, db4-Level5, db5-Level4, and db5-Level5) were compared within an outer Leave-One-Subject-Out Cross-Validation (LOSOCV) framework using a three-subject, subject-wise inner holdout procedure; db5-Level5 achieved the highest mean inner-validation performance. Using the fixed db5-Level5 representation, DWT-derived statistical features were evaluated with an EEGNet-inspired compact convolutional model, CNN, LSTM, BiLSTM, GRU, Random Forest, KNN, and SVM under LOSOCV. The EEGNet-inspired model achieved the highest average performance, with an accuracy of 0.5779±0.0903 and a macro F1-score of 0.5295±0.1359. Nevertheless, all models showed limited subject-independent performance. These results highlight the importance of subject-independent validation for highly overlapping EEG segments and indicate that the proposed pipeline should be regarded as an exploratory comparative framework rather than a ready-to-use practical or forensic lie detection system.

Keywords: EEG, Lie Detection, Machine Learning, Deep Learning, Discrete Wavelet Transform.

* Corresponding Author.
E-mail: ugur.yuzgec@bilecik.edu.tr

Recieved : 2 Mar 2026
Revision : 13 Apr 2026
Accepted : 10 Aug 2026
Published Online : 15 Sep 2026

1. Introduction

Deception is a complex cognitive behavior that individuals may deliberately employ to conceal information, avoid responsibility, or manipulate social and legal outcomes. Accurate lie detection plays a critical role in various domains, including forensic investigations, security screening, psychological assessment, and human-computer interaction (Chanda & Mandal, 2023). Traditional lie detection approaches largely rely on behavioral cues such as facial expressions, body language, eye movements, voice modulation, and physiological signals like heart rate or skin conductance. However, these indicators can often be consciously controlled or masked, particularly by trained or emotionally regulated individuals, which significantly limits the reliability of such methods (Joshi et al., 2025; Uotinen et al., 2024).

In recent years, electroencephalography (EEG) has emerged as a promising modality for deception detection due to its ability to directly capture neural activity associated with cognitive processes (Taha et al., 2025). EEG signals reflect real-time brain dynamics and are less susceptible to voluntary manipulation compared to external behavioral indicators. Advances in portable EEG devices and signal processing techniques have further facilitated the practical use of EEG-based systems in experimental and applied settings. Consequently, leveraging EEG signals for automated lie detection has attracted increasing attention within the machine learning and signal processing communities (Nagale & Khandare, 2025).

Despite these advances, EEG-based deception detection remains a challenging task due to the non-stationary, noisy, and high-dimensional nature of EEG signals. Effective preprocessing, feature extraction, and classification strategies are therefore essential to achieve reliable and generalizable performance. In this context, time-frequency analysis methods and data-driven learning models have shown considerable potential for capturing subtle neural patterns related to deceptive behavior (Bhatt et al., 2023).

The primary challenge in EEG-based lie detection lies in accurately distinguishing deceptive and truthful cognitive states from complex and artifact-prone EEG recordings. Raw EEG signals are typically contaminated by physiological and environmental artifacts such as eye blinks, muscle movements, and external noise, which can obscure meaningful neural information. Furthermore, the limited size of available deception-related EEG datasets often restricts the learning capacity and generalization performance of data-hungry deep learning models (Raj et al., 2025).

Another critical issue is the effective representation of EEG signals. While time-domain features may capture global signal characteristics, they often fail to represent localized frequency-specific patterns that are closely associated with cognitive load and deception-related brain activity. Therefore, robust time-frequency decomposition methods are required to extract discriminative features from EEG signals (Amin et al., 2017). Additionally, there is a lack of comprehensive comparative studies that systematically evaluate both deep learning and traditional machine learning models under a unified preprocessing and feature extraction framework for EEG-based lie detection (Tarawneh et al., 2025).

Addressing these challenges requires an integrated approach that combines artifact removal, overlapping window-based EEG segmentation, advanced signal decomposition, rigorous validation, and comparative model evaluation to better assess the generalizability of EEG-based deception detection models.

It should be emphasized that the aim of this study is not to develop a deployable lie detection system. Instead, the study aims to examine how DWT-based EEG feature representations and different learning models behave under a subject-independent LOSOCV protocol. Therefore, the contribution of the work is methodological and comparative rather than practical or forensic deployment-oriented. The main contributions of this study can be summarized as follows:

- A unified comparative evaluation framework is constructed for EEG-based deception detection, integrating artifact removal, overlapping window-based segmentation, DWT-based feature extraction, classification, and subject-independent LOSOCV evaluation within the same experimental pipeline.
- EEG signals from the LieWaves dataset are preprocessed using Artifact Removal (ATAR) and segmented via the Overlapping Sliding Window (OSW) technique to capture localized temporal dynamics from continuous EEG recordings.
- Discrete Wavelet Transform (DWT) is employed to decompose EEG signals into multilevel frequency components, and different mother wavelet/decomposition-level configurations are experimentally compared to identify the most effective representation among the tested settings.
- An ablation analysis is conducted to evaluate the effects of the mother wavelet and decomposition depth by comparing db4-Level4, db4-Level5, db5-Level4, and db5-Level5 DWT configurations.
- A wide range of deep learning and traditional machine learning models, including EEGNet, CNN, LSTM, BiLSTM, GRU, Random Forest, KNN, and SVM, are systematically evaluated under the same preprocessing and feature extraction pipeline using a subject-independent LOSOCV protocol.
- Experimental results show that db5-Level5 achieves the highest mean inner-validation performance among the tested DWT configurations in the ablation analysis conducted

within an outer LOSOCV framework using three-subject, subject-wise inner holdout validation, while EEGNet provides the highest average subject-independent performance among the evaluated classifiers under LOSOCV, although the overall performance remains limited.
- The study provides an empirical analysis of the strengths and limitations of different classifiers under subject-independent evaluation. Its main contribution lies not in developing a ready-to-use lie detection system, but in systematically evaluating DWT-based EEG representations and multiple classifiers within a unified LOSOCV-based comparative framework on the LieWaves dataset.

2. Related Work

Lie detection research has traditionally relied on behavioral and physiological indicators. Early approaches primarily examined observable cues such as facial expressions, body language, and speech patterns. However, these indicators are often subject to voluntary control and contextual variability, which may reduce reliability in real-world scenarios (Ekman, 2009; Vrij, 2008). Comprehensive reviews on verbal and nonverbal deception further emphasize that behavioral cues alone rarely provide consistently high detection accuracy due to individual differences and situational factors (Vrij et al., 2010).

Beyond behavioral observations, neurophysiological methods have gained attention due to their ability to capture cognitive processes directly associated with deception. One of the earliest and most influential EEG-based approaches is the use of event-related potentials (ERPs), particularly the P300 component, for concealed information detection (Farwell & Donchin, 1991). Subsequent research demonstrated both the effectiveness and limitations of P300-based lie detection paradigms, including the possibility of countermeasures affecting performance (Rosenfeld et al., 2004). These studies established EEG as a promising modality for objective deception assessment.

With the rapid advancement of computational intelligence, machine learning and deep learning techniques have increasingly been integrated into deception detection systems. In multimodal settings, Pérez-Rosas et al. (2015) introduced deception detection using real-life trial data, combining visual, acoustic, and textual features. Extending this line of work, Şen et al. (2022) applied multimodal deep learning approaches to the same real-life trial dataset and demonstrated improved performance under more comprehensive modeling strategies. Although multimodal systems enhance classification accuracy, they often require complex data acquisition setups and synchronization across heterogeneous modalities.

In the context of EEG analysis, wavelet-based feature extraction has been widely adopted for handling the non-stationary nature of brain signals. Subasi (2007) demonstrated that wavelet-based features combined with expert models can effectively classify EEG signals, highlighting the strength of time-frequency decomposition techniques. More recently, deep learning approaches have significantly improved EEG classification performance. Acharya et al. (2018) showed that convolutional neural networks (CNNs) can automatically learn discriminative representations from raw EEG signals, eliminating the need for extensive handcrafted feature engineering. Furthermore, Roy et al. (2019) provided a comprehensive systematic review of deep learning methods for EEG analysis, highlighting the potential of deep architectures to extract complex neural patterns across various cognitive tasks.

In addition to general-purpose CNN architectures, EEG-specific deep learning models have also been developed to improve classification performance with compact and interpretable structures. Lawhern et al. (2018) proposed EEGNet, a compact convolutional neural network designed for EEG-based brain-computer interface applications. EEGNet employs depthwise and separable convolutional operations to learn temporal and spatial EEG features with a relatively small number of parameters. Due to its compact structure and EEG-oriented design, EEGNet has become a widely used benchmark architecture in EEG classification studies.

Specifically in EEG-based deception detection, recent studies have proposed advanced hybrid and graph-based deep learning frameworks. Aslan et al. (2024b) introduced the LSTMNCP model, combining recurrent and convolutional structures to enhance lie detection performance using EEG signals. Similarly, Rahmani et al. (2024) proposed a novel framework integrating type-2 fuzzy sets with deep graph convolutional networks for automatic lie detection, demonstrating the potential of advanced neural architectures in modeling deception-related brain dynamics.

Despite these advances, prior studies often focus on a limited set of models or evaluate architectures within isolated experimental settings. In contrast, the present study aims to provide a systematic and comparative evaluation of EEG-specific, general deep learning, and conventional machine learning models within a unified preprocessing and feature extraction pipeline. By integrating artifact removal, overlapping sliding window-based segmentation, multilevel wavelet-based feature extraction, DWT configuration ablation, and subject-independent LOSOCV evaluation, this work seeks to offer a clearer understanding of the strengths and limitations of different classification strategies for EEG-based lie detection.

3. Materials and Methods

3.1. Dataset

In this study, the LieWaves dataset, a publicly available EEG dataset specifically developed for deception detection research, was utilized. The dataset was originally introduced by Aslan, Baykara, and Alakus (2024a) and provides EEG recordings acquired under controlled experimental conditions for the purpose of distinguishing truthful and deceptive cognitive states.

The LieWaves dataset consists of EEG signals obtained from 27 voluntary participants, each of whom participated in two separate experimental sessions. In these sessions, participants alternated between truthful and deceptive roles, ensuring that EEG data corresponding to both conditions were available for each subject. This experimental design reduces inter-subject variability and enables a balanced comparison between truthful and deceptive brain responses.

EEG signals in the dataset were acquired using the Emotiv Insight portable EEG device, which employs five electrodes positioned over frontal, temporal, and parietal scalp regions. The EEG signals in the LieWaves dataset were recorded at a sampling frequency of 128 Hz. Therefore, the Nyquist frequency of the recorded EEG signals was 64 Hz. The spatial configuration of the electrodes is illustrated in Figure 1.

