IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 10 April 2024, accepted 2 May 2024, date of publication 8 May 2024, date of current version 17 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399114

TOPICAL REVIEW

Signal to Image Conversion and Convolutional Neural Networks for Physiological Signal Processing: A Review

K. E. CH VIDYASAGAR[1,2], K. REVANTH KUMAR[2], (Student Member, IEEE), G. N. K. ANANTHA SAI[2], (Student Member, IEEE), MUNAGALA RUCHITA[2], (Student Member, IEEE), AND MANOB JYOTI SAIKIA[1,3], (Member, IEEE)

[1]Biomedical Sensors and Systems Laboratory, University of North Florida, Jacksonville, FL 32224, USA
[2]Department of Biomedical Engineering, University College of Engineering, Osmania University, Hyderabad 500007, India
[3]Department of Electrical Engineering, University of North Florida, Jacksonville, FL 32224, USA

Corresponding author: Manob Jyoti Saikia (manob.saikia@unf.edu)

ABSTRACT Physiological signals obtained from electroencephalography (EEG), electromyography (EMG), and electrocardiography (ECG) provide valuable clinical information but pose challenges for analysis due to their high-dimensional nature. Traditional machine learning techniques, relying on hand-crafted features from fixed analysis windows, can lead to the loss of discriminative information. Recent studies have demonstrated the effectiveness of deep convolutional neural networks (CNNs) for robust automated feature learning from raw physiological signals. However, standard CNN architectures require two-dimensional image data as input. This has motivated research into innovative signal-to-image (STI) transformation techniques to convert one-dimensional time series into images preserving spectral, spatial, and temporal characteristics. This paper reviews recent advances in strategies for physiological signal-to-image conversion and their applications using CNNs for automated processing tasks. A systematic analysis of EEG, EMG, and ECG signal transformation and CNN-based analysis techniques spanning diverse applications, including brain-computer interfaces, seizure detection, motor control, sleep stage classification, arrhythmia detection, and more, are presented. Key insights are synthesized regarding the relative merits of different transformation approaches, CNN model architectures, training procedures, and benchmark performance. Current challenges and promising research directions at the intersection of deep learning and physiological signal processing are discussed. This review aims to catalyze continued innovations in effective end-to-end systems for clinically relevant information extraction from multidimensional physiological data using convolutional neural networks by providing a comprehensive overview of state-of-the-art techniques.

INDEX TERMS Biomedical signal analysis, convolutional neural networks, deep learning, machine learning, physiological signals, signal-to-image conversion.

I. INTRODUCTION

Physiological signals obtained from electroencephalography (EEG), electromyography (EMG), and electrocardiography (ECG) provide valuable insights into various aspects of human health and function. Accurate signal analysis is the keystone upon which our understanding of human physiology enables us to unravel intricate narratives inscribed in these physiological signals. However, the analysis of these signals poses significant challenges due to their multidimensional nature comprising spatial, spectral, and temporal information and are often contaminated with noise and artifacts from various sources like powerline interference, motion artifacts, baseline wander, electrode movement, etc. which must be filtered, denoised and pre-processed. For multichannel data (e.g., multi-lead ECG and high-density EEG/EMG),

The associate editor coordinating the review of this manuscript and approving it for publication was Henry Hess.

TABLE 1. List of acronyms.

CNN	Convolutional Neural Network	WPT	Wavelet Packet Transform	RP	Recurrence Plot
STI	Signal-to-Image Transformation	DFT	Discrete Fourier Transform	DE	Differential entropy
STFT	Short-Time Fourier Transform	IMFs	Intrinsic Mode Functions	FC	Fully Connected
CWT	Continuous Wavelet Transform	PSD	Power Spectral Density	EMG	Electromyogram
FFT	Fast Fourier Transform	SVM	Support Vector Machine	EEG	Electroencephalogram
DWT	Discrete Wavelet Transform	BCI	Brain-Computer Interface	ReLU	Rectified Linear Unit Activation
EMD	Empirical Mode Decomposition	ECG	Electrocardiogram	LSTM	Long-short Term Memory
MI	Motor Imagery	RNN	Recurrent Neural Network	SGDM	Stochastic Gradient Descent Moment
AEP	Azimuthal Equidistant Projection	KNN	K Nearest Neighbor	MEMD	Multivariate EMD
CSP	Common Spatial Pattern	LBP	Local Binary Pattern	DNN	Deep Neural Network
Conv layers	Convolutional Layers	PCA	Principal Component Analysis	TF Analysis	Time-frequency
SGD	Stochastic Gradient Descent Optimizer	ViT	Vision Transformer	RMSprop	Root Mean Squared Propagation Optimizer

encoding spatial relationships is important, which cannot be captured in 1D feature vectors [1]. Signals acquired at high sampling rates yield sizable datasets that can strain computational resources, often requiring dimensionality reduction before deploying machine learning models. Hand-crafting feature representations from complex morphology of physiological signals is challenging. Manually designed features could be sub-optimal or omit useful discriminative information [2]. The spectral information and morphology of these physiological signals dynamically vary over time. In applications where real-time processing is essential, dealing with non-stationarity signal becomes critical. Delayed or outdated information due to non-stationarity can impact the effectiveness of real-time monitoring and decision-making [3], [4]. This necessitates specialized time-frequency (TF) analysis instead of static frequency analysis. Innovative strategies like signal-to-image conversion coupled with deep learning models like Convolutional Neural Networks (CNNs) that can automatically learn robust features from 2D representations of raw physiological data can be used to work around these constraints.

Raw signals can be converted directly to images, bypassing manual preprocessing. The preprocessing and feature extraction process in biomedical time series signals is complex, involving various features from different domains like spatial, temporal, and frequency. Unlike traditional approaches that necessitate laborious manual extraction of optimal features from complex time series data, CNNs autonomously learn crucial features directly from raw data [5], [6].

Analysis of full image representations prevents potential loss of discriminative information with pre-selected 1D features [7]. Large physiological datasets also become more tractable as images. Additionally, visual inspection of generated 2D representations provides intuitiveness. Several innovative signal-to-image (STI) transformation techniques have been proposed to address these challenges. These include TF representations such as scalograms, recurrence plots, spectrograms [3], [8], [9], [10], [11] and spatial feature representations such as azimuthal equidistant projection (AEP) [12] and common special patterns (CSP) [13], [14], [15]. The resulting 2D images can capture spectral, spatial, and temporal characteristics depending on the method, providing CNN models with richer information for feature learning. CNNs can then automatically learn hierarchical features from 2D visualizations of raw data, overcoming hand-crafted feature engineering. CNNs have shown remarkable performance on various signal-processing tasks [16], [17], [18], [19]. The convolutional kernels in 2D CNNs are especially suited for extracting spatial patterns and relationships from 2D imagery, which is valuable for encoded multi-channel physiological data. Minimal preprocessing is required since CNNs have the ability to learn from raw, visualized signals in an end-to-end manner [20]. The flowchart in Figure 1 offers a technical overview of signal acquisition-to-classification. 2D CNNs have achieved state-of-the-art results on various physiological classification tasks like arrhythmia detection [21], [22], seizure detection [23] and motor decoding [24] when applied to transformed signal images. Recent trends depicted in Figure 2 underscore the escalating utilization of CNN methods for autonomously acquiring robust features from raw physiological signals.

This Review Paper synthesizes key insights from the literature regarding the relative merits of different signal transformations, CNN architectures, and training procedures for EEG, EMG, and ECG signal processing. By providing a holistic overview of various methods and comparing their results based on metrics like accuracy and average accuracy, where accuracy represents the classification performance for a specific case, while average accuracy computes the mean classification performance by averaging the accuracies across multiple cases or conditions, this paper aims to catalyze continued innovations in the design of effective end-to-end systems for extracting clinically relevant information from multidimensional physiological data.

A. PHYSIOLOGICAL SIGNALS

Biomedical signals encompass diverse physiological data and play a pivotal role in understanding human health and function. These signals provide valuable insights into the complex workings of the human body and are extensively

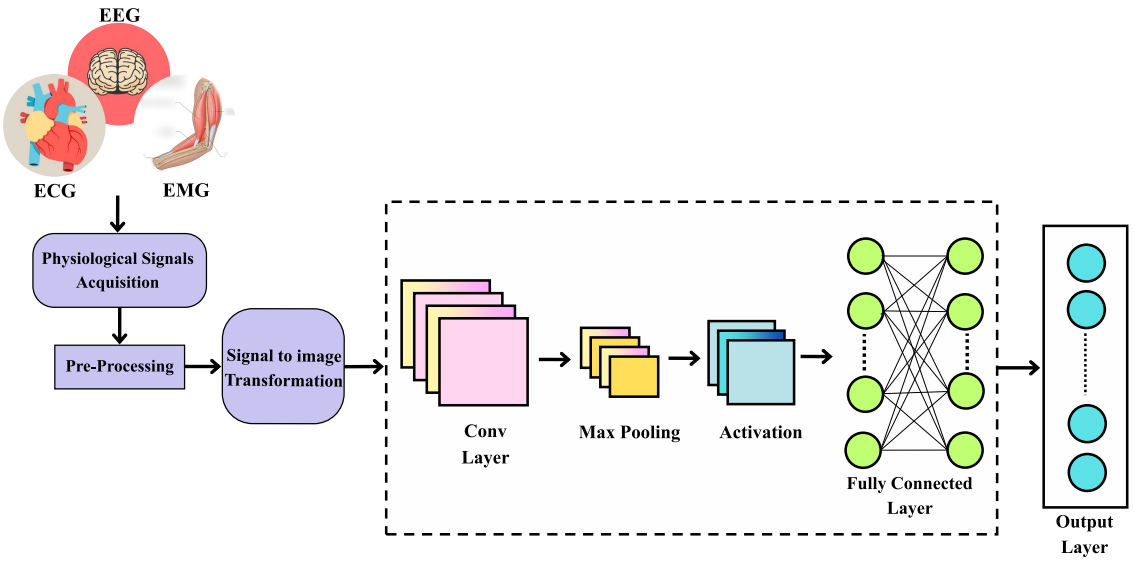


FIGURE 1. Basic CNN framework utilized in physiological signal processing.

used in medical diagnosis, treatment, and research. Among the various types of biomedical signals, electroencephalogram (EEG), electromyogram (EMG), and electrocardiogram (ECG) are of particular significance due to their ability to capture crucial information related to the brain, muscles, and heart, respectively.

Electroencephalography (EEG) is a non-invasive neuroimaging technique that records electrical activity generated by the brain. It measures the voltage fluctuations resulting from ionic current flowing within the neurons of the brain. EEG is widely used in neuroscience and clinical applications to study brain function and cognitive processes and diagnose neurological disorders. EEG signals are complex and dynamic, comprising spatial, spectral, and temporal patterns, which present challenges in their analysis [25].

Electromyography (EMG), on the other hand, focuses on the electrical activity produced by muscles. EMG signals are acquired by placing electrodes on the skin surface or directly within the muscles to measure the electrical potential generated during muscle contractions. EMG is utilized in various applications, including understanding motor control [26] diagnosing neuromuscular disorders [27], and developing assistive technologies such as prosthetic control and gesture recognition systems [28].

Electrocardiography (ECG) is a fundamental tool in cardiology, and it is used to record the electrical activity of the heart over time. It measures the depolarization and repolarization of the cardiac muscle, providing critical information about heart rate, rhythm, and overall cardiac health. ECG is essential for diagnosing and monitoring arrhythmias [29], ischemic heart disease [30], and other cardiac abnormalities.

Despite the invaluable information in EEG, EMG, and ECG signals, their analysis poses significant challenges. Traditional machine learning techniques often rely on handcrafted features extracted from fixed analysis windows, which may not fully capture the rich and discriminative patterns present in these multidimensional signals. However, deep learning methods, particularly CNNs, have emerged as a promising approach for automatically learning informative representations from raw physiological data in recent years.

B. CONVOLUTIONAL NEURAL NETWORKS (CNN)

Convolutional neural networks (CNNs) are a specialized class of artificial neural networks that have proven highly effective for processing physiological signal data transformed into images or spectra [6]. The unique architecture of CNNs is designed to take advantage of the 2D structure of such converted signals to automatically learn local spatial features and patterns through convolutional filters.

A core building block of CNNs is the convolutional layer (Conv layer), where filters (small matrix of weights) are convolved across the input image to produce feature maps that encode local motifs in the data. By stacking multiple Conv layers, the network can extract hierarchical feature representations, capturing higher-level concepts derived from lower-level features [31], [32], [33].

CNNs also incorporate pooling layers, typically max pooling, that down sample the feature maps to reduce computational requirements and enable the network to capture translational invariances. Pooling enhances robustness to small shifts in the input data [34], [35].

Later layers in the CNN architecture are often fully connected layers that integrate spatial information and perform high-level reasoning required for classification or regression tasks [36], [37], [38]. The network is trained via backpropagation to iteratively optimize the filter weights to minimize a loss function [39].

Various CNN architectures have been devised for physiological signal processing, differing in depth, filter sizes, strides, and other hyperparameters. Popular models based on

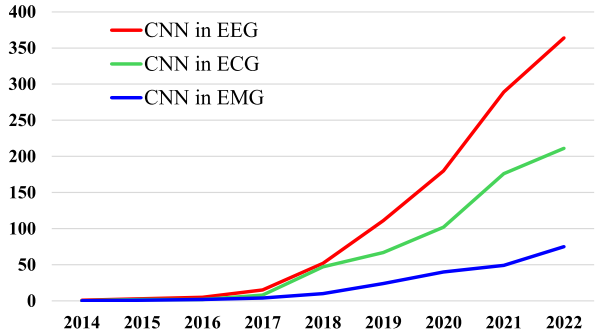


FIGURE 2. Recent trends in the number of published papers (according to PubMed).

transfer learning, pre-trained on large image datasets include AlexNet, VGGNet, ResNet, and Inception. Transfer learning is commonly employed to leverage these models for new tasks with limited training data, avoid overfitting, and reduce processing complexity [40]. Key advantages of CNNs for STI conversion approaches include automatic feature learning, reduced manual preprocessing, and extraction of spatial relationships from multi-channel or multi-lead physiological recordings. Limitations include large dataset requirements and long training times for deeper models [41].

1) HYBRID CNN MODELS

Hybrid models, combining CNNs with other architectures, have become powerful tools in diverse domains, including physiological signal processing. These models exploit CNNs' strengths in capturing spatial features from signal data alongside specialized modules to address specific challenges or temporal dependencies.

One notable hybrid model is CNN-SVM, merging CNNs' feature extraction with SVMs' discriminative power for classification tasks. By leveraging CNNs to learn relevant features and feeding them into an SVM classifier, CNN-SVM achieves robust performance in disease diagnosis and anomaly detection [42].

Another promising approach is the CNN-LSTM model, which combines CNNs with LSTM networks to handle sequential data. With LSTM layers integrated after convolutional layers, CNN-LSTM effectively captures temporal dependencies within physiological signal sequences, enabling tasks such as forecasting and event detection [43].

Recent advancements have introduced CNN-ViT models, integrating CNNs with Vision Transformer architectures. Unlike traditional CNNs, ViTs use self-attention mechanisms to capture global relationships, making them suitable for tasks requiring long-range dependencies and context understanding. By merging CNNs' spatial feature extraction with ViTs' global contextual information, CNN-ViT models exhibit enhanced performance in multi-modal physiological signal analysis and cross-domain learning. [44]

C. SIGNAL TO IMAGE TRANSFORMATION TECHNIQUES

In signal processing for physiological data, various transformative techniques have emerged as powerful tools to convert time-domain signals into image representations, enabling effective feature engineering for CNNs [45]. Although 1D feature extraction methods like time and frequency domain analysis are essential for understanding signal characteristics, they have inherent limitations. Time domain techniques, while effective at capturing statistical properties, often struggle to represent intricate signal details [8], [46], [47]. Meanwhile, the frequency domain enables spectral content analysis but lacks time localization, statistical features, averaging effects, and the ability to capture transients. To address more comprehensive representation, 2D feature extraction methods become crucial [48].

In this section, we present an overview of key feature representation methods. These include Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), Fast Fourier Transform (FFT), Welch's Method, Discrete Wavelet Transform (DWT), Empirical Mode Decomposition (EMD), Reshaping, and Power Spectral Density (PSD).

1) TIME DOMAIN FEATURE EXTRACTION TECHNIQUE

Time domain feature extraction represents one of the foundational methods in signal analysis. It offers simplicity and computational efficiency advantages, making it suitable for real-time applications. However, its reliance on statistical and morphological features can be limiting when dealing with complex signals or artifacts, necessitating more advanced techniques such as frequency domain analysis or mathematical transformations to capture intricate details accurately. Statistical methods like zero-crossing and autoregression were widely employed to extract important features from the biomedical signals [49].

2) FREQUENCY DOMAIN FEATURE EXTRACTION TECHNIQUES

Frequency domain analysis provides valuable insights into signal characteristics and aids in feature extraction, offering distinct advantages when preparing physiological data for classification using CNNs. It uncovers spectral information, aids in noise reduction, detects periodic patterns and extracts complex features that enrich the input data. Additionally, it enables hardware optimization, enhances visualization, and facilitates the identification of sharp transitions, collectively enhancing CNN's performance in classifying intricate physiological signals compared to time domain methods. In biomedical signal analysis, the Fourier Transform is the most often utilized transformation [50], [51].

a: FAST FOURIER TRANSFORM (FFT)

FFT is an algorithm that efficiently implements the Discrete Fourier Transform (DFT) to obtain frequency information. The spectral information computed via FFT can be reshaped into a matrix and represented as a grayscale image. When providing physiological signals as input for CNNs, it is critical to consider specific constraints associated with approaches such as DFT/FFT, which frequently do not incorporate all

TABLE 2. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
The paper's keywords should include "CNN" or "convolutional neural network" [OR] "Machine / Deep Learning" [AND] "Signal-to-Image transformation" [AND] "Physiological/Biomedical signal" [OR] "EEG / ECG / EMG".	Non-peer-reviewed journal papers and conference papers are [NOT] considered.
The studies should apply 2D-CNN to the analysis of physiological signals in the context of specific tasks such as seizure detection, motor imagery classification, emotion recognition, gesture recognition, arrhythmia detection, heart-beat classification, and related areas within the biomedical domain.	Papers that did not use either "Convolutional Neural Networks" [OR] "Convolutional Neural Networks in combination with other ML / DL algorithms".
A database of Journal Papers published only between 2019 and 2023 from the Google Scholar portal was curated.	Studies employing 1D-CNN and other one-dimensional deep learning algorithms exclusively were excluded from the review because 1D algorithms do not accept image inputs and, therefore, do not involve the utilization of 'Signal to Image' conversion techniques.

time-domain samples when implementing their mathematical operations. This omission could lead to the loss of critical information in the signal. Furthermore, because its foundational functions (sine and cosine) lack the ability to reliably capture non-stationary aspects of complex biomedical data, DFT/FFT fails to give the level of temporal detail required for analyzing non-stationary physiological signals [52].

DFT computation transforms any finite signal, represented by N samples (x_0, x_1, x_2, …, x_{N-1}), into frequency domain samples. Each DFT coefficient (X_k) encapsulates both the amplitude and phase information of its corresponding signal component (x_n) at a specific discrete frequency (k).

$$FFT = \sum_{n=0}^{n-1} x_n e^{-j2\pi kn/N}$$

b: POWER SPECTRAL DENSITY (PSD)

Power spectral density estimation is a non-parametric analysis method that indicates the strength of variations as a function of frequency. PSD for multiple-channel signals can be computed using techniques like the periodogram, which utilizes the FFT, among other methods.

$$PSD = |FFT(x(t))|^2 \quad [53]$$

c: WELCH'S METHOD

Welch's method is a prevalent technique in spectral density estimation that offers improved approaches to the traditional periodogram method for estimating spectral density parameters in the frequency domain [54].

Welch's method averages periodograms of overlapping windowed signals to obtain a power spectral density (PSD) estimate. The PSD matrix can be scaled to grayscale pixel intensities to produce an image.

The signal $x_i(m)$ is segmented into K overlapped segments of size N with an overlap parameter D. Utilizing Welch's method, a modified periodogram is computed using Fast Fourier Transform with a windowing function $w(m)$ to mitigate signal discontinuities, and the Power Spectral Density $P(f)$ is obtained by averaging the resulting periodograms [55].

$$x_i(m) = x_i(m + iD), \ where\ m = 0, \ldots .N-1 \quad and\ i = 0, \ldots .K-1$$

$$P(f) = \frac{1}{K}\sum_{i=0}^{K=1}\frac{1}{NU}\left|\sum_{m=0}^{N-1} w(m)\cdot x_i(m)\cdot e^{-j2\pi fm}\right|^2$$

$$U = \frac{1}{N}\sum_{m=0}^{N-1}(\omega(m))^2$$

While spectral estimation techniques are known for their resistance to noise and quantization effects, they have limitations when accurately estimating the densities of instantaneous frequency components. This is due to their finite windowed approach, which may not capture the non-stationarity of biomedical signals accurately. Instead, they provide a smoothed representation of frequency content within each window, which can obscure the fine details of instantaneous frequency changes.

Thus, frequency-domain methods like the DFT/FFT are not well-suited for capturing instantaneous frequency information in non-stationary signals. Additionally, spectral estimation techniques, such as Welch's method, may introduce an averaging effect, making them less effective in analyzing dynamic features. These limitations can hinder the accurate analysis of complex, time-varying physiological signals, emphasizing the need for complementary methods like joint TF techniques [56].

3) JOINT TIME-FREQUENCY DOMAIN FEATURE EXTRACTION TECHNIQUES

Time-frequency domain techniques transform one-dimensional time series data into two-dimensional image representations, simultaneously incorporating time localization information, statistical characteristics, and spectral features. TF analysis combines elements of both time domain and frequency domain representations allowing for the simultaneous examination of how a signal's frequency content changes over time. This combined representation enhances

our ability to capture intricate details within non-stationary biomedical signals. This transformation is quite advantageous when inputting data into CNNs. These localized 2D images can help CNNs learn to recognize complex characteristics that would not be visible using traditional time or frequency domain methods. This conversion to 2D images gives a visual representation of the data that aids CNNs in identifying patterns and structures.

Widely employed methods, such as STFT and Wavelet Transform, are indispensable for analyzing non-stationary signals, where traditional time and frequency domain methods fall short [57]. These approaches enable us to capture subtle signal characteristics often concealed in overlapping regions of multiple signal components. STFT, for instance, provides a local analysis framework by employing a moving-time window, but its fixed window width presents resolution challenges. On the other hand, Wavelet Transform offers a unique approach with varying window widths and the ability to sparsely represent data. Methods like Wavelet Packet Transform (WPT) have been developed to enhance feature extraction and signal analysis, enabling effective handling of non-stationary biomedical signals with various applications in signal processing, classification, and denoising.

a: SHORT-TIME FOURIER TRANSFORM (STFT)

STFT is a crucial technique to study the TF properties of signals. It entails applying a signal to a fixed-length time-frame, and calculating its FFT inside each of these windows. This imposes a compromise between time and frequency resolutions. It enables the analysis of the frequency content of brief time intervals. Longer windows provide greater frequency localization but inferior time localization, whereas shorter windows offer better time localization but inferior frequency localization [58]. In order to analyze non-stationary signals, STFT divides the signal into frames and uses a moving-time window for analysis. The magnitude of the STFT is used to create a spectrogram image.

The fixed window width of STFT allows for faster processing compared to other TF analysis methods. STFT is computationally efficient and can provide results quickly, making it a preference for real-time applications involving biomedical signals [59].

STFT can be either narrowband or wideband, depending on the windowing function selected. The uncertainty principle prevents it from producing an accurate Time-frequency representation and makes it easier to identify frequency intervals present during particular time intervals. However, its set window width restricts the extent to which it can capture all non-stationary properties, making it appropriate for unimodal, univariate signals with little noise and few complex components [52].

$$STFT\left\{x\left(t\right)\right\} = X\left(\tau,\omega\right) = \int_{-\infty}^{\infty} x\left(t\right)\cdot\omega\left(t-\tau\right)e^{-jwt}dt$$

where $\omega(t)$ is the analysis window and τ is a short time interval.

b: DISCRETE WAVELET TRANSFORM (DWT)

The wavelet transform has two primary variations, depending on whether orthogonal or non-orthogonal wavelets are employed as basis functions. The DWT is particularly noteworthy, as it decomposes signals into functions that are orthogonal with respect to both translation and scaling. This key feature has led to its extensive application in tasks such as denoising, signal processing, and data compression [60].

It excels at capturing transient features from non-stationary signals and provides a sparse representation that reduces noise interference. While it can effectively characterize both high- and low-frequency components, some significant downsides must be considered, including the lack of a standardized technique and the risk of information loss during inappropriate decomposition [61].

$$y\left(n\right) = \sum_{k=1}^{\frac{N}{2^j}} a_j\left(k\right)\phi_{j,k}\left(n\right) + \sum_{j=1}^{j}\sum_{k=1}^{\frac{N}{2^j}} d_j\left(k\right)\psi_{j,k}\left(n\right)$$
$$a_j = \left[y\left(n\right),\varphi_{j,k}\left(n\right)\right]$$
$$d_j = \left[y\left(n\right),\Psi_{j,k}\left(n\right)\right]$$

$\phi_{j,k}(n)$ and $\psi_{j,k}(n)$ represent the scaling and wavelet functions. J signifies the wavelet decomposition series, and N is the total number of coefficients. The approximate coefficient part is $a_j(k)$, and the detailed coefficient part is $d_j(k)$, expressed in the equation.

c: CONTINUOUS WAVELET TRANSFORM (CWT)

In CWT, the signal is multiplied with wavelet functions localized in both time and frequency, yielding wavelet coefficients that represent TF information. These coefficients can be arranged into a representation known as a scalogram that provides a visual representation of the way in which the frequency content of the signal changes over time, making it an instrumental approach for TF analysis. CWT utilizes non-orthogonal wavelets to offer variable resolution, potentially enhancing input representations for CNN models in biomedical applications [62].

$$W_{(a,b)}\left[y\left(t\right)\right] = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} y\left(t\right)\phi^{*}\left(\frac{t-b}{a}\right)dt$$

The CWT is a linear transformation to the time-domain signal y(t) stated above. Where φ(t−b/a) represents a basis wavelet function φ(t) that has been scaled and shifted. The scaling parameter a>0, controls the function's spread, while b represents the time shift parameter or the time instant indicating the moment for signal analysis.

The CWT holds significant advantages in signal analysis and differentiates itself by accounting for negative frequencies, a feature absents in the FFT, resulting in superior frequency resolution. CWT's adaptable window width excels

in capturing the nuances of non-stationary signals. Additionally, it offers flexible TF localization tailored to specific needs. CWT breaks down signals into manageable wavelet components, enabling in-depth analysis and potential signal dilation and compression. Nevertheless, its highly correlated output values may impact signal classification accuracy, especially at higher frequencies than the STFT [49], [63].