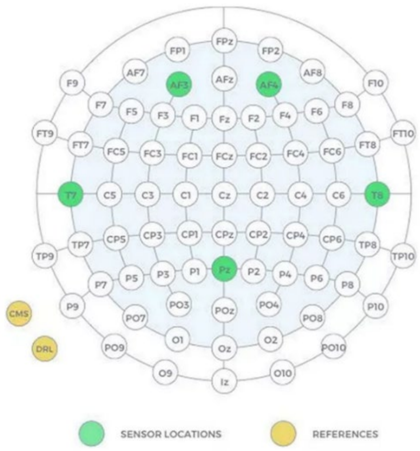


Figure 1. Electrode configuration of the Emotiv Insight EEG device (Aslan, Baykara, & Alakus, 2024a)

During each experimental session, participants were exposed to a sequence of visual stimuli consisting of bead images, preceded by a one-second black screen and followed by a two-second stimulus presentation. Each session included 25 stimuli, resulting in 75 seconds of EEG data per session.

Participants' responses were labeled according to predefined experimental protocols that distinguished truthful and deceptive answers. These annotations, provided as part of the original dataset, enable supervised learning approaches for EEG-based lie detection without requiring additional data collection.

3.2. Preprocessing

EEG signals are inherently susceptible to noise and artifacts arising from eye blinks, muscle movements, electrode movements, and environmental interference. To enhance signal quality and extract meaningful neural patterns, a multi-stage preprocessing pipeline was applied. The pipeline consisted of bandpass filtering, ATAR-based artifact removal, OSW-based segmentation, and DWT-based time-frequency decomposition.

3.2.1. Artifact Removal (ATAR)

Automatic and Tunable Artifact Removal (ATAR) was applied to reduce non-neural components from the EEG signals before the OSW-based segmentation stage. EEG signals were first filtered using a 0.5-45 Hz bandpass filter, and the ATAR algorithm was then applied to suppress artifacts caused by eye movements, muscle activity, and environmental noise sources.

ATAR is based on Wavelet Packet Decomposition (WPD) and removes artifact-related wavelet packet components according to a predefined thresholding strategy. Specifically, ATAR was applied with its default settings, using a threshold value of 0.1 and soft thresholding as the wavelet filter (Aslan, Baykara, & Alakus, 2024a). The same artifact removal procedure was applied consistently to all subjects, sessions, and EEG channels. The ATAR-preprocessed EEG signals were subsequently used as the input for OSW-based segmentation and DWT-based feature extraction. The effectiveness of the ATAR procedure in suppressing non-neural artifacts is illustrated in Figure 2, which shows EEG signals before and after artifact removal.

3.2.2. Overlapping Sliding Window (OSW)

Following artifact removal, the Overlapping Sliding Window (OSW) method was used to segment the EEG signals into overlapping temporal windows. In this study, OSW was treated as an overlapping segmentation procedure rather than a conventional data augmentation method. Although OSW increases the number of available EEG segments by dividing continuous EEG recordings into partially overlapping windows, it does not generate independent new EEG recordings. Therefore, adjacent OSW segments should be interpreted as partially dependent observations because they share a substantial proportion of signal samples.

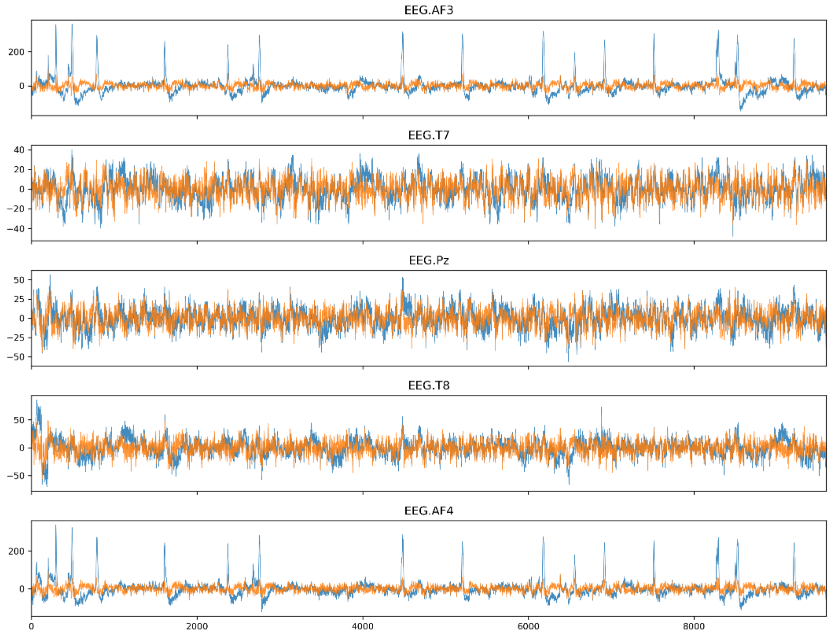


Figure 2. EEG signals before and after ATAR for Subject 1 in the first experimental session (S1S1) across five channels (orange: before ATAR, blue: after ATAR).

Each EEG recording consisted of 9600 samples. A window length of 384 samples and a step size of 32 samples were used in the OSW procedure. Accordingly, the overlap between two consecutive windows was 352 samples, corresponding to an overlap ratio of approximately 91.67%, calculated as (384-32)/384. With these parameters, 289 overlapping windows were generated from each EEG recording.

Since each subject in the LieWaves dataset has two recording sessions, the number of windows obtained from one subject was 289×2=578. For the complete dataset consisting of 27 subjects, the total number of generated windows was 289×2×27=15,606. The main parameters used in the OSW-based segmentation procedure, including the window length, step size, overlap ratio, and the number of generated windows are summarized in Table 1. Figure 3 shows the schematic representation of the OSW-based segmentation procedure.

Table 1. Parameters of the overlapping sliding window-based segmentation procedure

Parameter	**Value**
Recording length	9600 samples
Window length	384 samples
Step size	32 samples
Overlap length	352 samples
Overlap ratio	91.67%
Number of windows per recording	289
Number of sessions per subject	2
Number of windows per subject	578
Number of subjects	27
Total number of windows	15,606

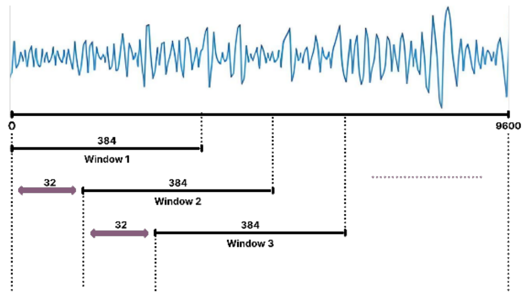


Figure 3. Schematic representation of the OSW method

3.2.3. Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) is a widely used time-frequency analysis technique that enables the decomposition of non-stationary signals into localized frequency components. Unlike classical Fourier-based methods, DWT provides simultaneous temporal and spectral resolution, making it particularly suitable for analyzing EEG signals, which exhibit transient and time-varying characteristics.

Let $x[n]$ denote a discrete EEG segment obtained from a specific window and channel. In a multiresolution DWT representation, the signal is decomposed into approximation and detail components across different resolution levels. Let $A_j(k)$ denote the approximation coefficients and $D_j(k)$ denote the detail coefficients at decomposition level j, where k is the coefficient index. The scaling and wavelet basis functions at level j and shift k are denoted by $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$, respectively. The multiresolution representation of the signal can be expressed as:

$$x(t)=\sum_{k} A_j(k)\,\phi_{j,k}(t)+\sum_{j}\sum_{k} D_j(k)\,\psi_{j,k}(t) \tag{1}$$

where j denotes the maximum decomposition level. In practical filter-bank implementation, DWT is computed using a low-pass filter $h[n]$ and a high-pass filter $g[n]$, followed by down-sampling by a factor of two. Starting from $A_0[n]=x[n]$, the approximation and detail coefficients at level j are obtained as:

$$A_j[n]=\sum_{k} x[k]\,h[2n-k] \tag{2}$$

$$D_j[n]=\sum_{k} x[k]\,g[2n-k] \tag{3}$$

where $A_j[n]$ denotes the approximation coefficients, $D_j[n]$ denotes the detail coefficients, $h[n]$ and $g[n]$ denote the low-pass and high-pass decomposition filters, respectively, j denotes the decomposition level, n denotes the down-sampled coefficient index, and k denotes the summation index.

In this study, different Discrete Wavelet Transform (DWT) configurations were evaluated to investigate the effects of the mother wavelet and decomposition depth on EEG-based lie detection performance. Specifically, four DWT configurations were considered: db4-Level4, db4-Level5, db5-Level4, and db5-Level5.

The terms db4 and db5 refer to the fourth and fifth-order Daubechies mother wavelets, respectively, whereas Level-4 and Level-5 denote the decomposition depth of the DWT. A Level-4 decomposition produces one approximation sub-band and four detail sub-bands, namely A4, D4, D3, D2, and D1. Similarly, a Level-5 decomposition produces one approximation sub-band and five detail sub-bands, namely A5, D5, D4, D3, D2, and D1.

Considering the sampling frequency of 128 Hz, the theoretical frequency ranges of the DWT sub-bands for the Level-5 decomposition were defined as follows: D1 corresponds to 32-64 Hz, D2 to 16-32 Hz, D3 to 8-16 Hz, D4 to 4-8 Hz, D5 to 2-4 Hz, and A5 to 0-2 Hz. For the Level-4 decomposition, the corresponding sub-bands were D1: 32-64 Hz, D2: 16-32 Hz, D3: 8-16 Hz, D4: 4-8 Hz, and A4: 0-4 Hz. Since the EEG signals were bandpass-filtered between 0.5 and 45 Hz during preprocessing, the practically retained frequency content was limited to this filtered range.

The use of different Daubechies wavelets and decomposition levels enables the analysis of EEG signals at different time-frequency resolutions. Lower-order Daubechies wavelets such as db4 can capture relatively sharper transient changes, whereas higher-order wavelets such as db5 provide smoother basis functions. Similarly, increasing the decomposition level allows the EEG signal to be represented through a finer set of low and high-frequency sub-bands. Therefore, the comparative evaluation of db4/db5 and Level-4/Level-5 configurations provides a systematic basis for identifying the most effective DWT representation for the LieWaves dataset. The multi-level decomposition process of an EEG signal using low-pass and high-pass filtering followed by down-sampling is schematically illustrated in Figure 4.

The multi-level decomposition enabled the extraction of both low and high-frequency components from EEG signals recorded during deceptive and truthful responses. DWT was applied separately to each windowed segment and each EEG channel, resulting in hierarchical time-frequency representations of EEG activity across multiple sub-bands. These decomposed sub-bands served as the basis for subsequent statistical feature extraction.