Despite the fact that CWT stands out as an excellent method for converting signals to images, it's essential to recognize that CWT can demand relatively substantial computational resources when dealing with intricate signals, particularly in large-scale or real-time signal processing scenarios. Nonetheless, CWT excels in converting signals to images, effectively capturing intricate details in signals characterized by non-stationary or varying frequency content [59].

d: RECURRENCE PLOT

A recurrence plot is a 2D representation of the recurrence of states in a dynamic system. It was originally proposed by [64]. It is a graphical technique to visualize and analyze recurring patterns and non-stationarity in time series data [65].

$$R_{i,j}^{m,\varepsilon_i} = \Theta\left(\varepsilon_i - \|\vec{x}_i - \vec{x}_j\|\right), \vec{x}_i \in R^m, i, j = 1 \ldots..N$$

where, $R_{i,j}$ is the recurrence matrix, with i and j being time indices. X_i and X_j are reconstructed phase space vectors at times i and j. $||$. $||$ is a norm, usually the Euclidean norm. ε is a threshold distance. Θ is the Heaviside step function: $\Theta(x) = 0$ if $x < 0$, and 1 otherwise.

When dealing with a signal that exhibits specific recurrent patterns or long-term dependencies, RPs may be more effective at capturing these features. On the other hand, if the signal components manifest as transient events with varying frequencies, CWT can be more effective at capturing such patterns.

4) DECOMPOSITION AND SPARSE DOMAIN FEATURE EXTRACTION

In contrast to all the previous methods, Decomposition and Sparse Domain Feature extraction techniques offer distinct advantages, providing a more localized and interpretable approach to feature extraction.

a: EMPIRICAL MODE DECOMPOSITION (EMD)

EMD decomposes a time series signal into a set of oscillatory components called Intrinsic Mode Functions (IMFs). Each IMF represents a specific oscillatory mode or component present in the original time series [66]. EMD was extended to its multivariate version, known as Multivariate EMD (MEMD), which is capable of deconstructing multi-channel signals to overcome the mode-mixing in the IMFs, particularly while analyzing signals with closely spaced frequencies and measurement noise [67].

These IMFs are one-dimensional signals, not images, but can be transformed by applying mathematical operations such as DE and Hilbert transform to extract features that can be further stacked to input a 2D CNN. EMD excels at separating complex signals into their constituent components or modes, making it easier to isolate and analyze specific features within a signal. Additionally, it is robust to noise, reduces dimensionality, and offers customization options for different signal types and applications [68].

$$X(t) = \sum_{i=1}^{N} I_i(t) + R_N(t)$$

where N is the total number of IMFs, $R_N(t)$ is the residue signal and $I_i(t)$ is the i^{th} order IMF.

5) RESHAPING

Reshaping refers to reorganizing the 1D time series signal into a 2D matrix by dividing time samples into rows and channels into columns. This generates an image representing multidimensional spatial-temporal information and simplifies the complex computing procedure [69].

$$X_{image} = reshape(X_{signal}, [N_{channels}, N_{samples}])$$

II. CNN AND SIGNAL-TO-IMAGE CONVERSION IN EEG APPLICATION DOMAINS

Electroencephalography (EEG) is a pivotal tool in neuroscience, with applications ranging from Brain-Computer Interfaces (BCIs) to commercial domains. Traditionally, EEG analysis relied on machine learning for neural classification and neuroimaging. Recent advances, including the proliferation of EEG datasets and the rise of Convolutional Neural Networks (CNNs), mark a new era. 2D CNNs play a central role, offering sophistication and automatic, context-rich EEG signal classification. They capture intricate EEG patterns, revealing nuances in brain functionality. Signal-to-image conversion is equally crucial, transforming raw EEG data into structured 2D representations, enabling deeper insight and bridging the gap between complex neural data and practical applications. This integration redefines EEG analysis, making it more accessible and less dependent on specialized expertise [70].

Figure 3 outlines the datasets used in the EEG analysis. Among these are well-established public datasets such as DEAP, SEED, and BCIC. Others represent 40%, while smaller datasets contribute to the remaining 27%.

A. SEIZURE DETECTION

Seizures are abnormal and sudden bursts of electrical activity in the brain, which can result in various symptoms and behaviors. EEG signals provide valuable insights into the patterns of electrical activity in the brain and are a crucial tool for diagnosing and managing seizure disorders. Seizure EEG signals are commonly categorized into interictal, preictal, and ictal phases. An interictal signal is the EEG data recorded between seizures, representing baseline brain activity. In contrast, the preictal signal is the EEG data recorded immediately before a seizure, leading up to its onset. The ictal signal is the EEG

data recorded during an active seizure event, capturing the seizure dynamics. The ictal signals can be further categorized as focal and non-focal EEG signals, where Focal (Partial) Seizure EEG Signals originate in a specific brain region, with localized abnormal activity in EEG, and Non-Focal (Generalized) Seizure EEG signals involve widespread and symmetric activity, affecting both hemispheres from the start, which help in diagnosing focal and non-focal epilepsy in patients [71], [72].

CNN has been found to be ideal for image-based classification due to its self-feature learning capability and excellent classification results on multi-class classification problems [6]. CNNs are being widely used for the application of seizure detection and prediction with various STI conversion techniques proposed in recent studies [73]. Reference [74] accomplished the conversion of a time domain EEG signal to an image by extracting EEG signal features such as correlation coefficient, STFT (spectrogram), and mutual information. The image representations of the EEG waveform were used to train AlexNet, resulting in better performance with 99.33% accuracy using STFT compared to statistical transforms such as correlation and mutual information. Reference [75] also employed STFT on EEG signals to generate 3-channel RGB spectrogram images, which were subsequently utilized as input for a 3-layer CNN architecture for classifying normal, preictal, and seizure states, achieving an accuracy of 98.22%. Another study that uses STFT to represent epileptic EEG signals as 2D spectrograms is [73], in which the spectrograms from 19 channels of EEG were stacked to form a single input image. This image was passed through 10 pre-trained CNNs for transfer learning-based feature extraction. The extracted image features were then classified into 8 groups representing seizure types using a support vector machine (SVM) classifier. The study reported the highest classification accuracy of 88.3% when using image features from the Inception-V3 network, classified by the SVM model. With 48 layers, Inception-V3 is one of the deepest models evaluated in the study. The increased depth enables learning more complex feature representations needed to distinguish different seizure types.

STFT is a commonly used method for representing signals as images. However, its fixed resolution can limit its effectiveness in representing biomedical signals with non-stationary or varying frequency content. In contrast, CWT offers variable resolution, making it a potentially better choice for providing improved input representations to CNN models in biomedical applications. For instance, in [76], the authors compared STFT spectrograms and CWT scalograms for EEG signal classification in seizure detection and found that CWT scalograms outperformed STFT. Their study proposed five classifiers, and the fine-tuned VGG16 model with CWT STI conversion demonstrated superior performance with an accuracy of 99.21% compared to 98.94% by STFT. Similarly, in [77], CWT was employed for binary to five class classification of seizure EEG signals, using a 2-layer CNN classifier, and yielded successful results, including an accuracy of 99.5% for Normal/Seizure classification, 98.5% for epileptogenic/seizure classification, and 99.0% and 93.6% accuracy for three-class eyes-open/hippocampus/seizure and five-class eyes-open/eyes-closed/ epileptogenic/hippocampus/seizure classifications, respectively. On the other hand, [78] introduced a Multi-Channel Vision Transformer (MViT), a transformative deep learning architecture designed for seizure prediction using EEG data. Operating on multiple channels simultaneously, MViT leveraged datasets such as CHB-MIT Scalp EEG, Kaggle/AES Invasive EEG, and Kaggle/Melbourne University Invasive EEG, demonstrating superior performance compared to state-of-the-art methods. The model incorporated an efficient CWT-based pre-processing step, converting EEG signals into scalograms. With preictal and interictal EEG activities as output classes, MViT aims to predict seizures. Across diverse EEG datasets, MViT showcases outstanding results, achieving a notable 99.8% sensitivity, 99.7% specificity, and 99.8% accuracy on the CHB-MIT Scalp EEG dataset. Furthermore, on the Kaggle/AES Invasive EEG dataset, the model attained 90.28% sensitivity with AUC values of 0.940 and 0.885 on public and private test sets, respectively. Similarly, on the Kaggle/Melbourne University Invasive EEG dataset, MViT achieved a sensitivity of 91.15% and a notable AUC of 0.924. The proposed MViT method has the capability to precisely and promptly predict upcoming seizure onsets, offering patients the chance to promptly administer fast-acting medications and implement safety measures in times of heightened susceptibility to seizures. These findings highlight the advantages of CWT in transforming biomedical signals into image representations for CNN-based classification tasks.

Although the CWT stands out as an advantageous method for converting signals to images due to its variable resolution, effectively capturing fine details in signals with non-stationary or varying frequency content, it is worth noting that CWT's computational intensity may pose challenges in large-scale or real-time signal processing tasks. Therefore, STI conversion using DWT is another common approach. In another study by [79], the authors proposed using DWT to decompose seizure EEG signals into various frequency bands. They applied the DWT-transformed images as input to a 2-layer CNN, successfully classifying EEG signals into focal vs. non-focal categories, as well as distinguishing among three classes: Healthy, InterIctal, Ictal signals with accuracies of 99.70% and 98.89%, respectively. In [80], PSD analysis was utilized to generate 1D PSD curves for each EEG channel, which were subsequently integrated over frequency bands to form 1D arrays of aggregated PSD values. These arrays were then combined channel-wise into a 2D array representing the Power Spectral Density Energy Diagrams (PSDEDs). These PSDEDs were applied to pre-trained DCNN models (Inception-ResNet-v2, Inception-v3, and ResNet152), fine-tuned for feature extraction with Online Hard Example Mining (OHEM) loss function. The method successfully distinguished four epileptic states:

TABLE 3. A summary of Signal-to-image transformations and CNNs in Seizure detection applications.

Ref.	Dataset	STI transform method	Classifier	Output classes	Results (%)
[74]	BONN University	STFT, Correlation Coefficient, Mutual information	AlexNet	3	Accuracies: - (3 classes): healthy/pre-ictal/seizure - STFT = 99.33, - Correlation Coefficient = 95.33 - (2 classes): healthy/normal vs seizure - Mutual Information = 97.50
[75]	BONN University	STFT	3-layer CNN	3	Average Accuracy: - (3 classes): healthy/preictal/ictal = 98.22
[73]	TUH EEG	STFT	Combination of InceptionV3 feature extractor and SVM classifier	8	Accuracy - (8 classes): SP/CP/FN/GN/AB/TN/TS/NS = 88.30
[76]	BONN University	STFT, CWT	Fine-tuned VGG16	2	Average Accuracies - (2 classes) - Normal/Inter-Ictal: CWT = 99.21 STFT = 98.94
[77]	BONN University	CWT	2-layer CNN	2, 2, 3, 5	Accuracies: BONN University Dataset (5 classes) - eyesOpen/eyesClosed/epileptogenic/hippocampus/seizure: - (2 classes) - eyes-closed/seizure= 99.5 - (2 classes) - epileptogenic/seizure = 98.5 - (3 classes) - eyes-open/hippocampus/seizure = 99 - (5 classes) - eyes-open/eyes-closed/epileptogenic/hippocampus/seizure = 93.6
[78]	- CHB-MIT Scalp EEG, - Kaggle/AES Invasive EEG, - Kaggle/Melbourne University Invasive EEG	CWT-	Multi-Channel Vision Transformer (MViT)	2	(2 classes) - Preictal/Interictal- -CHB-MIT Scalp EEG = 99.8 Accuracy -Kaggle/AES Invasive EEG = 90.28% sensitivity -Kaggle/MelbourneUniversity Invasive EEG = 91.15
[80]	CHB-MIT	PSD	EESC (combination of Inception-ResNet-v2, Inception-v3, and ResNet152)	4	Accuracy - (4 classes) - Interictal/Preictal-I/Preictal-II /Seizure = 92.6
[13]	CHB-MIT	CSP	2-layer CNN	2	Accuracy: - (2 classes) - Inter-ictal/Preictal= 90
[79]	Bonn EEG, Bern-Barcelona	DWT	2-layer CNN	3, 2	Accuracies: - (3 classes) - Healthy/InterIctal/Ictal = 98.89 - (2 classes) - Non-focal/focal = 99.70
[81]	Freiburg EEG	DTF	6-layer CNN	2	- (2 classes) - Interictal/Preictal Sensitivity = 90.8
[82]	CHB-MIT, AES	Patch Embedding (In patch embedding, the 1D data is divided into fixed-size segments or patches, and each patch is then embedded into a higher-dimensional space. These embeddings are arranged in a grid-like structure, resembling a 2D image.)	HViT-DUL (Hybrid Visual Transformer-Data Uncertainty Learning)	2,2	Sensitivity: CHB (2 classes) - Seizure / Non-Seizure = 87.9 ± 2.2 AES (2 classes) - preictal/interictal = 78.9 ± 5.3
[83]	CHB-MIT	Multi-channel EEG signals from all patients were segmented with a 25% overlap, forming 3D matrices where each segment is represented as a 2D matrix with dimensions for channels and sequence length.	Lightweight Convolution Transformer (LCT)	2	Accuracy: (2 classes) - Ictal/Interictal = 96.31
[84]	CHB-MIT	Sequential signals from multiple channels were mapped to be interpreted as 2D images.	Temporal multi-channel vision transformers (TMC-ViT), MLP, CNN+Bi-LSTM, CNN, TMC-T	2	Accuracies: (2 classes) - preictal/interictal: - TMC-ViT = 95.73 ± 3.56 - MLP = 75.02 ± 11.87 - CNN+Bi-LSTM = 92.07 ± 7.09 - CNN = 95.59 ± 4.36 - TMC-T = 93.74 ± 5.45

interictal, preictal I, preictal II, and seizure, with an accuracy of 92.6%.

References [13] and [81] particularly focus on seizure prediction by distinguishing between interictal and preictal EEG signals. In [13], the authors proposed Common spatial patterns (CSP), a spatial filtering technique commonly used for feature extraction in motor imagery-based brain-computer interfaces for STI conversion. The EEG signal was decomposed using wavelet packet decomposition into 8 frequency sub-bands. Then CSP was applied to each frequency sub-band and original signal to maximize the variance between classes. The resulting feature matrix was fed to a shallow 2-layer CNN designed to distinguish between interictal and preictal seizure EEG and attained an accuracy of 90% and sensitivity of 92.2%. Reference [81] utilizes the Directed Transfer Function (DTF), a method known for analyzing the directed flow of information between different brain regions in EEG data by quantifying the causal relationships between different brain regions, providing insights into brain network interactions. This study applied DTF to extract connectivity features between intracranial EEG channels and creates 2D channel-frequency maps of DTF features, which serve as input for a 6-layer CNN; the model successfully classifies preictal and interictal seizure EEG with a sensitivity of 90.8%. While DTF is valuable for investigating the connectivity patterns in the brain and CSP is effective in classifying motor imagery tasks, further study is required to assess the direct applicability of these techniques for STI conversion of seizure EEG signal.

Patch embedding is another unique approach in ViT-based methods, converting and mapping one-dimensional signals into grid-like structures resembling images, enabling ViTs to leverage self-attention for capturing dependencies. Reference [82] used a CHB-MIT dataset comprising 844 hours of pediatric scalp EEG signals, including 182 seizure events. Additionally, the AES dataset involves 627.7 hours of intracranial EEG recordings with 48 seizure events from dogs and patients, with varied electrode configurations and sampling frequencies. The HViT takes raw EEG segments as input, representing them as a 3D matrix with dimensions $T \times N$, where T is 5 seconds, and N is the number of electrodes. The patch embedding module enhances local feature perception by concatenating features from large and small convolutions with different kernels. The resulting feature maps are then processed by the C2T (CNN-to-Transformer) module, which includes a lite bottleneck block and a transformer module with a separable multi-headed self-attention mechanism. The HViT-DUL (Hybrid Visual Transformer-Data Uncertainty Learning) demonstrates superior performance compared to all baseline models, achieving an AUC of 0.937 ± 0.004 and 0.889 ± 0.004, sensitivity of $87.9 \pm 2.2\%$ and $78.9 \pm 5.3\%$, and FPR of 0.056 ± 0.006/h and 0.049 ± 0.008/h in CHB-MIT and AES (American Epilepsy Society) datasets, respectively. Notably, HViT-DUL significantly reduces FPR and enhances AUC compared to ViT, with a notable boost of 4.2% (0.899→0.937) and 6.7% (0.833→0.889) on CHB-MIT and AES datasets. While in [83], the CHB-MIT scalp EEG dataset, containing recordings from 24 pediatric patients with intractable seizure disorders, was utilized. 18 channels across all patients were selected, and the multi-channel EEG signals were segmented into overlapping segments of varying lengths, with a 25% overlap. Each segment is represented as a 2D matrix, where the dimensions are the number of channels and sequence length, and further formed into a 3D matrix with a new dimension representing the number of ictal or interictal segments. This 3D matrix serves as the input to the proposed Lightweight Convolution Transformer (LCT), which incorporated a convolution tokenizer instead of patches and attention-based pooling instead of a classification token. This enabled the framework to learn spatial and temporal correlated information simultaneously from multi-channel EEG signals. This capability helps identify high- and low-frequency features in ictal and interictal periods. The proposed LCT model achieved an accuracy of 96.31% on seizure detection in the cross-patient case at 0.5-second segment length. Additionally, the performance metrics showed that the inclusion of convolutional layers and attention-based pooling in the model enhances the performance and reduces the number of Transformer encoder layers, significantly reducing the computational complexity. Similarly, [84] explores definitions of preictal and interictal states in the CHB-MIT Scalp EEG Database, evaluating 30 and 60 minutes before seizures. The interictal state spans 4 hours after the last seizure and 4 hours before the next, with a 5-minute pre-seizure window for timely alerts. Two EEG sample sizes, 5-second and 20-second with a 5-second overlap, were tested for DL architectures. Sequential signals from multiple channels were mapped to be interpreted as 2D images. The TMC-ViT model used a CNN at the input for embeddings, reducing the input matrix to 21×21 dimensions compatible with 16×16 images and employing 3×3 patches. The CNN had 16, 32, 64, and 64 filters of dimensions 1×20, 1×20, 1×10, and 3×3, followed by batch normalization and max-pooling layers. While learnable embeddings handled position encoding, convolutional networks managed token embedding. With 4 attention heads and 8 Transformer layers, the model concluded with dense layers of dimensions 2,048 and 1,024. TMC-ViT demonstrated superior performance with the highest accuracy (95.73%), AUC (97.55%), and sensitivity (96.46%) when compared to other models like MLP (75.02 ± 11.87), CNN+Bi-LSTM (92.07 ± 7.09), CNN (95.59 ± 4.36), and TMC-T (93.74 ± 5.45) in terms of accuracy. Additionally, it exhibited the lowest standard deviation across patients for these metrics, showcasing the model's robustness in learning EEG signal nuances from diverse seizures and patients.

B. MOTOR IMAGERY DECODING

Motor imagery is a cognitive process in which an individual envisions performing a specific motor action without overt movement. Decoding motor imagery offers a unique

TABLE 4. A summary of signal-to-image transformations and CNNs in motor imagery decoding applications.

Ref.	Dataset	STI transform	Classifier	output classes	Results (%)
[86]	BCI Competition-IV (Dataset 2b)	STFT	VGG-16	2	Average Accuracy - (2 classes): Left-hand motor imagery/ Right-hand motor imagery= 74.2
[87]	BCI Competition-IV (Dataset 2a), HGD (High Gamma Dataset)	STFT	Channel-Projection Mixed-Scale CNN (CP-MixedNeT)	4 (BCI IV 2a), 4 (HGD)	Accuracies: - BCI IV 2a Dataset (4 classes): left/right/feet/ tongue= 74.6, - HGD (4 classes): left/right/feet/ rest = 93.7
[61]	BCI Competition-III (Dataset IV-A)	STFT, CWT	FT-AlexNet (5 layer)	2	Accuracies: - (2 classes): Right hand/Right foot - STFT = 98.7 - CWT = 99.35
[88]	BCI Competition-IV (Dataset 2a) BCI Competition-IV (Dataset 2b)	CWT	Multilevel and multiscale feature fusion convolutional neural network (MLMSFFCNN)	4 (dataset 2a), 2 (dataset 2b)	Accuracies: - BCI Competition-IV Dataset 2a - (4 classes): left hand/right hand/foot/ tongue = 92.95 - BCI Competition-IV Dataset 2b - (2 classes): Left hand/Right hand = 97.03
[89]	Physionet MI EEG Dataset	CWT (MW transform)	6-layer CNN	4	-(4 classes): Left fist/right fist/both fists/ both feet -94.5 Average accuracy across 10 subjects (original dataset) - 92.7 Average accuracy after adding 4 new subjects to the original dataset
[43]	BCI Competition-IV (Dataset 2a)	CWT (MW Transform)	Inception V3 + LSTM	4	Accuracy: -(4 classes): left hand/right hand/ feet /tongue = 92
[90]	BCI Competition-III (Datasets IV-a, IV-b, V) GigaDB	CWT	FT - ShuffleNet	2 (Dataset IV-a, IV-b), 3 (Dataset V), 2 (GigaDB)	- BCI Competition-III: - Dataset IV-a (2 classes): Right hand/Right foot = 99.52 - Dataset IV-b (2 classes): left hand/right foot = 99.52 - Dataset V (3 classes): left hand/right hand/word association = 97.77 - GigaDB dataset (2 classes) = 87.69 average accuracy across all subjects
[91]	BNCI Horizon 2020 website publicly available dataset with EEG and motion data for 6 upper limb movements	CWT	3-layer 3D-CNN	2	Average Accuracies: - Premovement vs Rest = 90.3
[92]	The dataset involved 20 healthy right-handed participants with an average age of 27.9 ± 2.9 years. Ethical approval for the study was obtained from the Monash University Human Research Ethics Committee (MUHREC)	CWT	Standard ViT, Residual ViT (ResViT), TWINS model, ResNet150	4	Average Accuracies: (4 classes): right/Left plantar flexion, left/Right dorsiflexion. - Standard ViT = 89.82 - ResViT = 93.39 - TWINS = 97.33 - ResNet150 = 96.03
[93]	- BCI 2A (dataset IIA from BCI competition 4), - BCI 2B (dataset IIB from BCI competition 4), - WEIBO2014.	Data was divided into segments and then randomly combined to make 2D while keeping time order.	EEGNet, ViT, Spatio-Temporal CNN + ViT (st-CViT)	2, 2, 2	Accuracies: (2 classes): Left-hand/right-hand motor imagery tasks - BCI IV 2a dataset: - EEGNet = 68.25 - ViT = 55.09 - Spatio-Temporal CNN + ViT = 80.44 - BCI IV 2b dataset: - EEGNet = 72.00 (nested LOSO) - ViT = 53.03 - Spatio-Temporal CNN + ViT = 74.73 (nested LOSO) - Weibo dataset: - EEGNet = Not mentioned in the study - ViT = 57.83 - Spatio-Temporal CNN + ViT = 78.44

window into understanding the underlying neural mechanisms involved in action planning, execution, and control. By harnessing physiological signals such as EEG and coupling them with advanced deep-learning architectures, researchers have endeavored to unravel the intricate relationship between mental representations of movement and corresponding brain activity patterns. In that regard, methods like STFT and CWT offer nuanced spectral-temporal

analysis, accommodating complex EEG patterns effectively, hence remaining more adaptable and scalable [85]. [86] converted Raw motor EEG signals to TF images using STFT, alpha (8-13 Hz), and beta (16-32Hz) frequency bands are extracted from the full spectrums, as these show discriminative patterns for different MI tasks. These spectral extracts were stacked and used for fine-tuning an ImageNet pre-trained VGG-16 CNN. The model attained 74.2% accuracy in classifying left vs. right-hand motor imagery tasks. Reference [87] used amplitude-perturbation data augmentation that includes the STFT algorithm to extract the spectral images of EEG recordings aided by a channel projection mixed-scale CNN (CP-MixedNet), which consists of three blocks: the primary spatial and temporal representation learning block, the mixed-scale convolutional block for capturing mixed-scale temporal information, and the classification block for classifying EEG tasks. The approach sought to enhance decoding accuracy by capturing spatial dependencies and information at various temporal scales. The framework was tested on two public EEG datasets. The BCI Competition IV 2a dataset, with four output classes (left hand, right hand, feet, tongue movements), achieved 74.6% accuracy. The High Gamma dataset, featuring four classes (left hand, right hand, feet, rest), yielded an accuracy of 93.7%.

While both STFT and CWT are commonly used for STI conversion of motor EEG, CWT offers distinct stationarity [59], multi-resolution analysis capturing diverse frequency bands, and flexibility in wavelet selection, making it a solid choice for dynamic neural signals such as motor EEG advantages like TF localization, adaptability to non-signals with complex characteristics. The superiority of CWT over STFT was investigated by [61]. The paper contrasts CWT scalograms and STFT spectrograms as 2D input representations for a Fine-Tuned AlexNet comprising 5 layers. The proposed CNN was tested on IV-A of BCI Competition-III with 2 classes (right hand, right foot), utilizing CWT scalograms as input, outperformed STFT, achieving an accuracy of 99.35% compared to STFT's 98.7%.

Reference [88] employed CWT and the Clough-Tocher (CT) interpolation algorithm for multidimensional MI-EEG imaging. Each electrode's time-frequency matrices were generated using the CWT. The feature matrices were interpolated to align with their corresponding coordinates, resulting in a Wavelet transform-based multidimensional image (WTMI) for which an MLMSFFCNN (multilevel and multiscale feature fusion CNN) was specifically designed, achieving a notable performance with an accuracy of 92.95% on the BCI Competition IV 2a dataset containing 4 classes (left hand, right hand, feet, tongue movements) and even higher accuracy of 97.03% on the BCI Competition IV 2b dataset containing 2 classes (left-hand, right-hand movement).

Among the assortment of wavelet functions used in the CWT framework, the Morlet Wavelet (MW) emerges as a standout as it offers a well-balanced combination of oscillatory frequency representation and time-domain localization, making it particularly advantageous for capturing dynamic signal patterns such as motor-related EEG signals making MW a popular choice for representing intricate temporal and spectral features in signals, especially when preparing input representations for CNNs. However, The MW might not be as effective when dealing with signals without clear oscillatory patterns. Additionally, choosing the wavelet's central frequency and scale could impact the result, requiring careful parameter tuning [94]. The applicability of MW CWT for STI transformation is investigated by [43] and [89].