3.3. Feature extraction

Following the DWT decomposition, statistical features were extracted from each wavelet sub-band to construct discriminative representations of the EEG signals. Feature extraction was performed for each window, channel, and decomposition level independently. For each DWT sub-band obtained from a given EEG window and channel, let x_i denote the $i-th$ sample of the sub-band signal, where $i=1,2,\ldots,n$, and n denotes the total number of samples in that sub-band. The mean of the sub-band samples is denoted by $\bar{x}$. For median calculation, $x_{(\mathrm{i})}$ denotes the $i-th$ sample after sorting the sub-band samples in ascending order. The minimum value represents the lowest amplitude of the signal within a window and provides information about extreme negative deviations. It is defined as:

$$\min(x)=\min_{1\le i\le n} x_i \tag{4}$$

The maximum value indicates the highest signal amplitude and reflects extreme positive variations in the EEG signal:

$$\max(x)=\max_{1\le i\le n} x_i \tag{5}$$

The median represents the central tendency of the signal and is less sensitive to outliers than the mean. For an ordered sequence $x_1,x_2,\cdots,x_n$, the median is defined as:

$$\operatorname{median}(x)=\begin{cases} x_{((n+1)/2)}, & if\ n\ is\ odd \\ \frac{1}{2}\left(x_{(n/2)}+x_{((n/2)+1)}\right), & if\ n\ is\ even \end{cases} \tag{6}$$

The mean value describes the average amplitude of the signal and is computed as:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

Energy quantifies the signal power within a window and is calculated as the sum of squared amplitudes:

$$E = \sum_{i=1}^{n} {x_i}^2 \tag{8}$$

Standard deviation measures the dispersion of signal values around the mean and reflects signal variability:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9}$$

Variance represents the squared standard deviation and provides another measure of signal spread:

$$\mathrm{var}(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{10}$$

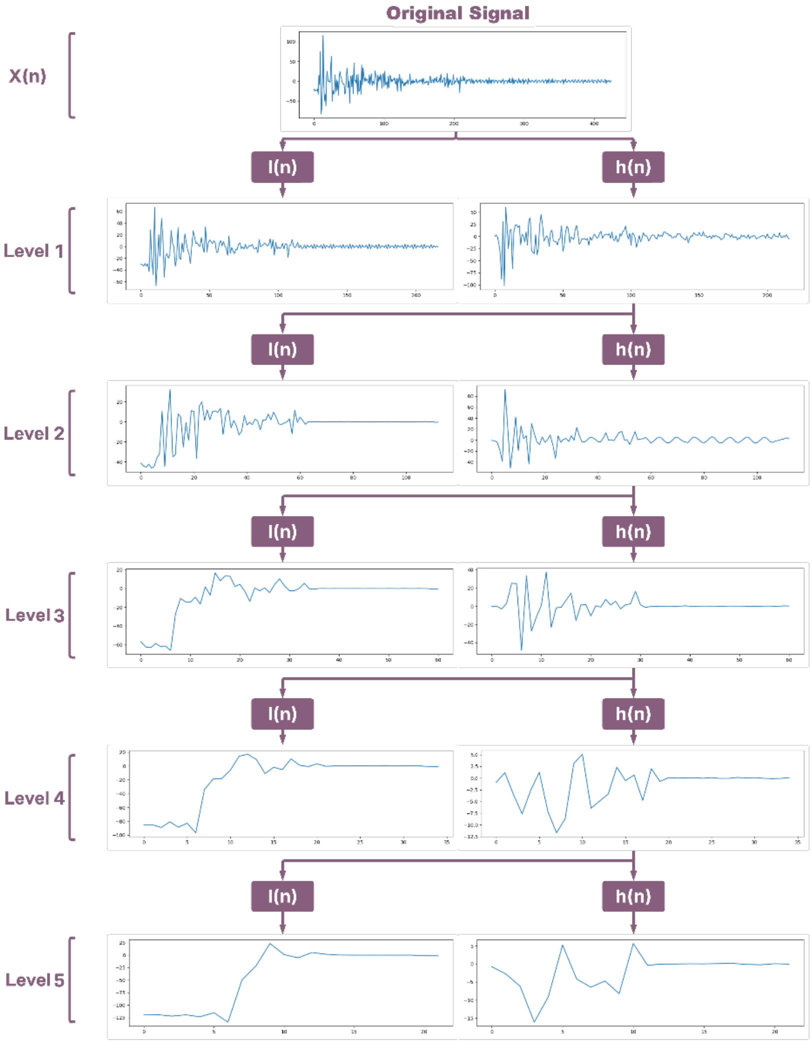


Figure 4. Multi-level DWT decomposition of an EEG signal showing approximation and detail coefficients across successive levels

It should be noted that standard deviation and variance are mathematically related dispersion measures, since variance is the square of the standard deviation. Therefore, they do not represent fully independent statistical descriptors. In this study, both features were retained to preserve the consistency of the predefined feature extraction pipeline and the reported feature dimensionality across all models and DWT configurations. However, their redundancy was considered when interpreting the feature set, and no independent discriminative contribution was attributed separately to these two measures.

Skewness describes the asymmetry of the signal distribution. Positive skewness indicates a right-skewed distribution, while negative skewness indicates a left-skewed distribution:

$$\mathrm{skewness}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3 \tag{11}$$

Skewness was computed only when $\sigma > 0$. In this study, the population-style normalization factor $1/n$ was used for standard deviation and variance calculations. Entropy measures the randomness or complexity of the sub-band signal. Let p_i denote the probability of the $i-th$ amplitude bin or quantized amplitude level, and let B denote total number of non-empty amplitude bins. Entropy is defined as:

$$H = -\sum_{i=1}^{B} p_i \log(p_i) \quad (12)$$

where only non-zero probability terms are included in the summation. Relative energy RE_i represents the proportion of energy contained in the $i-th$ wavelet sub-band relative to the total energy across all sub-bands:

$$RE_i = \frac{E_i}{\sum_{j=1}^{M} E_j} \quad (13)$$

where E_i denotes the energy of the $i-th$ sub-band, E_j denotes the energy of the *j-th* sub-band, and M denotes the total number of sub-bands for the corresponding DWT configuration. For each EEG recording, statistical features were extracted from all wavelet sub-bands obtained through the evaluated DWT configurations across five EEG channels. This process resulted in a high-dimensional feature matrix that captures both temporal and spectral characteristics of the EEG signals. For Level-5 decomposition, six sub-bands were obtained from each EEG channel: A5, D5, D4, D3, D2, and D1. Since ten statistical features were extracted from each sub-band and the LieWaves dataset contains five EEG channels, the final feature vector for each Level-5 window consisted of 5×6×10=300 features. For Level-4 decomposition, five sub-bands were obtained from each EEG channel: A4, D4, D3, D2, and D1. Therefore, the final feature vector for each Level-4 window consisted of 5×5×10=250 features. These feature representations were used consistently in the ablation analysis of different DWT configurations.

Prior to classification, the extracted features were standardized using StandardScaler. The partition used to estimate the scaler parameters depended on the evaluation stage. During DWT-configuration selection, the scaler was fitted using the 23 inner-training subjects and was applied without refitting to the three inner-validation subjects. During the classifier-comparison experiments, StandardScaler was fitted using the complete outer-training partition and was applied without refitting to the held-out outer-test subject. Because the ordered 10% validation subset was defined after standardization of the outer-training partition, these validation samples contributed to the estimation of the scaler parameters; however, the held-out outer-test subject did not.

3.4. Classification models

To evaluate the effectiveness of different learning paradigms for EEG-based lie detection, both deep learning-based and traditional machine learning-based classifiers were employed. The evaluated models included EEGNet, CNN, LSTM, BiLSTM, GRU, RF, KNN, and SVM. This comparative approach enables a comprehensive assessment of each model's capability to capture discriminative patterns from wavelet-based EEG features under identical experimental conditions.

The extracted feature vectors were represented differently depending on the classifier type. It should be noted that the classifiers were not trained directly on raw EEG signals or raw wavelet coefficients. Instead, all models used statistical features extracted from DWT sub-bands. Therefore, in the context of this study, the term EEGNet refers to an EEGNet-inspired compact convolutional model adapted to structured DWT-based statistical feature matrices, rather than the canonical EEGNet architecture operating on raw or minimally processed EEG trials. For traditional machine learning models, including RF, KNN, and SVM, the extracted features were used directly as flat input vectors.

For deep learning models, the feature vectors were reshaped into structured input representations. In the Level-5 configuration, the 300-dimensional feature vector was arranged as a (30,10) representation, where 30 corresponds to the ordered channel-sub-band components obtained from 5 EEG channels and 6 DWT sub-bands, and 10 corresponds to the statistical features extracted from each component. For Level-4 configurations, the corresponding 250-dimensional feature vector was constructed from 5 EEG channels, 5 DWT sub-bands, and 10 statistical features.

For EEGNet and CNN models, the structured feature matrix was used to capture local dependencies among channel, sub-band, and feature dimensions through convolutional operations. For LSTM, BiLSTM, and GRU models, the same representation was interpreted as a sequential input to model dependencies among the ordered channel-sub-band components. For the recurrent models, this sequence should not be interpreted as the original temporal EEG signal. Instead, each sequential step corresponded to an ordered channel-sub-band component represented by its statistical feature vector. Therefore, LSTM, BiLSTM, and GRU modeled dependencies among structured wavelet-derived feature components rather than temporal dependencies directly from raw EEG samples.

3.4.1. EEGNet

EEGNet is a compact convolutional neural network architecture specifically developed for EEG-based classification tasks. Lawhern et al. (2018) introduced EEGNet as an EEG-specific CNN model for brain-computer interface applications and designed it to generalize across different EEG paradigms while using a relatively small number of trainable parameters. The architecture incorporates temporal convolution, depthwise convolution, and separable convolution operations to efficiently learn EEG-related representations.

In this study, EEGNet was included as a contemporary EEG-specific benchmark model to evaluate subject-independent lie detection performance. Unlike its original use on raw EEG trials, the EEGNet model in this study was trained on structured DWT-based statistical feature representations. For the db5-Level5 configuration, each EEG window produced a 300-dimensional feature vector derived from 5 EEG channels, 6 DWT sub-bands, and 10 statistical features. This vector was reshaped into a 5×60×1 input representation, where 5 corresponds to the EEG channels and 60 corresponds to the ordered sub-band-feature components for each channel. The model consisted of temporal convolution, depthwise convolution, separable convolution, average pooling, dropout, and a final softmax output layer for binary classification.

The original EEGNet architecture was designed to learn temporal and spatial filters directly from raw or minimally processed EEG trials. In contrast, the present study employed an EEGNet-inspired compact convolutional model on structured DWT-based statistical feature representations. Therefore, the model input in this study should not be interpreted as raw EEG data. Instead, it corresponds to a wavelet-derived statistical feature matrix constructed from EEG channels, DWT sub-bands, and statistical descriptors. Although the model retains the compact convolutional design principles of EEGNet, its use in this work does not fully conform to the original architectural assumptions of canonical EEGNet. Consequently, the EEGNet results reported in this study should be interpreted as the performance of an EEGNet-inspired compact convolutional model applied to wavelet-derived statistical feature matrices, rather than as a standard EEGNet evaluation on raw EEG signals.

In the experimental tables, this model is denoted as EEGNet for brevity; however, it should be understood as the EEGNet-inspired compact convolutional model described above.

3.4.2. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are widely used for automatic feature learning due to their ability to capture local spatial patterns through convolutional operations. In this study, the CNN architecture was not trained directly on raw EEG signals; instead, it was trained on structured matrices constructed from DWT-based statistical feature representations. Convolutional layers were used to capture local dependencies among channel, sub-band, and feature dimensions, followed by pooling layers to reduce dimensionality and improve generalization. Fully connected layers were then applied to perform the final classification. CNNs are particularly effective in modeling structured feature representations derived from multi-channel EEG signals. Figure 5 shows the CNN model architecture used in this study.

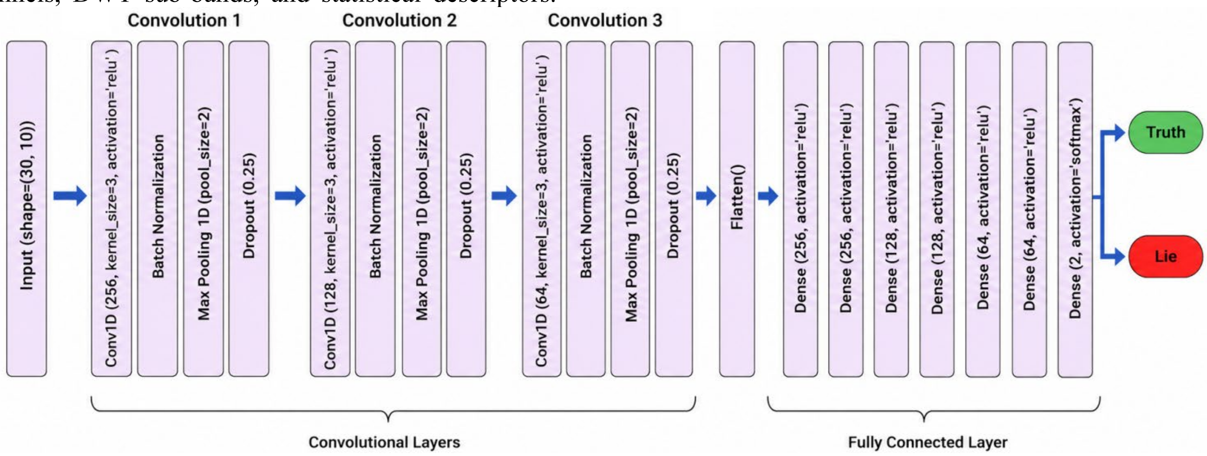


Figure 5. Architecture of the CNN model applied to structured DWT-based statistical feature representations

The CNN architecture consists of three convolutional blocks, each including convolution, batch normalization, ReLU activation, dropout, and max-pooling layers. The convolutional blocks are followed by six fully connected layers, and a final softmax output layer is used for binary classification.

3.4.3. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a specialized form of recurrent neural networks designed to model long-term dependencies in sequential data. LSTM architectures utilize memory cells and gating mechanisms to mitigate the vanishing gradient problem encountered in traditional RNNs. In this study, the LSTM model was employed to capture dependencies within the sequential representation of DWT-based statistical EEG features, rather than directly processing raw EEG signals. Thus, the sequence dimension of the LSTM input represents the ordered channel-sub-band structure of the DWT-derived feature matrix, not the original EEG time axis. Figure 6 shows the architecture of the LSTM model used in this study.

The LSTM-based model takes an input sequence of size (30, 10) and consists of two stacked LSTM layers with 512 and 128 hidden units, respectively, both configured with return sequences enabled. Each LSTM layer employs dropout and recurrent dropout to improve generalization. The learned sequential feature representations are passed to a fully connected layer with ReLU activation, followed by a dropout layer and a softmax output layer for binary classification.

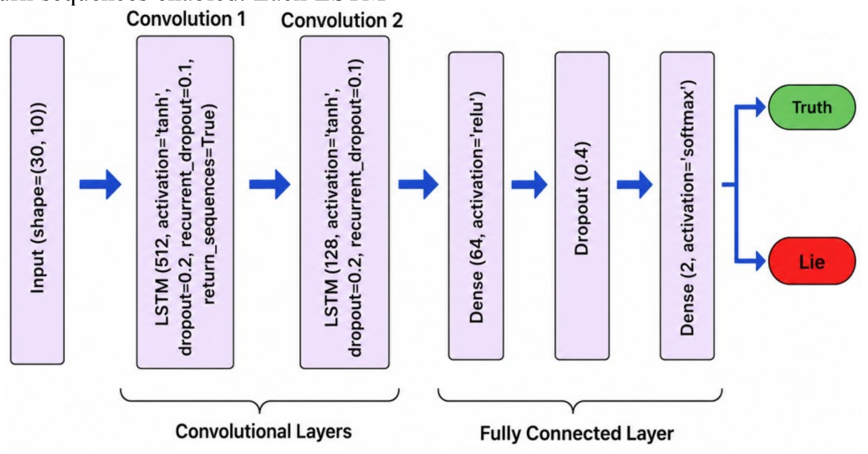


Figure 6. Architecture of the LSTM model applied to the ordered channel-sub-band-feature representation

3.4.4. Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional LSTM (BiLSTM) networks extend the standard LSTM architecture by processing input sequences in both forward and backward directions. In this study, this bidirectional processing was applied to the ordered channel-sub-band-feature sequence, not to the original temporal EEG signal. Therefore, the forward and backward directions should be interpreted as opposite traversal directions over the structured DWT-derived feature representation rather than over the raw EEG time axis. In the context of EEG-based lie detection, the BiLSTM model was used to capture bidirectional dependencies within the sequential representation of DWT-based statistical EEG features. Figure 7 shows the BiLSTM model used in this study.

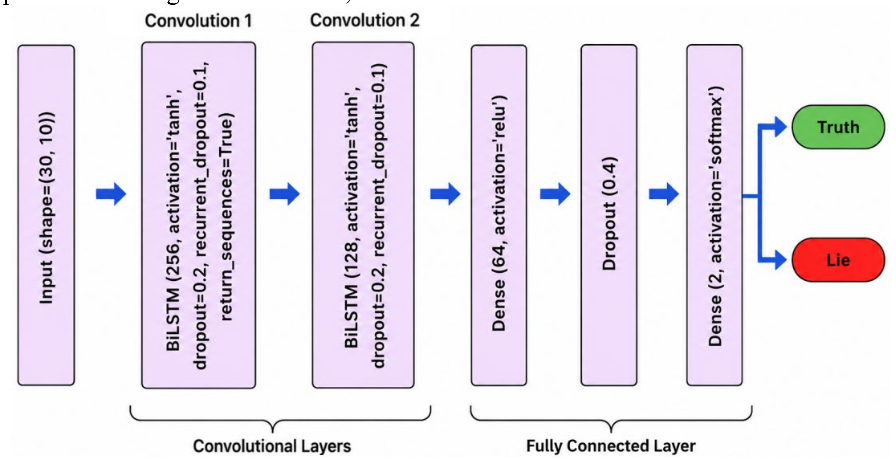


Figure 7. Architecture of the BiLSTM model applied to the ordered channel-sub-band-feature representation

The model takes an input of shape (30, 10) and comprises two stacked layers with 256 and 128 units (tanh activation, dropout=0.2), followed by a dense layer with 64 units (ReLU), a dropout layer (0.4), and a final softmax output layer with two neurons.

3.4.5. Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) networks are a simplified variant of LSTM architectures that employ fewer gating mechanisms while retaining the ability to model sequential dependencies. In the present input setting, these sequential dependencies correspond to relationships among ordered channel-sub-band-feature components rather than temporal dependencies in the original EEG waveform. GRUs consist of update and reset gates, enabling efficient learning with reduced computational complexity. In this study, GRU-based models were utilized to evaluate whether comparable performance to LSTM-based architectures could be achieved with fewer parameters and faster training. The hierarchical GRU structure allows for effective modeling of dependencies within the structured DWT-based statistical feature representation. Figure 8 shows the GRU model used in this study.

The GRU-based model receives an input sequence of shape (30,10), derived from the structured DWT-based statistical feature representation. The first GRU layer consists of 512 units with tanh activation and is configured with return sequences enabled to propagate the full sequential representation to the next layer. Batch normalization and dropout are applied to improve training stability and reduce overfitting. A second GRU layer with 128 units further summarizes the learned dependencies, followed again by batch normalization and dropout. The learned representation is then fed into a fully connected layer with 128 neurons and ReLU activation, followed by a dropout layer, and finally a two-neuron softmax output layer to perform binary classification.

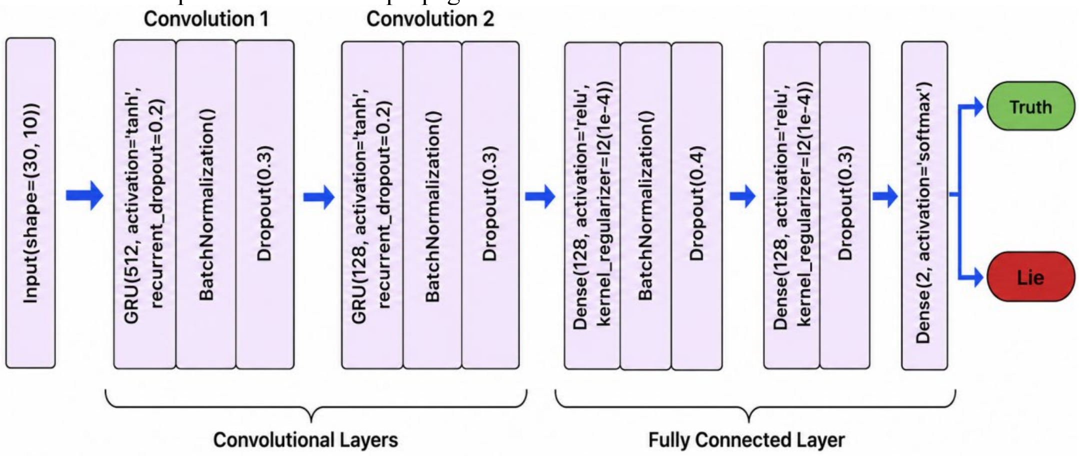


Figure 8. Architecture of the GRU model applied to the ordered channel-sub-band-feature representation

3.4.6. Random Forest (RF)

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and performs classification based on majority voting. RF models are robust to overfitting and capable of handling high-dimensional feature spaces. In this study, RF classifiers were employed to evaluate the performance of ensemble-based machine learning approaches on wavelet-derived EEG features.