In the research by [89], the EEG signal undergoes a transformation into an image using scout EEG source imaging (ESI) methodology. This methodology involves selecting a region of interest (ROI) within the motor cortex, which is divided into ten scouts. From the time series of these scouts, features are extracted using an MW CWT technique. The resultant TF maps for each scout are then employed as image inputs for a custom 6-layered CNN. Remarkably, scout R5 situated in the right motor cortex demonstrated the highest accuracy of 94.5% in classifying 4 classes, namely Left fist/right fist/both fists/ both feet on the Physionet database, with an average across ten subjects [43] also used CWT with MW, converting the signals into 2D images to input to CNN. Various CNN configurations were proposed, encompassing a customized two-block CNN combined with LSTM, ResNet50 with LSTM, and Inception V3 with LSTM. Among these, the latter demonstrated the most optimal performance, which was evaluated on BCI Competition IV dataset 2a with 4 classes (Left hand, right hand, both feet and tongue movements) achieving an accuracy of 92% [90] proposed a new motor EEG classification framework using 10 pre-trained CNNs like AlexNet, SqueezeNet, ShuffleNet, GoogLeNet, ResNet, DenseNet, MobileNetV2, InceptionV3, etc. Raw EEG signals were denoised with MSPCA and converted to 3D scalograms using CWT and stacking channels. The CNN models are implemented and fine-tuned via transfer learning on the scalograms. Comparative experiments use varying learning rates, optimizers, model sizes, and noisy vs denoised data. The best performance was obtained by the ShuffleNet model, coupled with the RMSprop optimizer, achieving an accuracy of 99.52% for the motor imagery datasets IV-a (Right hand/Right foot) and IV-b (left hand/right foot) and 97.77% accuracy for the multiclass mental imagery dataset V (left hand/right hand/word association) from BCI Competition III, and 87.69% accuracy for GigaDB dataset. Meanwhile, the [91] approach combines beamforming and CWT to generate 2D scalogram representations for each EEG source signal. These were then stacked into a 3D (time $\times$ frequency $\times$ source) matrix to retain spatial source location relationships for the 3-layer 3D DCNN. This method achieved an average accuracy of 90.3% for classification between movement preparation vs rest epochs. This indicates the model can reliably detect the intention to move from resting EEG. The approach also attained an average accuracy of 62.47% for classifying different movement preparation epochs. Although not high enough for practical use, this is above chance levels (50%), suggesting real differences

exist in the EEG during the preparation of different sub-movements. Meanwhile, [92] explored the application of ViT models to decode movement preparation from electroencephalography (EEG) signals. It involved 20 healthy right-handed participants with an average age of 27.9 ± 2.9 years. Ethical approval for the study was obtained from the Monash University Human Research Ethics Committee (MUHREC). The EEG signals were transformed into time-frequency maps using continuous wavelet transform and converted to RGB images as input for the ViT models. Three ViT variants were assessed: standard ViT, Residual ViT (ResViT), and the TWINS model, which combined the Pyramid Vision Transformer and Spatially Separable Vision Transformer. Additionally, a ResNet150 model was used for comparison. Among these approaches, the TWINS model demonstrated the highest performance, achieving 97.33% accuracy, 97.32% F1-score, 97.30% recall, and 97.36% precision in the four-class classification task. The results of this research illustrate the efficacy of employing a deep learning methodology utilizing EEG signals in advancing potential Brain-Machine Interfaces (BMI) for lower limb rehabilitation.

On the other hand, [93] investigated motor-imagery tasks using publicly available EEG data from four different datasets: WEIBO2014, Physionet, BCI 2A (dataset IIA from BCI competition 4), and BCI 2B (dataset IIB from BCI competition 4). The training data is divided into N_s segments and then randomly concatenated to transform the 1D data into 2D while maintaining the original time sequence. In this case, N_s was configured to be 3. Focused on left-hand and right-hand tasks, specific datasets were chosen, and continuous EEG data was divided into 4-second trials for each imagery task post-onset of mental imagery. The study evaluated various models, including CNNs, EEGNET, ViT, Spatial CNN + ViT, Temporal CNN + ViT, and Spatio-Temporal CNN + ViT (st-CViT). Performance assessment employed LOSO cross-validation, augmented by nested cross-validation for unbiased evaluation, especially with distinct subject-specific samples. The outer loop assessed the model, while the inner loop fine-tuned hyperparameters. Results showed that the Spatio-Temporal CNN + ViT model outperformed alternative models across BCI IV 2a (80.44%), 2b (74.73%), and Weibo datasets (78.44%), highlighting its potential for practical implementation in BCI applications.

C. EMOTION RECOGNITION

Emotion detection, an intriguing domain within affective computing, aims to discern and comprehend human emotional states. Numerous studies have attempted to unravel the complex connections between emotional states and patterns of cerebral activity by utilizing the potential of cutting-edge deep-learning architectures and merging them with EEG readings involving the analysis of the electrical activity of the brain. The DEAP dataset [95] is classified into two emotional states, valence and arousal, which are used to characterize emotions on a 2D scale. Valence refers to how positive or negative an emotion is. It ranges from unpleasant to pleasant, while Arousal refers to the intensity of the emotion and ranges from inactive (calm/bored) to active (excited/stimulated). The SEED dataset is classified into three emotional states: positive, negative, and neutral. It contains EEG and eye movement data, allowing for a more comprehensive analysis of emotion.

This section delves into the fundamentals of EEG-based emotion detection, exploring the transformation of neural signals into interpretable images and the role of CNNs in decoding emotions. Furthermore, we will examine key studies and innovative techniques adopted. One distinct onset for feature extraction from EEG signals involves PSD estimation through FFT. The resulting features can be subsequently mapped using various methods and techniques. Reference [96] introduced a novel method for extracting features from EEG signals, encompassing time domain features such as RAW (original amplitude) and NORM (normalized amplitude), as well as Power Spectral Density (PSD) features using FFT. This fusion of time and frequency features resulted in two combined feature sets: FREQRAW and FREQNORM. It was observed that the deep CNN models, particularly CVCNN, achieved the highest performance when employing the combined TF features, achieving 88.76% accuracy in Low vs High valence and 85.57% in Low vs High arousal binary classification tasks for the DEAP dataset. Similarly, [97] also used FFT to extract PSD features; the feature vectors computed from average power within each frequency band across EEG channels are subsequently mapped onto a 2D grid through Azimuthal-Equidistant Projection (AEP). The proposed combined CNN (to extract spatial features) and LSTM (to capture temporal variations) model achieved an accuracy of 90.62% for valence, 86.13% for arousal, 88.48% for dominance, and 86.23% for liking on the DEAP dataset. Meanwhile, [98] applied independent component analysis (ICA) to decompose EEG signals into distinct components and remove EOG and EMG components. They introduced offset variables following a Gaussian distribution for each EEG channel to address biased electrode coordinates and projected 3D coordinates to 2D using AEP to generate images representing energy distribution. After EEG feature extraction, the Clough-Tocher interpolation scheme estimates discrete energy. The GECNN (Graph-Embedded CNN) aims to extract both local CNN features and global functional features from EEG-based images. Local features were acquired via trunk and attention branches, while the global features were extracted using dynamic graph filtering. These extracted features are then fused together for emotion recognition. The proposed GECNN achieved commendable results on various datasets: On the SEED dataset (3 classes - positive, neutral, negative), subject-dependent accuracy reached 92.93%, subject-independent accuracy was 82.46%, with higher accuracy than PSD using differential entropy (DE) features. On the SDEA dataset (3 classes - neutral, funny, angry), subject-dependent accuracy peaked at 79.69% using PSD, and subject-independent accuracy

at 55.01%. Hilbert-Huang Spectrum features outperformed PSD. On the MPED dataset (7 classes - joy, funny, anger, sadness, fear, disgust, neutral), accuracy was 40.98%. On the DREAMER dataset (2 classes - high/low valence, high/low arousal), valence classification achieved 95.73% accuracy, while arousal classification reached 92.79% accuracy.

Out of various methods, DDE allows us to measure how the information content or uncertainty of signals evolves over time, providing a dynamic perspective in understanding and capturing not only frequency-domain features but also the time-domain variations in EEG signals, enhancing the distinction between different emotions. Reference [99] proposed EMD to obtain frequency features and Differential Entropy (DE) feature extraction for CNN input representation. The emotion EEG signal undergoes EMD to break down the signal into intrinsic mode functions (IMFs). This allows for extracting frequency information. Differential entropy (DE) features are calculated from each IMF to capture the frequency components. The DE features from all IMFs are concatenated into a TF feature vector called DDE. A custom 2-layer CNN then classifies the extracted DDE feature vectors into 2 classes (positive/negative emotions) with an accuracy of 97.56%. Reference [104] while DDE holds great promise, its full potential has yet to be fully explored in various applications. In contrast, STFT is a well-established and widely accepted technique with a proven track of high-frequency resolution and precise TF localization. As such, [100] computed EEG signals from the time domain to frequency using a 256-point STFT without overlapping the Hanning window of one second. Differential entropy (DE) images based on the Gaussian distribution are extracted from the EEG signals and fed to a 2-layered CNN for each frequency band. The FC (fully connected) features of the CNN are extracted and split into time intervals, allowing the model to learn the sequential information of the EEG features. Experiments on the SEED dataset, consisting of 62-channel EEG signals and 3 output classes (Positive, Negative, Neutral), yielded an accuracy of 90.41%. Likewise, [101] normalized the STFT outputs to transform these signals into image representations referred to as Electrode-Frequency distribution maps (EFDMs). These grayscale images, depicting the frequency distribution across various electrodes in the EEG signals, were then fed into a CNN featuring 4 residual blocks. The proposed model achieved an accuracy of 90.59% on the SEED dataset. Furthermore, fine-tuning the CNN pre-trained on SEED facilitated the learning of more emotion-related features in the DEAP dataset, resulting in an accuracy of 82.84%. Reference [102] utilized STFT and proposed a simplified CNN architecture with channel selection based on the DenseNet-201 model. The channel selection approach identified the top 10 channels from the original 32, namely Frontocentral 2 (FC2), Frontocentral 6 (FC6), Temporal 7 (T7), Central 3 (C3), Central zero (Cz), Parietal 3 (P3), Parietal zero (Pz), Occipital 1 (O1), Occipital zero (Oz), and Occipital 2 (O2). The architecture featured a simplified 4-layer CNN. The evaluation encompassed intra-subject and inter-subject classification of valence and arousal in low/high categories using a Gaussian Naive Bayes classifier. In intra-subject testing with 32 channels, the accuracy reached 97.4% for valence and 97% for arousal, whereas using the 10 selected channels resulted in an inter-subject accuracy of 92.4% for valence and 93.4% for arousal. For inter-subject testing with 32 channels, the accuracy reached 98.3% for valence and 96.7% for arousal, whereas using the 10 selected channels resulted in inter-subject accuracy of 92.1% for valence and 92.2% for arousal. Though it offers valuable insights, one of the primary constraints of STFT lies in choosing the length and type of window and its fixed time and frequency resolutions, which may not adequately capture the diverse frequency components present in EEG signals, especially during emotional states characterized by dynamic and rapidly changing neural activity. Reference [105] CWT emerges as a compelling alternative to address this limitation and enhance the effectiveness of EEG emotion detection. To this end, a comprehensive comparison between STFT and CWT was performed by [42]. He utilized CWT to convert EEG signals into EEG scalogram images since it achieved higher scores in all performance evaluation criteria than STFT. Deep features were extracted using the GoogleNet model. They were classified into emotion categories and frequency resolutions, which may not adequately capture the diverse frequency components present in EEG signals, especially during emotional states characterized by dynamic and rapidly changing neural activity. The classification methods considered were GoogLeNet, k-Nearest Neighbors (k-NN), SVM, and Extreme Learning Machine (ELM), evaluated on the GAMEEMO datasets. The results show that SVM outperformed other classifiers, yielding the highest Sensitivity, F1 score, and accuracy of 98.78%, while the k-NN classifier had the highest specificity (99.61%) and precision (99.60%). Using the DEAP dataset as a comparison, the proposed method achieved 91.2% and 93.7% accuracy scores for the High/Low-Valence and High/Low-Arousal categories, respectively. The ML classifiers performed better in classifying the deep features obtained from the GoogLeNet model than the GoogLeNet classifier itself, which was why GoogLeNet was primarily used for feature extraction, and the extracted features were then classified using different ML classifiers. On the other hand, [103] analyzed 3 CWT wavelets - Morse wavelet, Bump, and Amor (Analytic Morlet), revealing better results with the Morse wavelet. The first experiment generated scalograms using the 10 frontal electrodes—fP1, fP2, F3, F4, F7, F8, FC5, FC6, FC1, and FC2 solely due to the emotional relevance of the frontal brain. In contrast, the second experiment involved the utilization of all EEG electrodes in generating scalograms. The proposed 2-layer CNN demonstrated slightly higher accuracy by employing these 10 frontal channels compared to using all channels, achieving accuracies of 61.50% for valence and 58.50 for arousal on the DEAP dataset and 56.22% on the

TABLE 5. A summary of signal-to-image transformations and CNNs in emotion detection applications.