3.4.7. k-Nearest Neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm is a distance-based classification method that assigns class labels based on the majority class among the nearest neighbors in the feature space. Despite its simplicity, KNN can be effective for EEG classification tasks when meaningful feature representations are available. In this study, KNN was included as a baseline model to assess the discriminative power of the extracted features.

3.4.8. Support Vector Machines (SVM)

Support Vector Machines (SVMs) aim to find an optimal decision boundary that maximizes the margin between different classes in a high-dimensional feature space. SVMs are particularly effective for non-linear classification problems when combined with kernel functions. In this work, SVM classifiers were used to evaluate the separability of truthful and deceptive EEG feature representations.

4. Experimental Setup

This section presents the experimental configuration adopted in this study, including the data splitting strategy, model hyperparameters, and evaluation metrics. Providing detailed parameter settings and mathematical definitions ensures reproducibility and transparent comparison across all classification models.

4.1. Data Splitting Strategy

To evaluate subject-independent generalization and to reduce the risk of data leakage caused by highly overlapping OSW segments, LOSOCV was adopted as the primary evaluation protocol. In each LOSOCV fold, all windows belonging to one subject were held out as the test set, while the windows of the remaining 26 subjects were used for training.

Since each subject had two recording sessions and each recording produced 289 windows, each LOSOCV fold contained 289×2=578 test windows. Across 27 subjects, the LOSOCV procedure resulted in 27 folds and a total of 27×578=15,606 test predictions. This protocol ensured that no window from the test subject appeared in the training set, thereby providing a more realistic assessment of subject-independent generalization.

For the deep learning models, a validation subset was created only from the training subjects within each fold. The held-out test subject was not used during model training, validation, or hyperparameter selection.

4.2. Model Hyperparameters

All classification models were trained using identical feature representations obtained from wavelet-based EEG preprocessing. The hyperparameter settings for deep learning and traditional machine learning models are summarized in Table 2. Hyperparameters were selected based on empirical evaluation and commonly adopted configurations in EEG-based classification studies.

Table 2. Hyperparameter settings of the classification models

Model	Key Parameters
KNN	Number of neighbors=5, distance metric: Euclidean
SVM	Kernel = RBF
RF	Number of trees = 100
CNN	Three convolutional blocks (Conv + BatchNorm + ReLU + Dropout + MaxPooling), six fully connected (dense) layers, Softmax output layer
LSTM	Two stacked layers (512 and 128 units, respectively, activation: tanh), Dropout (0.2), Dense (64, ReLU), Dropout (0.4), Softmax output layer
BiLSTM	Two stacked layers (256 and 128 units, respectively, activation: tanh), Dropout (0.2), Dense (64, ReLU), Dropout (0.4), Softmax output layer
GRU	Two stacked layers (512 and 128 units, respectively, activation: tanh), Dropout (0.3), Dense (128, ReLU), Dropout (0.4-0.3), Softmax output layer
EEGNet	F1=8, D=2, F2=16, kernel length=64, dropout=0.5, depthwise/separable convolutions, ELU activation, max-norm=0.25, Softmax output layer

All deep learning models were trained for 100 epochs with a batch size of 100. The Adam optimizer was used for model optimization. In preliminary empirical trials, SGD and RMSprop optimizers were also tested with different parameter settings; however, the Adam optimizer provided more stable training behavior and was therefore selected for the final experiments. The default learning rate of the Adam optimizer was used. Early stopping was not applied in order to maintain a fixed and consistent training schedule across all deep learning models. These settings were kept identical for EEGNet, CNN, LSTM, BiLSTM, and GRU models to ensure a fair comparison.

4.3. Evaluation Metrics

To assess the classification performance, standard metrics commonly used in binary classification problems were employed. Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy (Acc) measures the overall proportion of correctly classified samples:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Precision (P) represents the proportion of correctly predicted positive samples among all predicted positives:

$$P = \frac{TP}{TP + FP} \tag{15}$$

Recall (R), also known as sensitivity, measures the proportion of correctly identified positive samples:

$$R = \frac{TP}{TP + FN} \tag{16}$$

The F1-score ($F1$) is the harmonic mean of precision and recall, providing a balanced performance measure:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{17}$$

In addition to these metrics, confusion matrices were analyzed to examine class-wise classification behavior and identify potential misclassification patterns.

4.4. Implementation and Reproducibility Details

All experiments were implemented in Python 3.13.9. The main libraries used in the implementation were TensorFlow/Keras 2.21.0 for deep learning models, scikit-learn 1.7.2 for traditional machine learning models and preprocessing, NumPy 2.1.3 for numerical operations, and PyWavelets 1.9.0 for wavelet decomposition. To improve reproducibility, the random seed was fixed to 42 where applicable.

To address potential DWT configuration selection bias, configuration selection was embedded within each outer LOSOCV fold using a subject-wise inner holdout validation scheme. In each outer fold, one subject was reserved exclusively as the outer-test subject. Of the remaining 26 subjects, 23 were used for inner training and three were used for inner validation. All windows belonging to a given subject were assigned exclusively to one inner partition. The same inner subject split was used to evaluate the db4-Level4, db4-Level5, db5-Level4, and db5-Level5 configurations using the CNN model. StandardScaler was fitted exclusively on the 23 inner-training subjects and was then applied without refitting to the three inner-validation subjects. The DWT configuration achieving the highest inner-validation macro F1-score was selected. Subsequently, a new StandardScaler was fitted using all 26 outer-training subjects, the CNN model was retrained using the selected DWT representation, and the held-out outer-test subject was evaluated once. Therefore, the outer-test subject was not involved in feature standardization, DWT configuration selection, or model training. The subject-wise inner holdout procedure described above was used specifically for DWT configuration selection. The classifier-comparison experiments retained the ordered 10% validation procedure described below.

For the deep learning models, 10% of the ordered outer-training samples were used as a validation subset for validation-loss-based checkpoint selection. Keras selected the final 10% of the input samples before training-time shuffling. Because the feature files were organized as contiguous subject-wise blocks, the validation subset was not constructed by randomly distributing windows from all outer-training subjects. Instead, it consisted primarily of complete terminal subject blocks, although one subject block could be divided at the training-validation boundary. The held-out outer-test subject was never included in training, validation, feature standardization, or checkpoint selection.

The traditional machine learning classifiers were implemented using scikit-learn. The SVM classifier used an RBF kernel with C=1.0 and gamma=scale, corresponding to the default scikit-learn settings. The RF classifier was trained with 100 trees, max_depth=None, and min_samples_split=2. The KNN classifier used k=5 with uniform weighting and the Minkowski distance metric with p=2, which corresponds to the Euclidean distance.

Hyperparameters were selected based on preliminary empirical evaluation and commonly adopted default configurations in EEG-based classification studies. The held-out LOSOCV test subject was not used during hyperparameter selection. Therefore, the test fold remained isolated from all training, validation, feature standardization, and model selection procedures.

5. Results and Discussion

This section presents the experimental results obtained from the evaluated EEG-based deception detection pipeline and provides a comparative discussion of the classification models. The performance of deep learning-based and traditional machine learning-based approaches is analyzed using standard evaluation metrics under the LOSOCV protocol.

5.1. Ablation Analysis of DWT Configurations

A DWT-configuration ablation analysis was conducted within an outer LOSOCV framework using three-subject, subject-wise inner holdout validation. Four DWT configurations were compared: db4-Level4, db4-Level5, db5-Level4, and db5-Level5. This analysis aimed to determine how different time-frequency representations influence subject-independent classification performance on the LieWaves dataset.

As shown in Table 3, the db5-Level5 configuration achieved the highest overall performance among the tested DWT settings under the subject-wise inner holdout evaluation conducted within the outer LOSOCV framework. Specifically, db5-Level5 achieved the highest mean inner-validation accuracy of 0.5335 and the highest mean inner-validation macro F1-score of 0.4929. The db5-Level4 configuration achieved the second-highest performance, with an accuracy of 0.5266 and a macro F1-score of 0.4893. The db4-Level4 configuration yielded an accuracy of 0.5260 and a macro F1-score of 0.4834, whereas db4-Level5 produced the lowest accuracy and macro F1-score values of 0.5204 and 0.4636, respectively.

Table 3. Mean inner-validation performance of the DWT configurations across the 27 outer LOSOCV folds using three-subject, subject-wise inner holdout validation

DWT configuration	Accuracy	Macro Precision	Macro Recall	Macro F1-score
db4-Level4	0.5260	0.5323	0.5260	0.4834
db4-Level5	0.5204	0.5339	0.5204	0.4636
db5-Level4	0.5266	0.5344	0.5266	0.4893
db5-Level5	**0.5335**	**0.5368**	**0.5335**	**0.4929**

These findings indicate that the effect of decomposition depth varied according to the selected mother wavelet. For the db5 wavelet, Level-5 achieved higher accuracy and macro F1-score than Level-4. In contrast, for the db4 wavelet, Level-4 outperformed Level-5. Among all evaluated configurations, db5-Level5 achieved the highest overall performance, although the differences between the leading configurations remained relatively small. Therefore, the result should be interpreted within the scope of the tested DWT configurations, the LieWaves dataset, the CNN-based evaluation setting, and the adopted nested subject-independent evaluation protocol. The performance of db5-Level5 should not be generalized as evidence that it is universally optimal for EEG-based lie detection.

Following fold-specific DWT-configuration selection, the CNN model was retrained using all 26 outer-training subjects and evaluated once on the corresponding held-out outer-test subject (Table 4).

Across the 27 outer LOSOCV folds, the final bias-controlled outer-test evaluation yielded an accuracy of 0.5144, macro precision of 0.5303, macro recall of 0.5144, and macro F1-score of 0.4631. After concatenating all 15,606 outer-test predictions, the pooled macro F1-score was 0.5075. These results represent the final outer-test performance obtained after DWT-configuration selection had been conducted exclusively within the corresponding outer-training partition.

Table 4. Final outer-test performance after fold-specific DWT-configuration selection

Accuracy	Macro Precision	Macro Recall	Macro F1-score	Pooled Macro F1
0.5144	0.5303	0.5144	0.4631	0.5075

Mean values were calculated across the 27 outer LOSOCV folds. Pooled macro F1-score was calculated after concatenating all 15,606 outer-test predictions and therefore differs from the mean of the 27 fold-level macro F1-scores.

The lower outer-test accuracy compared with the earlier non-nested CNN estimate indicates that evaluating DWT configurations using the same outer folds can produce an optimistic performance estimate. The subject-wise inner holdout procedure provides a more conservative assessment by preventing the outer-test subject from influencing DWT-configuration selection.

5.2. Quantitative Performance Evaluation

The classifier-comparison results reported in Table 5 were obtained using the fixed db5-Level5 representation and should therefore be interpreted separately from the bias-controlled fold-specific DWT-selection analysis reported in Table 4. To assess subject-independent generalization, all classification models were evaluated using the LOSOCV protocol. The results are reported as mean ± standard deviation across 27 folds. In each fold, one subject was used as the test set, while the remaining 26 subjects were used for training. Since each subject contributed 578 test windows, the LOSOCV evaluation produced 15,606 test predictions across all folds.