Ref.	Dataset	STI Transform	Classifier	Output Classes	Results (%)
[96]	DEAP	FFT	CVCNN (Computer Vision CNN)	2	Accuracies: - Low/High Valance = 88.76, - Low/High Arousal = 85.57,
[97]	DEAP	PSD + AEP	CNN + LSTM	2, 2, 2, 2	Accuracies for Low vs High binary classification: - Low/High Valence = 90.62, - Low/High Arousal = 86.13 - Low/High Dominance = 88.48, - Low/High Liking = 86.23
[98]	SEED, SDEA, MPED, DREAMER	AEP	Graph Embedded CNN	SEED = 3, SDEA = 3, MPED = 7, DREAMER = 2	Accuracies: - SEED dataset (3 classes): - Subject-dependent Positive/Neutral/Negative emotion = 92.93, - Subject-independent Positive/Neutral/Negative emotion = 82.46 - SDEA dataset (3 classes): - Subject-dependent Angry/Neutral/Funny = 79.69, - Subject-independent Angry/Neutral/Funny = 55.01 - MPED dataset (7 classes): - Funny/Joy/Anger/Fear/Disgust/Sadness/Neutrality = 40.98 - DREAMER dataset (2 classes): - Low/High Valence = 95.73, - Low/High Arousal = 92.79
[99]	SEED	Dynamic Differential Entropy (DDE)	2-layer CNN	2	SEED Dataset (2 classes) - Positive/Negative emotion = 97.56.
[100]	SEED	STFT	2-layer CNN	3	SEED dataset (3 classes): - Positive/Neutral/Negative emotion = 90.41 (Average Accuracy),
[101]	SEED, DEAP	EFDM through STFT	Residual Block CNN	3, 3	SEED dataset (3 classes): - Positive/Neutral/Negative emotion = 90.59 (Average accuracy), DEAP dataset (3 classes): - Positive/Neutral/Negative emotion = 82.84 (Average Accuracy)
[102]	DEAP	STFT	Simplified 4-layer CNN	2	Accuracies: - Intra-subject (32 channels): Low/High valence = 97.4, Low/High arousal = 97 - Intra-subject (10 channels): Low/High valence = 92.4, Low/High arousal = 93.4 - Inter-subject (32 channels): Low/High valence = 98.3, Low/High arousal = 96.7 - Inter-subject (10 channels): Low/High valence = 92.1, Low/High arousal = 92.2
[42]	GAMEEMO, DEAP	CWT, STFT	GoogLeNet (as Feature Extractor) + SVM	2, 2	Accuracies: GAMEEMO dataset (2 classes): - Positive/Negative emotion with STFT= 97.95 - Positive/Negative emotion with CWT= 98.78 DEAP dataset (2 classes): - Low/High Valance CWT = 91.2 - Low/High Arousal CWT = 93.7 (STFT was not applied to the DEAP dataset)
[103]	DEAP, SEED	CWT (Morse wavelet)	2-layer CNN	2, 2	Accuracies: DEAP dataset (2 classes): - For 10 Frontal electrodes: - Low/High Valence = 61.50, - Low/High Arousal = 58.50 - All 32 electrodes: - Low/High Valence = 59.50, - Low/High Arousal = 58.00 SEED dataset (2 classes): - For 10 Frontal electrodes: Low/High Valence = 56.22, - All 62 electrodes: Low/High Valence = 53.68

SEED dataset in a 3-class classification scenario involving Positive, Neutral, and Negative emotions.

D. OTHER EEG APPLICATIONS

In addition to these pivotal applications, a cluster of lesser-explored yet promising domains, such as schizophrenia detection, sleep stage classification, dementia analysis, etc., has harnessed the power of signal-to-image transformation and CNNs to advance their respective frontiers.

1) SCHIZOPHRENIA DETECTION

Schizophrenia (SZ) is a chronic mental disorder that affects how a person thinks, feels, and behaves. According to WHO [106] Schizophrenia affects approximately 24 million people or 1 in 300 people (0.32%) worldwide. There is a pressing demand for accurate and timely SZ identification. EEG waves can reveal changes in brain activity and provide information on brain changes during SZ, which can be analyzed to detect SZ. Reference [108] utilized STFT

to create visual spectrograms from raw multi-channel EEG data. These spectrogram pictures are fed into a VGG-16 CNN architecture, which extracts features and classifies them into healthy or schizophrenic categories. On Dataset A (children) and Dataset B (adults), the approach achieved 95% and 97% accuracies, respectively. Grad-CAM visualization shows that CNN mostly relies on changes in mid-range frequency patterns in spectrograms to distinguish schizophrenia patients. Reference [107] used STFT and CWT and smoothed pseudo-Wigner-Ville distribution (SPWVD) to obtain spectrograms, scalograms, and SPWVD-based TFRs, respectively.

To maintain the uniformity, the same parameters were applied to each method. AlexNet, VGG16, and ResNet50 networks are used for feature extraction and classification. The accuracy of CNN (93.36%) is the highest compared to other deep networks, such as AlexNet (93.33%), VGG16 (93.09%), and ResNet50 (93.34%), even after fine-tuning [108] employed a dataset of 14-channel EEG recordings from 14 schizophrenia patients and 14 healthy controls, with 12 minutes of data sampled at 250 Hz for each participant. The raw EEG signals are converted into images using CWT. These images are fed into pre-trained CNNs, such as AlexNet, ResNet18, InceptionV3, and VGG19 for feature learning. Only the convolutional and pooling layer features are collected from these CNNs, bypassing the fully-connected layers to prevent overfitting on the short dataset. Using an SVM classifier, the output feature vectors are categorized as schizophrenia or normal. ResNet18 had the highest average accuracy of 88% across all EEG channels. The proposed transfer learning strategy employing ResNet18 and SVM obtained 98.6±2.29% accuracy by merging relevant frontal, central, parietal, and occipital areas.

A novel method for mapping EEG data related to schizophrenia was introduced by [44] by using the IBIB PAN dataset from the Department of Methods of Brain Imaging and Functional Research of Nervous System. Initially, 1D EEG sequences were transformed into 3D images by generating EEG spatial feature matrices and conducting channel segmentation based on cerebral lobes. Subsequently, temporal feature patch merging was performed using convolutional layers. The study proposed a lightweight Vision Transformer model (LeViT), integrating four convolutional layers and attention mechanisms to learn spatial-temporal features within the mapped 3D images. This approach achieved an average accuracy of 98.99% in subject-independent criteria and 85.04% in subject-dependent criteria.

2) DROWSINESS DETECTION

Combining diverse feature extraction improves accuracy and robustness compared to individual techniques. Hybrid systems leverage the strengths of different feature extraction methods, making them versatile and accurate. Reference [109] used three feature extraction mechanisms (building blocks) to extract features from EEG signals and STFT spectrograms of EEG: Building Block 1 extracts energy/zero-crossing distributions and spectral entropy/instantaneous frequency features. Building Block 2 uses pre-trained deep CNNs (AlexNet and VGG16) to extract deep features from spectrogram images. Building Block 3 uses tunable Q-factor wavelet transform (TQWT) to decompose EEG signals into sub-bands and extract statistical features from the instantaneous frequencies. The extracted features from each block are fed into separate LSTM classifiers whose outputs are combined using majority voting. The proposed hybrid method was evaluated on the MIT-BIH Polysomnographic EEG dataset for binary classification of awake vs drowsy states, achieving 94.31% accuracy.

3) ATTENTION-DEFICIT/HYPERACTIVITY DISORDER (ADHD)

With a prevalence of 5.9% in youth and 2.5% in adults, untreated ADHD can result in a range of negative consequences [110] emphasizing the need for early detection. Reference [111] proposed a simple and novel method of STI conversion. In the study, the 19-channel EEG signals were segmented and passed through three 4th-order Butterworth bandpass filters to extract theta (4-8Hz), alpha (8-12Hz), and beta+gamma (12-40Hz) rhythms, resulting in 3 separate 2D matrices (of size 19 channels x 512 timepoints each) that contain the amplitude values for each of the 3 frequency bands respectively. These 3 matrices were then converted into the R, G, and B channels of the EEG image. The authors employed a custom CNN with Conv layers for the binary classification of ADHD/Normal and demonstrated an average accuracy on subjects of 97.81%, along with high precision, recall, and F1-scores.

Another approach by [112] involves recording EEG with the Starstim system from 7 positions covering the primary hubs of the fronto-parietal executive control network (Fp1, Fp2, F3, Fz, F4, P3, and P4) controls. Independent component analysis (ICA) was utilized to identify and remove noise components. The Wavelet transform was applied to signals using EEGlab's newtime function to create ERSP (Event-Related Spectral Perturbation). The performance of the deep learning architectures was assessed using accuracy and area under the curve (AUC). The four-layered (combining filtering and pooling) CNN trained with ERSPs achieved the highest accuracy of 88% and AUC of 96%, outperforming the RNN with stacked LSTM cells and SNN, indicating the effectiveness of CNN in discriminating between ADHD and healthy control groups.

4) CONSCIOUSNESS

Within the realm of consciousness studies, research conducted by [69] aimed at predicting levels of Depth of Anesthesia (DOA), specifically categorized as anesthetic light (AL), anesthetic OK (AO), anesthetic deep (AD), and signal polluted (SP), utilizing EEG signals. The study employed an enhanced STFT technique featuring a time-varying window function to generate EEG images. Three distinct CNN architectures were evaluated—CifarNet, AlexNet, and

VGGNet—. The findings revealed the superiority of the Modified STFT over the fixed-window STFT, resulting in enhanced classification accuracy from 84.7% to 92.3%. Moreover, the study demonstrated enhanced performance with deeper CNNs, with CifarNet achieving an accuracy of 87.5%, AlexNet achieving 92.35% accuracy, and VGGNet attaining the highest accuracy at 93.3%.

5) SLEEP STAGE CLASSIFICATION

Reference [113] proposed an orthogonal CNN (OCNN) with orthogonal weight regularization for single-channel EEG sleep state classification. The OCNN utilizes raw EEG data and transforms them into TF representations using the Hilbert-Huang transform. This technique employs EMD to break down the EEG into intrinsic functions and uses the Hilbert transform to extract instantaneous frequencies and amplitudes for each function. The proposed OCNN model achieved an accuracy of 88.4% on the UCD dataset and 87.6% on the MIT-BIH dataset when classifying 5 distinct sleep stages: wakefulness, Stage 1 (S1), Stage 2 (S2), slow-wave sleep (SWS), and rapid eye movement (REM) sleep. Incorporating orthogonal regularization in the OCNN proves pivotal, enabling reliable classification of sleep stages from EEG data.

6) DAYTIME SLEEPINESS ESTIMATION FOR APNEA PREDICTION

Predicting daytime sleepiness is crucial for enhancing public safety, as it allows for proactive measures to combat drowsy driving accidents and workplace errors caused by fatigue. For this purpose, [56] developed a CNN classifier to estimate daytime sleepiness in patients suspected of having obstructive sleep apnea (OSA). The classifier utilizes 2-channel EEG, 1-channel EOG, and 1-channel chin EMG signals from overnight polysomnography (PSG) to estimate the results of the multiple sleep latency test (MSLT). The approach involves applying Welch's PSD method within a frequency range spanning 0.3 to 30.3Hz on the 4 channels. Subsequently, the PSD estimates were converted to a dB scale and organized into spectrogram images where one column corresponds to an individual epoch. The study employs a custom CNN architecture, comprising 8 convolutional and 4 max-pooling layers, for both 4-class (severe, moderate, mild, normal) and binary (sleepy and non-sleepy) classifications based on mean sleep latency (MSL). In the 4-category classification, the overall accuracy of the CNN was 60.6%, and the model performed best for moderate sleepiness (66.9% accuracy) and worst for normal (52.0%), while the binary classification accuracy was 77.2%.

7) RECOGNITION OF GRAMMATICAL CLASS OF IMAGINED WORDS

Reference [114] recorded EEG using a Neuroscan 64-channel Quik cap of the extended 10-20 system, including electrodes for measuring eye movement (VEOG and HEOG). STFT was used to calculate spectrograms from EEG signals. A shorter Hanning window was used to improve temporal resolution. Brain signals were selected from three different brain areas. Electrodes in group 1 cover Broca's and Wernicke's areas. Electrode group 2 covers a large part of the frontal lobe. Electrode group 3 covers the Occipital lobe and the Parietal lobe. The spectrograms were subjected to baseline normalization and fed to Multichannel CNN with three channels. Each channel comprised three blocks containing 2D Conv layers for efficient feature extraction. The features extracted from all three blocks were combined using a concatenate layer to classify Verb vs Noun EEG signals, resulting in a recognition rate of 84.6%.

8) DETECTION OF FOCAL/NON-FOCAL EEG SIGNALS

Non-focal EEG signals signify normal brain activity or abnormalities that are not localized to a specific location. In contrast, focal EEG signals show abnormal brain activity focused on a specific brain region. Reference [115] mentions the use of the Bern-Barcelona-Dataset (BBD) for this analysis. Existing 1D EEG signals were converted into 2D RGB-scaled images with varied texture patterns using the RP (Recurrence Plot) technique. Several models like AlexNet, VGG16, VGG19, ResNet18, ResNet50, and ResNet101 were used for feature extraction, and the best features were obtained through VGG16. The VGG16 feature extractor was combined with various binary classifiers, such as SoftMax layer, decision tree (DT), random forest (RF), KNN, naïve Bayes (NB), and SVM to form ensemble models, which were then used to train and validate on the classification of Normal and Focal EEG signals. Performance comparison revealed that VGG16 alone achieved an accuracy of 96.36% while combining it with different classifiers yielded varying results: VGG16+DT (decision tree) achieved 95.78%, VGG16+KNN reached 96.98%, VGG16+Naive Bayes achieved 95.51%, VGG16+SVM attained 95.94%, and VGG16+RF (random forest) yielded the highest accuracy of 96.99%, surpassing all other combinations.

9) DEMENTIA STAGE CLASSIFICATION

Classifying dementia stages aids in optimizing treatment strategies, predicting disease progression, and fostering a better understanding of dementia's complexities. Reference [116] proposed a CNN framework using 2D spectral representations of EEG PSD for classifying Dementia stages, namely Alzheimer's disease (AD), mild cognitive impairment (MCI), and healthy control (HC). A modified periodogram method with rectangular windowing was employed aimed at reducing spectral leakage and achieving better resolution. The periodogram was computed for each epoch of the 19-channel EEG, providing information about the power present in different frequency bands. The estimated PSD values are then used to create PSD images. Each PSD image corresponds to an EEG epoch. The matrix is essentially a heatmap with higher PSD values represented by brighter pixels—a grayscale image. The study utilized a 1-layer CNN

for binary and 3-way classification, obtaining accuracies of 92.95% in AD vs HC, 84.62% in AD vs MCI, 91.88% in MCI vs HC, and 83.33% in 3-way (AD vs MCI vs HC) classification.