As shown in Table 5, EEGNet achieved the highest mean accuracy of 0.5779±0.0903 and the highest macro F1-score of 0.5295±0.1359 among the evaluated models. CNN achieved the second-highest mean accuracy of 0.5336±0.0773, followed by SVM with 0.5227±0.0783. Among the traditional machine learning models, KNN produced the highest macro F1-score of 0.4964±0.0563, followed by RF with 0.4797±0.0933 and SVM with 0.4646±0.0997.

To verify whether the reported accuracy of the best-performing model was statistically above the chance level, a fold-level chance analysis was conducted using the 27 LOSOCV accuracy values of EEGNet. The chance level was set to 0.50 for the binary lie/truth classification task. Since one fold had accuracy exactly equal to 0.50, it was treated as a tie and excluded from the sign-based comparison. Among the remaining 26 non-tied folds, 20 folds achieved accuracy values above 0.50 and 6 folds were below 0.50. An exact one-sided binomial sign test showed that the number of above-chance folds was statistically significant (20/26, $p=0.00468$). In addition, a one-sided Wilcoxon signed-rank test against the chance level of 0.50 also indicated statistically above-chance performance ($W=311.0$, $p=0.000289$). These results suggest that EEGNet achieved statistically above-chance accuracy under the LOSOCV protocol. However, considering the modest absolute accuracy value, this finding should not be interpreted as evidence of practical or forensic applicability.

Table 5. LOSOCV-based classifier-comparison performance using the fixed db5-Level5 DWT representation, reported as mean ± standard deviation across 27 folds

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score
EEGNet	**0.5779 ± 0.0903**	**0.5741 ± 0.1387**	**0.5779 ± 0.0903**	**0.5295 ± 0.1359**
CNN	0.5336 ± 0.0773	0.5734 ± 0.1242	0.5336 ± 0.0773	0.4611 ± 0.1125
SVM	0.5227 ± 0.0783	0.5227 ± 0.1255	0.5227 ± 0.0783	0.4646 ± 0.0997
GRU	0.5167 ± 0.0620	0.4546 ± 0.1668	0.5167 ± 0.0620	0.4134 ± 0.1099
BiLSTM	0.5141 ± 0.0678	0.5109 ± 0.1653	0.5141 ± 0.0678	0.4318 ± 0.1026
LSTM	0.5074 ± 0.0924	0.4678 ± 0.1677	0.5074 ± 0.0924	0.4408 ± 0.1166
RF	0.5069 ± 0.0872	0.5202 ± 0.1120	0.5069 ± 0.0872	0.4797 ± 0.0933
KNN	0.5043 ± 0.0504	0.5031 ± 0.0555	0.5043 ± 0.0504	0.4964 ± 0.0563

Overall, all models exhibited limited performance under the LOSOCV setting, with mean accuracy values remaining close to chance level. This result indicates that subject-independent EEG-based lie detection on the LieWaves dataset is substantially more challenging than sample-level classification. The findings also highlight the importance of using subject-independent validation protocols when evaluating EEG-based deception detection models, particularly when highly overlapping window-based segmentation is applied.

5.3. Statistical Significance Analysis

To further examine the statistical robustness of the LOSOCV-based model comparison, non-parametric statistical tests were performed on the fold-wise results obtained from the 27 LOSOCV folds. Since each fold corresponded to the same held-out subject across all models, the comparisons were treated as paired repeated measurements.

It should be noted, however, that the fold-level statistical results should be interpreted with caution. Although LOSOCV prevents subject-level data leakage by ensuring that all windows of the held-out subject are excluded from training, validation, and hyperparameter selection, the fold-wise results are still obtained from the same dataset. Moreover, for any pair of distinct LOSOCV folds, the corresponding outer-training sets share 25 of the 27 subjects, which provides an additional source of dependence among the fold-level performance estimates. In addition, the EEG samples were generated using a highly overlapping window-based segmentation procedure. In particular, the 91.67% overlap ratio means that adjacent windows share most of their samples, which may reduce the effective independence of window-level observations and should be considered when interpreting statistical significance. Therefore, the fold-level performance values may not fully satisfy the strict independence assumption required by some non-parametric statistical tests. For this reason, the Friedman and Wilcoxon signed-rank tests with Holm correction were used as supportive statistical evidence rather than as fully independent confirmatory tests.

First, the Friedman test was applied to determine whether there were statistically significant differences among the evaluated classifiers. As shown in Table 6, the Friedman test indicated a significant difference among the models for both accuracy ($\chi^2(7) = 16.575$, $p = 0.0204$) and macro F1-score ($\chi^2(7) = 29.166$, $p = 0.0001$). Therefore, post-hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test. Since EEGNet achieved the highest mean accuracy and macro F1-score in the LOSOCV evaluation, it was used as the reference model in the pairwise comparisons. Holm-Bonferroni correction was applied to adjust p-values for multiple comparisons.

Table 6. Friedman test results for LOSOCV-based model comparison

Metric	χ^2	df	p-value	Interpretation
Accuracy	16.575	7	0.0204	Significant
Macro F1-score	29.166	7	0.0001	Significant

The post-hoc results are presented in Table 7. EEGNet achieved significantly higher accuracy than all other evaluated models after Holm correction. For macro F1-score, EEGNet significantly outperformed SVM, GRU, BiLSTM, and LSTM. However, the differences between EEGNet and CNN, RF, and KNN in terms of macro F1-score were not statistically significant after Holm correction.

Table 7. Holm-corrected Wilcoxon signed-rank test results comparing EEGNet with the other models

Comparison	Accuracy mean difference	Accuracy p-adjusted	Accuracy significant?	Macro F1 mean difference	Macro F1 p-adjusted	Macro F1 significant?
EEGNet-CNN	0.0443	0.0105	Yes	0.0684	0.0560	No
EEGNet-SVM	0.0552	0.0028	Yes	0.0649	0.0148	Yes
EEGNet-GRU	0.0612	0.0067	Yes	0.1161	0.0092	Yes
EEGNet-BiLSTM	0.0638	0.0075	Yes	0.0977	0.0214	Yes
EEGNet-LSTM	0.0706	0.0057	Yes	0.0887	0.0219	Yes
EEGNet-RF	0.0711	0.0075	Yes	0.0499	0.1811	No
EEGNet-KNN	0.0736	0.0035	Yes	0.0331	0.1811	No

These findings indicate that although EEGNet obtained the highest average performance, the statistical evidence is stronger for accuracy than for macro F1-score. In other words, the statistical analysis supports the overall superiority of EEGNet in terms of accuracy; however, EEGNet did not significantly outperform all models in terms of macro F1-score after multiple-comparison correction. Therefore, the results should be interpreted cautiously, particularly with respect to class-balanced performance.

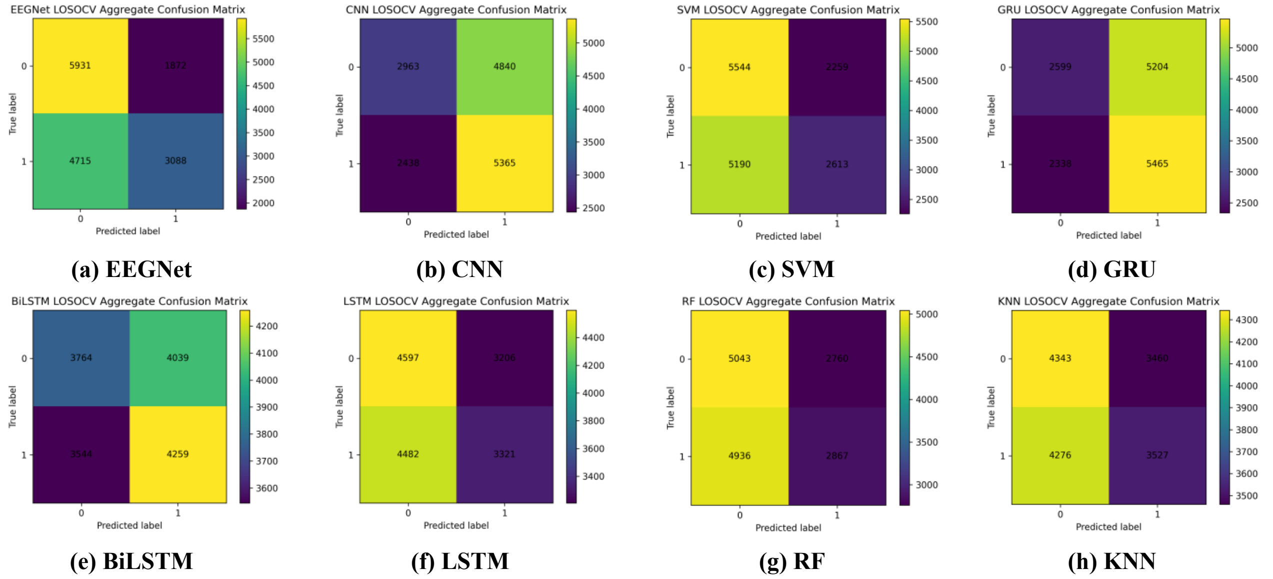
5.4. Confusion Matrix Analysis

Figure 9 presents the aggregate confusion matrices obtained by combining the test predictions from all 27 LOSOCV folds. Since each LOSOCV fold contained 578 test windows, each aggregate confusion matrix represents a total of 15,606 test predictions.

The confusion matrices reveal different class-wise prediction tendencies among the evaluated models. EEGNet achieved the highest overall performance among the tested classifiers. It correctly classified 5931 class-0 samples out of 7803, while 1872 class-0 samples were misclassified as class 1. For class 1, EEGNet correctly classified 3088 samples out of 7803, while 4715 class-1 samples were misclassified as class 0. This result indicates that although EEGNet achieved the best overall LOSOCV performance, its predictions were not fully balanced across the two classes.

CNN and GRU showed a stronger tendency to predict class 1, whereas EEGNet, SVM, RF, and LSTM showed a stronger tendency toward class 0. KNN produced a relatively more balanced distribution across the two classes; however, its overall performance remained close to chance level. These class-wise prediction patterns explain why the aggregate accuracies remained limited despite some models performing relatively better for one class.

Overall, the confusion matrix analysis supports the LOSOCV-based quantitative findings reported in Table 5. The results show that subject-independent EEG-based lie detection on the LieWaves dataset remains challenging and that the evaluated models do not achieve consistently balanced performance across the two classes.



(a) EEGNet **(b) CNN** **(c) SVM** **(d) GRU**

(e) BiLSTM **(f) LSTM** **(g) RF** **(h) KNN**

Figure 9. Aggregate confusion matrices obtained by combining test predictions from all 27 LOSOCV folds for the evaluated classification models.