10) AUTOMATIC DETECTION OF AUTISM SPECTRUM DISORDER

Reference [118] proposed a method in which the initial phase involves refining raw EEG data through re-referencing, filtering, and normalization. Next, the preprocessed signals are divided into 3.5-second segments, which undergo STFT, and the generated spectrogram plots are saved as images for classification using both ML and DL methods. In the ML method, Spectrogram images were processed using the tCENTRIST algorithm, extracting textural feature vectors whose dimensionality was reduced by employing PCA. Six ML classifiers, including Naive Bayes, Linear Discriminant Analysis, Random Forest, KNN, Logistic Regression, and SVM, were tested, with SVM attaining the highest accuracy of 95.25%. In the DL-based process, three different CNN models are used for classification, out of which Model 3 performed best using a batch size of 64 with an accuracy of 99.15% and an F1 score of 1.0.

11) ALZHEIMER'S DETECTION

Reference [119] introduced a novel approach for Alzheimer's disease (AD) classification using EEG signals, presenting a Dual-Input Convolution Encoder Network (DICE-net). Utilizing recordings from 36 AD patients, 23 Frontotemporal dementia (FTD) patients, and 29 healthy individuals, the signals underwent preprocessing, including denoising and extraction of Band power, Coherence features, and time-frequency domain characteristics using methodologies like DWT. These features were inputted into the DICE-net architecture, comprising 2 Conv layers, 2 Transformer Encoder layers, and Feed-Forward layers. Results demonstrated DICE-net's efficacy, achieving an accuracy of 83.28% in distinguishing AD from healthy controls, surpassing baseline models, and exhibiting good generalization performance. The findings suggest that this convolution transformer network can effectively capture the intricate features of EEG signals, facilitating improved AD diagnosis and potential expansion to other dementia types like FTD. This approach holds promise for enhancing early detection accuracy and advancing interventions for AD.

12) CLINICAL DEPRESSION ANALYSIS

Utilizing integrated audio spectrograms alongside multiple frequencies of EEG signals, [120] reported a significant enhancement in diagnostic performance. This end-to-end framework was developed using the Multimodal Open Dataset for Mental Disorder Analysis (MODMA) dataset. STFT was applied to extract spectrograms from both EEG signals and audio data. Various pre-trained CNN architectures, including ResNet, DenseNet, and EfficientNet, were explored for feature extraction, with DenseNet demonstrating slightly superior overall performance. Notably, a proposed ViT method emerged as the top performer, achieving an accuracy of 97.31%, precision of 97.71%, and recall of 97.34% metrics, showcasing promising prospects for clinical depression diagnosis through advanced deep learning techniques.

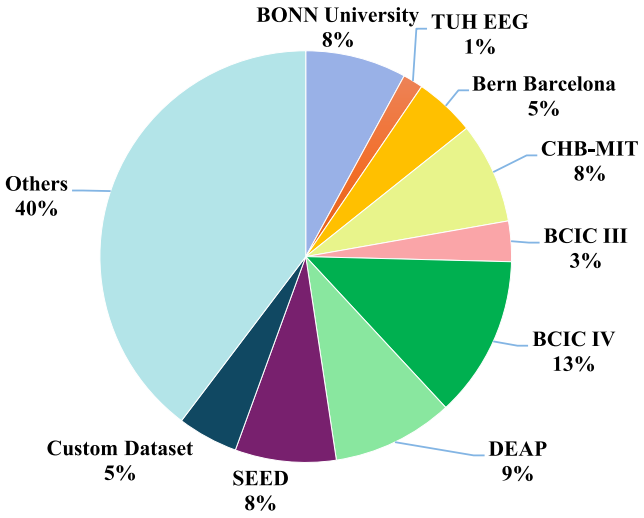


FIGURE 3. Distribution of datasets used in EEG classification tasks.

13) DELIRIUM PREDICTION

Using a limited dataset of critically ill older adults, [121] employed a Vision Transformer (ViT) model to detect delirium in EEG data. The cohort included 13 individuals aged 50 and above, requiring mechanical ventilation in ICUs, with seven experiencing delirium according to CAM-ICU assessment. EEG data underwent rigorous preprocessing, including artifact removal and noise reduction through Individual Component Analysis (ICA). Sampled every 4 milliseconds from eight sensors, the continuous data was segmented into various lengths (0.1s to 5s) and resized to 224 $\times$ 224 images for ViT input. Achieving training accuracies surpassing 99.9%, ViT demonstrated optimal testing accuracy of 97.58% using 5-second data slices, outperforming traditional methods like random forest and support vector machines in delirium classification. Notably, ViT's robust performance persisted even without ICA cleaning, suggesting the potential for accurate delirium prediction in EEG data without extensive feature engineering.

III. CNNs AND SIGNAL-TO-IMAGE CONVERSION IN EMG APPLICATION DOMAINS

Surface electromyography (sEMG) signals are recorded using electrodes placed on the skin surface above the muscles. sEMG signals can be recorded using two main approaches, each offering distinct advantages. High-density sEMG (HD-sEMG) employs a grid or array of electrodes with a substantial number, typically ranging from 100 to 200 electrodes, closely spaced at around

TABLE 6. A summary of signal-to-image transformations and CNNs in other EEG applications.

Ref.	Dataset	STI transform	Classifier	Output Classes	Results (%)
[117]	Dataset A (Mental Health Research Center), DatasetB (Institute of Psychiatry and Neurology in Warsaw, Poland)	STFT	VGG-16	2, 2	Accuracies: (2 classes) - Normal/Schizophrenic: Dataset A = 95, Dataset B = 97
[107]	Kaggle dataset (EEG data from the basic sensory tasks in Schizophrenia)	STFT, CWT, Smoothed pseudo-Wigner-Ville distribution (SPWVD)	AlexNet, VGG16, ResNet50, CNN	2	Accuracies: (2 classes) - Normal/Schizophrenic: SPWVD with CNN = 93.36 - Using STFT as STI: - STFT with AlexNet = 78.40 - STFT with VGG16 = 78.56 - STFT with ResNet50 = 79.18 - STFT with CNN = 79.17 - Using CWT as STI: - CWT with AlexNet = 90.28 - CWT with VGG16 = 89.20 - CWT with ResNet50 = 90.19 - CWT with CNN = 90.64 - Using SPWVD as STI: - SPWVD with AlexNet = 93.33 - SPWVD with VGG16 = 93.09 - SPWVD with ResNet50 = 93.34 - SPWVD with CNN = 93.36
[108]	Publicly available EEG Data approved by the Ethics Committee of the Institute of Psychiatry and Neurology in Warsaw, Poland	CWT (Morse Wavelet)	Inception-v3 + SVM, VGG-19 + SVM, AlexNet + SVM, ResNet18 + SVM	2	Highest Accuracy: (2 classes) - Normal/Schizophrenic: - ResNet18 + SVM = 98.6±2.29
[109]	MIT/BIH Polysomnographic EEG	STFT + TQWT (Tunable-Q Wavelet Transform)	AlexNet+VGG16+LSTM (Ensemble Model with outputs of each model combined using Majority Voting)	2	(2 classes) - Awake/Sleep-Stage-1: Accuracy = 94.31
[110]	First EEG Data Analysis Competition with Clinical Applications by the National Brain Mapping Laboratory of Iran	Signal amplitudes of different frequency bands were mapped to RGB color channels to generate images.	3-layer CNN	2	(2 classes) - ADHD/Normal: Accuracy = 97.81
[112]	The dataset included EEG data from 40 participants recorded with the starstim system: 20 healthy adults (10 males, 10 females) and 20 ADHD adult patients (10 males, 10 females). Each session comprised a total of 140 trials.	Wavelet Transform	2-layer CNN	2	(2-classes) - Healthy/ADHD: Accuracy of 2-layer CNN = 88
[69]	Raw EEG signals collected from the National Taiwan University Hospital (NTUH), Expert assessment of conscious level dataset (EACL)	Modified STFT	CifarNet, AlexNet, VGGNet.	NTUH = 4 EACL = 4	Average Accuracies across 10 subjects: - CifarNet = 87.53, - AlexNet = 92.64, - VGGNet = 93.5
[113]	University College Dublin (UCD), MIT-BIH	Hilbert-Huang Transform	Baseline Orthogonal CNN (OCNN), OCNN + SE block (squeeze-and-excitation)	5, 5	Accuracies: - UCD (5 classes) - Wake/S1/S2/ SWS/ REM: - Baseline OCNN = 84.7 - OCNN + SE block= 88.4, - MIT-BIH (5 classes) - Wake/S1/S2/SWS/REM: - OCNN + SE block = 87.6
[56]	Recordings conducted by Loewenstein Hospital—Rehabilitation Center, Israel.	Welch's Power Spectral Density	8-layer CNN	4, 2	Accuracies: - (4 classes) - Severe/Moderate/Mild/Normal = 60.6, - (2 classes) - Sleepy/Non-Sleepy = 77.2
[114]	EEG activity was recorded using a Neuroscan 64-channel Quik cap based on the extended 10–20 system. 19 participants imagined speech in response to 10 randomly presented words (5 nouns and 5 verbs). Each trial had a total duration of 4 seconds, including a 1-second blank screen before stimulus onset and a 1-second blank screen after the stimulus.	STFT	Multichannel CNN (each channel consists of 3 convolutional layers, and a concatenate layer combines their outputs)	2	(2 classes) - Noun/Verb: Accuracy = 84.6

TABLE 6. *(Continued.)* **A summary of signal-to-image transformations and CNNs in other EEG applications.**

[115]	Bern Barcelona Dataset	Recurrence Plot	- Only VGG16, - VGG16 + Decision Tree (DT), - VGG16 + Random Forest (RF), - VGG16 + KNN, - VGG16 + Naive Bayes, - VGG16 + SVM	2	Accuracies on Bern Barcelona Dataset (2 classes - Normal/Focal EEG): - Only VGG16 = 96.36 - VGG16 + DT = 95.78 - VGG16 + KNN = 96.98 - VGG16 + Naive Bayes = 95.51 - VGG16 + SVM = 95.94 - VGG16 + RF = 96.99 (Highest)
[116]	The dataset included 189 subjects from IRCCS Centro Neurolesi Bonino-Pulejo in Messina (Italy): 63 with Alzheimer's Disease, 63 with Mild Cognitive Impairment, and 63 with healthy controls. EEG recordings followed the standard 10–20 International System.	PSD	1-layer CNN	2, 3	Accuracies: - for AD vs HC = 92.95 - for AD vs MCI = 84.62 - for MCI vs HC = 91.88 - for 3-way classification (AD vs. MCI vs. HC) = 83.33
[118]	King Abdulaziz University (KAU) Hospital Dataset	STFT	4-layer CNN	2	(2 classes) - Healthy/Autism-spectrum-disorder: Accuracy = 99.15
[44]	IBIB PAN - Department of Methods of Brain Imaging and Functional Research of Nervous System (schizophrenia)	EEG data is transformed into a spatial feature matrix, followed by segmenting channels based on cerebral lobes and then merging temporal feature patches.	Lightweight Vision Transformer model (LeViT) (4 convolutional layers)	2	Average Accuracies: (2 classes) - healthy control/ schizophrenia patient: - 98.99 in subject-independent criteria - 85.04 in subject-dependent criteria.
[119]	Recordings of 36 AD, 23 Frontotemporal dementia (FTD), and 29 age-matched healthy individuals (CN) of AHEPA General University Hospital of Thessaloniki.	Discrete Wavelet transform	Dual-Input Convolution Encoder Network (DICE-net).	3	Accuracy - (3 classes) - AD/CN/FTD = 83.28

10 mm intervals. This configuration grants HD-sEMG exceptional spatial resolution, allowing for the precise observation of muscle activity at the motor unit level. Moreover, it captures signals at high sampling rates, typically exceeding 1kHz, which enables detailed tracking of rapid muscle activations. Conversely, sparse multi-channel sEMG utilizes a smaller number of electrodes, typically fewer than 16, with greater spacing, typically between 2-3 cm, resulting in coarser spatial resolution. The sampling rates are comparatively lower, around 100-200Hz, making recording simpler. Figure 4 visually distinguishes between HD-sEMG and sparse multi-channel sEMG, highlighting HD-sEMG's provision of instantaneous snapshots with higher-fidelity signals and fine spatial and temporal resolutions, in contrast to sparse multi-channel sEMG, which relies on temporal windows for contextual information while offering practical simplicity in data acquisition.