5.5. Training and Validation Behavior Analysis

Figure 10 presents the training and validation accuracy curves of the deep learning models for the S5 LOSOCV fold. This figure is provided as a representative example to illustrate the training behavior of the models under the subject-independent evaluation setting. Therefore, it should not be interpreted as the average training behavior across all 27 LOSOCV folds.

As shown in Figure 10, most deep learning models achieved increasing and relatively high training accuracy, whereas validation accuracy remained substantially lower. This gap indicates that the models were able to fit the training data but had difficulty generalizing to validation samples under the LOSOCV setting. EEGNet exhibited a more stable training curve compared with the other deep learning models, although its validation accuracy also remained limited.

CNN, GRU, BiLSTM, and LSTM showed larger gaps between training and validation curves. In particular, CNN, BiLSTM, and GRU reached high training accuracy, while their validation accuracy remained at considerably lower levels. This pattern suggests overfitting to training-subject-specific characteristics and limited generalization to unseen or held-out subject-related patterns.

These observations are consistent with the LOSOCV-based quantitative results reported in Table 5. Although deep learning models can learn discriminative patterns from the DWT-based structured feature representations, their subject-independent generalization performance remains limited on the LieWaves dataset. Therefore, the training and validation behavior supports the main finding that EEG-based lie detection remains challenging under subject-independent evaluation.

5.6. General Discussion

The experimental findings provide important insights into the subject-independent generalization capability of EEG-based lie detection models on the LieWaves dataset. The present study was initially motivated by and built upon the methodological pipeline introduced in the original LieWaves study, which combined ATAR-based artifact removal, OSW-based segmentation, DWT-based feature extraction, and deep learning-based classification. The original study demonstrated very high classification performance using this pipeline, particularly with DWT-based features and LSTM models. However, the use of highly overlapping OSW segments together with sample-level training, validation, and test splits may lead to optimistic performance estimates, since highly similar windows originating from the same subject or recording session can be distributed across different subsets.

A similar caution applies to the recurrent models evaluated in this study. Although LSTM, BiLSTM, and GRU architectures are commonly associated with temporal sequence modeling, they were not applied to the original EEG time series in the present framework. Instead, their inputs were ordered channel-sub-band-feature representations derived from DWT-based statistical features. Therefore, the recurrent models should be interpreted as modeling dependencies among structured wavelet-derived feature components rather than temporal dynamics directly from raw EEG signals. This distinction is important when interpreting their performance and comparing the present results with studies that apply recurrent architectures directly to raw or minimally processed EEG sequences.

In this study, the LOSOCV-based evaluation provides a stricter assessment of subject-independent generalization. Unlike sample-level evaluation, LOSOCV ensures that all windows belonging to the held-out subject are excluded from model training and validation and are used exclusively for testing. Thus, although LOSOCV eliminates subject-level leakage, it does not make the highly overlapping windows statistically independent observations.

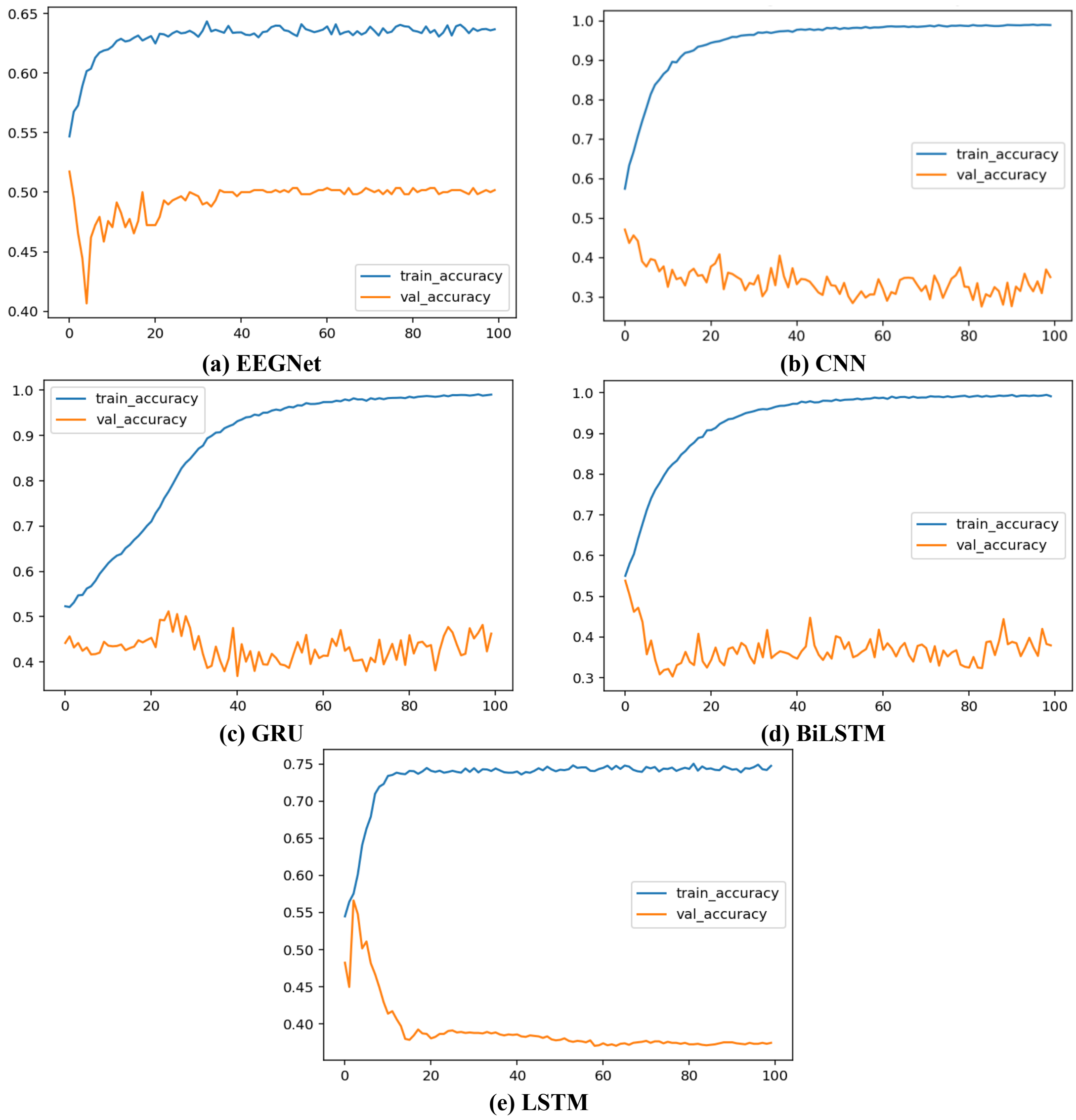


Figure 10. Training and validation accuracy curves of (a) EEGNet, (b) CNN, (c) GRU, (d) BiLSTM, and (e) LSTM for the S5 LOSOCV fold.

Because the 90%–10% validation boundary did not always coincide exactly with a subject boundary, one outer-training subject could be partially represented in both the training and validation subsets. Given the high overlap between adjacent OSW segments, a limited dependency may therefore exist between windows belonging to this boundary subject; however, the outer-test subject remained completely isolated. The substantial decrease in performance observed under LOSOCV indicates that EEG-based deception detection becomes considerably more challenging when subject-specific information is excluded from the training process. Therefore, the findings do not contradict the usefulness of the LieWaves dataset or the effectiveness of the original signal processing pipeline; rather, they highlight that the evaluation protocol plays a critical role in interpreting model performance.

A limitation of the statistical analysis is the partial dependency among the LOSOCV fold-level results. Although each fold evaluates a different held-out subject and therefore prevents direct subject-level leakage into the corresponding test set, the training sets of different LOSOCV folds overlap substantially. Specifically, any two folds share data from 25 of the 27 subjects, differing only in the subjects assigned to their respective test sets. Consequently, the fold-level performance estimates should not be regarded as fully independent observations. An additional source of dependence arises from the 91.67% overlap between adjacent OSW segments within each subject. Therefore, the strict independence assumptions underlying confirmatory interpretations of the Friedman and Wilcoxon signed-rank tests may not be fully satisfied. Accordingly, these statistical results are interpreted as supportive evidence for comparative model assessment rather than as definitive proof of model superiority.

Another limitation concerns the statistical feature set used in this study. Both standard deviation and variance were included as dispersion-related descriptors; however, these two measures are mathematically redundant because variance is directly derived from standard deviation. Although both descriptors were retained to maintain the consistency of the experimental pipeline and feature dimensionality, this redundancy may increase the feature space without providing fully independent information. Future studies should consider feature selection, dimensionality reduction, or redundancy-aware feature engineering strategies to obtain a more compact and potentially more robust EEG representation.

The use of highly overlapping OSW segments is an important factor to consider when interpreting EEG classification results. In this study, the overlap ratio between consecutive windows was approximately 91.67%, meaning that adjacent windows share a large proportion of signal samples. Under random sample-level splitting, highly similar windows from the same subject or session may appear in both training and test sets. In contrast, the LOSOCV protocol mitigates this issue by evaluating each model on completely unseen subjects.

Among the evaluated models, EEGNet achieved the highest mean accuracy and macro F1-score under the LOSOCV protocol. The statistical significance analysis further showed that EEGNet significantly outperformed all other models in terms of accuracy after Holm correction. However, for macro F1-score, EEGNet did not significantly outperform CNN, RF, and KNN after multiple-comparison correction. This suggests that EEGNet provides a more consistent advantage in terms of overall accuracy, whereas its superiority in class-balanced performance should be interpreted more cautiously.

However, a methodological limitation should also be noted regarding the use of EEGNet. Although EEGNet achieved the highest average performance in this study, it was applied to DWT-based statistical feature matrices rather than raw EEG trials. This input representation differs from the original assumptions of EEGNet, which was designed to exploit temporal and spatial patterns directly from EEG signals. Therefore, the EEGNet results should be interpreted with caution, as the performance of an EEGNet-inspired compact convolutional model on wavelet-derived statistical feature matrices, and should not be generalized as evidence of the superiority of the canonical EEGNet architecture for raw EEG-based lie detection.

The confusion matrix analysis also supports this interpretation. Although EEGNet achieved the best overall LOSOCV performance, its predictions were not fully balanced across the two classes. Similar class-wise tendencies were observed for other models, indicating that subject-independent EEG-based lie detection remains difficult not only in terms of overall accuracy but also in terms of balanced classification of truthful and deceptive responses.

The DWT-configuration ablation analysis conducted within the outer LOSOCV framework using three-subject, subject-wise inner holdout validation showed that db5-Level5 achieved the highest mean inner-validation performance among the tested configurations when evaluated using the CNN model. This suggests that both the choice of mother wavelet and decomposition depth affect the discriminative representation of EEG signals under subject-independent evaluation. In particular, the effect of decomposition depth differed between the two mother wavelets. Level-5 performed better than Level-4 for db5, whereas Level-4 performed better than Level-5 for db4. Nevertheless, db5-Level5 achieved the highest mean inner-validation performance among the evaluated DWT configurations. However, the limited LOSOCV performance indicates that improving the wavelet representation alone is not sufficient to ensure robust subject-independent generalization. Therefore, DWT configuration selection should be considered an important but not sufficient component of EEG-based lie detection systems.

The bias-controlled outer-test performance obtained after fold-specific DWT selection was lower in accuracy than the earlier fixed-configuration CNN estimate. This difference demonstrates the importance of separating DWT-configuration selection from outer-test evaluation. Nevertheless, the subject-independent nested evaluation used a single three-subject inner holdout partition within each outer LOSOCV fold rather than a full inner LOSOCV procedure. Therefore, the selected configuration may remain sensitive to the particular subjects assigned to inner validation.

Overall, the results emphasize the importance of rigorous validation protocols in EEG-based deception detection. While wavelet-based statistical features and deep learning architectures may capture informative

EEG patterns, their generalization to unseen subjects remains limited on the LieWaves dataset.

These findings should therefore be interpreted within the scope of an exploratory and comparative evaluation. Although EEGNet achieved the highest average LOSOCV performance among the tested models, the absolute accuracy and macro F1-score values remained modest and relatively close to chance level. Therefore, these results do not provide evidence that the proposed pipeline is suitable for practical or forensic lie detection. Rather, the main value of the study lies in showing how DWT-based EEG feature representations and different classifiers perform when evaluated under a more rigorous subject-independent LOSOCV protocol. Thus, the results are best described as modest but informative findings obtained under a rigorous subject-independent evaluation setting. This interpretation is particularly important because subject-independent evaluation provides a more realistic assessment of generalization to unseen individuals than sample-level splitting.

Future studies should investigate larger and more diverse EEG datasets, alternative segmentation strategies with lower overlap ratios, subject-adaptive learning, domain generalization, and explainable artificial intelligence approaches to better understand deception-related EEG patterns and improve subject-independent performance. In particular, future work should incorporate XAI-based channel and sub-band importance analyses to identify which EEG electrodes and wavelet-derived frequency components contribute most to deception-related classification decisions.

6. Conclusions

This study presented a systematic and comparative evaluation of machine learning and deep learning models for EEG-based lie detection using the publicly available LieWaves dataset. Building upon the original LieWaves methodology, the study employed ATAR-based artifact removal, OSW-based segmentation, DWT-based feature extraction, and classification models within a unified experimental pipeline. Accordingly, the study should be viewed as a comparative subject-independent evaluation framework rather than as the development of a ready-to-use lie detection system.

In contrast to evaluations based only on sample-level splitting, this study assessed subject-independent generalization using the LOSOCV protocol. The results showed that all evaluated models exhibited limited performance when tested on unseen subjects. This finding indicates that EEG-based lie detection on the LieWaves dataset is substantially more challenging under subject-independent evaluation than under sample-level evaluation settings.

The comparative model evaluation showed that EEGNet achieved the highest average LOSOCV performance among the evaluated models, with an accuracy of 0.5779±0.0903 and a macro F1-score of 0.5295±0.1359. Statistical significance analysis further indicated that EEGNet significantly outperformed the other models in terms of accuracy after Holm correction. However, its superiority in terms of macro F1-score was not statistically significant relative to all models, suggesting that class-balanced performance should be interpreted cautiously. A fold-level chance analysis further confirmed that EEGNet accuracy was statistically above the 0.50 chance level; nevertheless, the absolute performance remained modest and should not be interpreted as sufficient for practical or forensic lie detection. It should also be noted that the EEGNet result corresponds to an EEGNet-inspired compact convolutional model applied to DWT-based statistical feature matrices, not to the canonical EEGNet architecture trained directly on raw EEG trials.

The DWT-configuration evaluation conducted within an outer LOSOCV framework using three-subject, subject-wise inner holdout validation showed that db5-Level5 achieved the highest mean inner-validation performance among the tested configurations. Following fold-specific DWT selection, the final bias-controlled CNN evaluation yielded an outer-test accuracy of 0.5144, a macro F1-score of 0.4631, and a pooled macro F1-score of 0.5075 across 15,606 outer-test predictions. This finding suggests that both the choice of mother wavelet and decomposition depth affect the representation capacity of DWT-based EEG features. However, the limited LOSOCV performance also shows that improving the wavelet representation alone is not sufficient to ensure robust subject-independent generalization. Therefore, the db5-Level5 result should be interpreted as dataset- and protocol-specific rather than as a universal recommendation for EEG-based lie detection.

In general, the findings highlight the critical role of the evaluation protocol in EEG-based deception detection. The results do not diminish the usefulness of the LieWaves dataset or the value of the original signal processing pipeline; rather, they show that subject-independent validation is necessary for realistically assessing model generalization. Future studies should focus on larger and more diverse EEG datasets, reduced-overlap segmentation strategies, subject-adaptive and domain-generalization approaches, and explainable artificial intelligence methods to improve both robustness and interpretability in EEG-based lie detection.

In summary, the contribution of this study is not the demonstration of a deployable EEG-based lie detection system, but the systematic comparison of DWT-derived EEG representations and multiple classifiers under a subject-independent LOSOCV framework. The modest performance values highlight the need for larger datasets, more robust representations, and careful validation before EEG-based deception detection can be considered for any practical application.

Ethical Statement and Data Availability

This study did not involve any new experiments with human participants, interventions, or data collection. All analyses were conducted using the publicly available LieWaves EEG dataset introduced by Aslan et al. (2024a). The original LieWaves study reported that the EEG recordings and the signal acquisition procedures were conducted in compliance with ethical approval granted by the relevant ethics committee. Therefore, no additional ethical approval was required for the present secondary analysis.

Considering the sensitive nature of EEG-based lie detection research, the findings of this study should be interpreted only within an academic, methodological, and comparative evaluation context. The reported results should not be used for practical, forensic, legal, or individual-level decision-making purposes.

The LieWaves dataset is publicly available through Mendeley Data with DOI: 10.17632/5gzxb2bzs2.2. The source code used in this study is publicly available at the following GitHub repository: https://github.com/rabiacansel/EEG-Based-Lie-Detection

Acknowledgment

This study was conducted within the scope of the Design Project course at the Department of Computer Engineering, Bilecik Şeyh Edebali University. The authors would like to express their gratitude to Aslan et al. for making the LieWaves dataset publicly available, which significantly contributed to the experimental evaluation of this research.

References

Acharya, U.R., Oh, S.L., Hagiwara, Y., Tan, J.H., Adeli, H., 2018. Deep convolutional neural network for the automated detection of seizures using EEG signals. Computers in Biology and Medicine, 100, 270-278.

Amin, H.U., Mumtaz, W., Subhani, A.R., Saad, M.N.M., Malik, A.S., 2017. Classification of EEG signals based on pattern recognition approach. Frontiers in Computational Neuroscience, 11, 103.

Aslan, M., Baykara, M., Alakus, T.B., 2024a. LieWaves: dataset for lie detection based on EEG signals and wavelets. Medical & Biological Engineering & Computing, 62(5), 1571-1588.

Aslan, M., Baykara, M., Alakus, T.B., 2024b. LSTMNCP: lie detection from EEG signals with novel hybrid deep learning method. Multimedia Tools and Applications, 83(11), 31655-31671.

Bhatt, P., Sethi, A., Tasgaonkar, V., Shroff, J., Pendharkar, I., Desai, A., Jain, N.K., 2023. Machine learning for cognitive behavioral analysis: datasets, methods, paradigms, and research directions. Brain Informatics, 10(1), 18.

Chanda, D., Mandal, R.K., 2023. A comprehensive review and future prospects of lie detection using machine learning. International Conference on Computational Technologies and Electronics, November 2023, pp. 64-77.

Ekman, P., 2009. Telling lies: Clues to deceit in the marketplace, politics, and marriage, 4th ed., W.W. Norton & Company, New York.

Farwell, L.A., Donchin, E., 1991. The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. Psychophysiology, 28(5), 531-547.

Joshi, G., Tasgaonkar, V., et al., 2025. Multimodal machine learning for deception detection using behavioral and physiological data. Scientific Reports, 15(1), 8943.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. Journal of Neural Engineering, 15(5), 056013.

Nagale, T., Khandare, A., 2025. Spontaneous deception detection using EEG signals: Advances, challenges, and future directions in deep learning frameworks. 2025 IEEE International Conference on Emerging Trends in Engineering and Computing (ETECOM), October 2025, pp. 1-6.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M., 2015. Deception detection using real-life trial data. Proceedings of the 2015 ACM International Conference on Multimodal Interaction, Seattle, USA, November 2015, pp. 59-66.

Rahmani, M., Mohajelin, F., Khaleghi, N., Sheykhivand, S., Danishvar, S., 2024. An automatic lie detection model using EEG signals based on the combination of type 2 fuzzy sets and deep graph convolutional networks. Sensors, 24(11), 3598.

Raj, V.A., Parupudi, T., Thalengala, A., Nayak, S.G., 2025. A comprehensive review of deep learning models for denoising EEG signals: challenges, advances, and future directions. Discover Applied Sciences, 7(11), 1268.

Rosenfeld, J.P., Soskins, M., Bosh, G., Ryan, A., 2004. Simple, effective countermeasures to P300-based tests of detection of concealed information. Psychophysiology, 41(2), 205-219.

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J., 2019. Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering, 16(5), 051001.

Şen, M.U., Perez-Rosas, V., Yanikoglu, B., Abouelenien, M., Burzo, M., Mihalcea, R., 2022. Multimodal deception detection using real-life trial data. IEEE Transactions on Affective Computing, 13(1), 306-319.

Subasi, A. (2007). EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Systems with Applications, 32(4), 1084-1093.

Taha, B.N., Baykara, M., Alakuş, T.B., 2025. Neurophysiological approaches to lie detection: A systematic review. Brain Sciences, 15(5), 519.

Tarawneh, A.A., Kamel, A., Naceur, M.S., 2025. Comparative performance analysis of machine and deep learning models for EEG-based biometric authentication. Informatica, 49(8).

Uotinen, M., Simola, P., Henttonen, P.J., 2024. Detecting deceit within a predominantly true statement using two parallel assessment methods: A pilot study. Journal of Military Studies, 1(1).

Vrij, A., 2008. Detecting lies and deceit: Pitfalls and opportunities, 2nd ed., John Wiley & Sons, Chichester.

Vrij, A., Granhag, P.A., Porter, S., 2010. Pitfalls and opportunities in nonverbal and verbal lie detection. Psychological Science in the Public Interest, 11(3), 89-121.