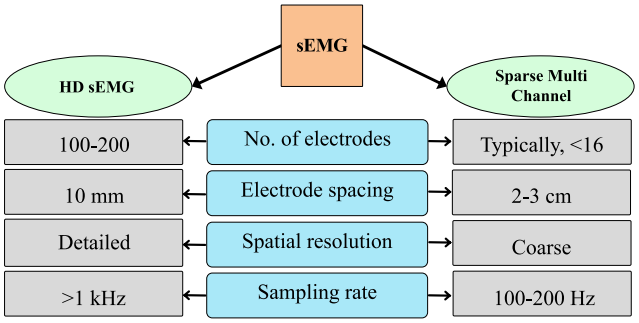


FIGURE 4. **Comparison of HD-sEMG and sparse multi-channel sEMG acquisition methods.**

A. CNNS IN EMG GESTURE RECOGNITION

Gesture Recognition is a vital field focused on developing systems that can interpret and classify EMG signals generated by muscle activity. It holds immense potential in enhancing the quality of life for individuals with diverse motor abilities. Studies have harnessed various methods to process and extract valuable information from EMG data. There are various fundamental tools in these investigations. One such is using the Fourier transform as an essential step in generating spectrograms for feature extraction. To begin with, [122] applied STFT on the 1D raw EMG signals acquired from NinaPro databases DB3 and DB4 to convert them into 2D functions with the resulting spectrogram computed by taking the square magnitude of the STFT. The CNNs automatically learn discriminative features from the spectrogram images, eliminating the need for feature engineering and selection. Two CNN architectures were compared: CNN-1 comprised one conv layer, while CNN-2 incorporated two conv layers. Results indicated that CNN-2 with two conv layers extracted superior features, yielding higher accuracies for both NinaPro DB3 and DB4 datasets. Specifically, CNN-2 achieved accuracies of 57.4% and 88.04% for NinaPro DB3 and DB4 datasets, respectively outperforming the first CNN architecture, which achieved

accuracies of 60.34% and 86.18% for the same datasets in classifying the 17 hand and wrist movements. Similarly, [123] collected 8-channel EMG data, encompassing three grasping gestures (fist, pinch, card) and three sign language gestures (ok, victory, good), utilizing an armband. A simple CNN was employed for classification, utilizing various input data representations comprising raw data, STFT, wavelet transform (WT), and scale average wavelet transform (SAWT) images. Among the STI conversion methods compared, CNN using SAWT images achieved the highest accuracy of 94.6% for classifying the selected hand gestures, outperforming WT (90.3%) and STFT (92.1%). Additionally, the SAWT approach imposed a lesser computational load than conventional multi-channel STFT or WT techniques. In another approach, [124] harnessed STFT to create frequency feature maps from multi-channel EMG signals, classifying 9 distinct classes using a CNN collected from three healthy men and an amputee. These classes included rest and 8 distinct postures (power grasp, precision grasp, lateral grasp, wrist flexion, wrist extension, wrist pronation, and wrist supination), each with 3 force levels. The results demonstrated that, in the case of healthy subjects, the proposed CNN method exhibited a 5-10% higher classification accuracy of 88% compared to the traditional ANN method, both before and after reseating the sensors. Furthermore, CNN outperformed the ANN by approximately 20% in classifying force levels, while the performance in posture classification remained similar. This suggests that CNN holds a distinct advantage in classifying finer differences within EMG signals. On a similar note, [125] investigated the utility of Convolutional-ViT (CViT) in the realm of EMG-based gesture classification, leveraging the integration of convolutive Blind Source Separation (BSS) preprocessing techniques. They obtained HD-sEMG signals from two publicly available datasets. They employed a 2D grid representation of raw input data denoted as $X \in RC \times S \times D$, where C represents the number of channels (set to 64), S denotes the window size (102 for DB1 and 50 for DB2), and D signifies the depth (maintained at 1). Using a Hann window, STFT was applied to capture changes in non-stationary EMG signals, yielding a 2D STFT image. The proposed CViT model was structured with 3 conv layers for spatial feature extraction and dimensionality reduction, followed by a transformer block housing 1 multi-head attention layer to capture long-range dependencies. The BSS-CViT model exhibited 96.61% and 91.98% accuracies on the two datasets, showcasing a significant 6.63% improvement over the shallow-CNN approach. This disparity underscores the limitations of CNNs in capturing the intricate long-term dependencies in raw HD-sEMG data. At the same time, the implementation of CViT proved highly effective in addressing this challenge.

Furthermore, [126] acquired EMG signals from the upper arm and upper body, specifically without considering wrist muscles from 7 healthy subjects (6 electrodes), 1 trans-radial amputee, and 1 wrist amputee using 8 surface electrodes. The signals were normalized and segmented. FFT was applied on segments to extract spectrogram features using the Hamming window. PCA was used to reduce the dimensionality to 25 channels and convert it to a 5×5 matrix while maintaining useful information fed to a 4-layered CNN. The proposed model was compared with traditional TDAR-SVM (Time Domain Auto-Regressive) in which features such as MAV, ZC, SSC, WL, and AR are extracted from the EMG signals and concatenated to be classified by SVM. While PCA-CNN outperformed TDAR-SVM's 61.6% accuracy for healthy subjects using 6 electrodes with 69.4%, TDAR-SVM was slightly better (62%) than PCA-CNN (58.6%) for amputees. This performance can be improved by providing more data. Similarly, [127] collected sEMG signals using an 8-channel Myo armband for 10 hand gestures. Wavelet transform is applied to remove noise, and a maximum value is derived from each of the 8 channels for every gesture, which serves as an additional auxiliary 9th channel. The 9-channel sEMG signals were converted into spectrogram images by mapping time from original time-domain signals and frequency from frequency-domain signals generated by the Fourier transform. The dataset was evaluated on 4 models: Model-1 is a single-label classifier, Model-2 is a multi-label classifier, Model-3 is a combination of both with 22 deepened network layers, and Model-4 in which pretreatments are carried out on the sEMG signals of each channel, such as removing noise. Out of all 4 models, Model-1 with 5 conv layers, when iterated 3100 times, yielded an accuracy of 94.06% when compared to 93% by SVM, while the multi-labelled CNN having 9 SoftMax outputs (one per channel) outputs a label obtained by the Majority Vote Algorithm.

On the other hand, many methods involve wavelet-based transformations. Reference [128], As such, CWT was used over STFT to convert raw signals into spectrograms because CWT can better adapt to the non-stationary characteristics of sEMG signals. These images were fed into EMGNet with 4 Conv layers to classify the movements. The 4-layer CNN model was evaluated on 2 different datasets: On the Myo Dataset consisting of 7 classes, it achieved an accuracy of 98.81%, while on the NinaPro DB5 Dataset consisting of 3 subsets, each with 12, 17, and 23 classes, it yielded accuracies of 69.62%, 67.42%, 61.63% respectively. On a similar note, [129] proposed a Region-based CNN (R-CNN) with WPT as an STI conversion method. WPT is a variation of DWT that provides a more flexible and comprehensive signal decomposition by allowing multiple sub-bands to be analyzed at each level of decomposition [130]. The proposed R-CNN is a combination of VGG-16-based CNN that jointly extracts informative features from EMG signals through convolutional and pooling layers and a Region Proposal Network (RPN), which subsequently identifies potential gesture regions using the RPN. This enhances the model's capacity for accurate recognition by combining feature learning with region-focused processing, resulting in an accuracy of 96.5%

in the classification of 4 hand gestures, namely close fingers, wave-in, fist, and gun.

While CWT is intricate and computationally demanding, it is worth noting that transforming EMG data into structured 2D representations through reshaping provides a simpler and computationally more efficient approach. Notably, [131] obtained 2D EMG-picture representations by reshaping the 1D EMG time-series data from the Myo armband sensor into a 50×8 matrix with 50 timesteps and 8 channels. A convolutional recurrent neural network (CRNN) model consisting of conv layers and a Gated Recurrent Unit (GRU) was used to extract spatial data from EMG pictures to classify 10 distinct dynamic hand gesture classes, namely rest, close, open, fist, right, left, thumb up, thumb down, supination, and pronation. In training subjects, the accuracy is 96.57%, whereas in fresh subjects, it is 95.10%.

Similarly, [132] introduced a Transfer Learning strategy for sEMG gesture recognition, collecting 128-channel EMG data from forearm and upper arm muscles for a source gesture dataset that included 30 distinct hand gestures involving diverse finger, wrist, and elbow positions. A sliding window was implemented to divide the EMG data points into multiple segments, and these segments were reshaped into a 16×8 image, where 16 rows represent the channels, and 8 columns represent the time steps. So, each source image captures the spatial pattern of muscle activation across 128 channels, which were used to train the source CNN and designed as a general gesture feature extractor. Transfer Learning was employed by transferring the initial layers of a source CNN to both the target CNN and CNN-LSTM networks. These networks were then fine-tuned using limited data from three distinct targets datasets: TG_BS, TG_New, and DB-a. CNN achieved an accuracy of 92.4%. The target CNN and target CNN-LSTM achieved accuracies of $92.13 \pm 4.5\%$, $93.32 \pm 3.9\%$, $91.18 \pm 6.25\%$, and $93.73 \pm 7.03\%$, $97.34 \pm 3.79\%$, $94.57 \pm 6.77\%$ for the TG_BS, TG_New, and DB-a datasets respectively. The study concluded that Transfer Learning significantly decreased average training times and remained effective even when dealing with new users, new gestures, and variations in data collection systems or locations. Additionally, due to the inclusion of temporal context, it was noted that the Target CNN-LSTM slightly outperformed the Target CNN. Reference [133] proposed a CNN model employing a 1-D convolution kernel to extract intricate abstract characteristics and enhance recognition accuracy. The evaluation utilized the NinaPro DB1 dataset, encompassing 52 distinct gestures. Three variations of sEMG images, including raw-sEMG images, sEMG-feature images, and multi-sEMG-features images, were generated using the mapping method of the multi-channel sEMG amplitude to the image pixel value referring to Du's coloring scheme. These images were then fed as inputs for the deep CNN model comprising 2 conv layers for classification. The amalgamation of the deep CNN with the multi-sEMG-features image achieved the highest average accuracy in gesture recognition, achieving a notable 82.54%. Reference [134] converted instantaneous signal samples to image pixels directly for HD-sEMG (as for csl-hdemg & CapgMyo datasets), while for sparse multi-channel sEMG (as for NinaPro), a time window is employed to sample the sEMG signals, and the signals recorded by C channels within an L-frame time window are converted to an sEMG image of size $L \times C$. The key idea is to use a ''divide and conquer'' strategy. The approach consists of two stages: the multi-stream decomposition stage and the fusion stage. In the multi-stream decomposition stage, the sEMG image is divided into equal-sized patches, and each patch is used as input for a single-stream CNN. In the fusion stage, the learned features from all streams are combined and fed into a fusion network for improved gesture recognition accuracy. The proposed CNN is compared to a standard single-stream CNN, which is similar to each stream in the multi-stream framework. For the ssCNN, the full raw sEMG image is provided as input rather than divided patches. msCNN outperforms ssCNN on all three datasets with 85% accuracy on the NinaPro dataset, 95.4% on the csl-hdemg dataset, and 99.8% on CapgMyo with a 150 ms time window using majority voting.

Amid the prevailing trends in TF analysis, Variational Mode Decomposition (VMD) offers a distinctive avenue dedicated to extracting essential features from complex multivariate time-series data. Reference [135] applied VMD, proposed by [137] as an enhancement over EMD, to extract spatial-temporal features from multi-channel sEMG signals. A two-stage classifier was used - the first stage classified gestures into 3 superclasses (Finger, Wrist, Functional movements) using 1D EMG and SVM. The second stage took Multivariate-VMD (MVMD) decomposed sEMG as input to a Separable CNN to predict the final 52 gesture classes within each superclass. The accuracies achieved by the approach on the NinaPro DB1 dataset across three electrode configurations were - DB1-E1 (12 classes): 93.95%, DB1-E2 (17 classes): 92.9%, and DB1-E3 (23 classes): 88.67%. MVMD overcomes EMD's robustness issues stemming from its dependence on extreme points and stopping conditions due to its mathematically less grounded approach. While the conventional emphasis lies on time or TF characteristics, few prefer PSD-based techniques as they offer a comprehensive view of the signal's spectral content. As such, [55] applied Welch's method, which uses FFT to generate PSD feature maps from 8-channel raw EMG signals. Subsequently, a 5-layer CNN classifier was used to recognize hand gestures from the EMG-PSD features, achieving a 99% accuracy on 6-gesture classification (Right, Left, Up, Down, Stop/fist, None/no gesture).

Meanwhile, [136] introduces a new deep learning architecture named Temporal Multi-Channel Vision Transformer (TMC-ViT) for hand gesture classification using surface electromyography (sEMG) signals. The TMC-ViT interprets the input as a 2D grid of data, which involves mapping raw sEMG signals into dimensions $N \times T$, where N represents the number of electrodes and T signifies the time steps enabling the capture of temporal patterns effectively. Adapting the

TABLE 7. A summary of signal-to-image transformations and CNNs in gesture recognition applications.

Ref.	Dataset	STI transformation	Classifier	Output Classes	Results (%)
[122]	Ninapro Database DB3 and DB4	STFT	1-layer CNN (CNN-1), 2-layer CNN (CNN-2)	17, 17	Average Accuracies across 11 and 10 subjects respectively: - NinaPro DB3 (amputee subjects) - 17 hand/ wrist movement classes - CNN-1 = 57.40 - CNN-2 = 60.34 - NinaPro DB4 (healthy subjects) - 17 hand/ wrist movement classes - CNN-1 = 86.18 - CNN-2 = 88.04
[123]	Performing 200 repetitions for each gesture and 1200 repetitions in total, the operations involve capturing 50-200 samples at 2-second intervals and using them as one.	STFT, WT, SAWT (scale average wavelet transform)	2-layer CNN	6	Average Accuracies across 3 subjects (6 classes): Fist/Pinch/Ok/Card/Victory/Good - STFT = 92.1 - WT = 90.3 - SAWT = 93.9
[124]	Involving 3 healthy men and 1 amputee, the 4 subjects underwent a protocol with a 4-second rest time and a 5-second keep time.	STFT	2-layer CNN	9	(9-classes) - rest/open-palm/power-grasp/precision-grasp/lateral-grasp/wrist-flexion/wrist-extension/wrist-pronation/wrist-supination: Accuracy of CNN = 88
[126]	7 healthy subjects, 1 trans-radial amputee, and one wrist amputee completed tasks with 10 repetitions each, utilizing the Southampton Hand Assessment Procedures (SHAP) for reaching-to-grasping tasks.	FFT and Hamming Window	Hybrid Network: PCA-CNN	6	Accuracies for PCA-CNN: (6 classes) - spherical/tripod/power/lateral/tip/extension. - for Healthy Subjects = 69.4 - for transradial amputees = 58.6 - for wrist amputees = 54.5
[127]	Within this dataset, 50 subjects utilized Myo devices, performing gestures with an incorporated relaxation time of 3 seconds.	Spectrograms created by mapping time and frequency	5-layer CNN	10	Accuracy = 94.06
[128]	Myo Dataset, NinaPro DB5 Dataset	CWT	4-layer CNN	Myo dataset = 7, Ninapro DB 5 = 12,17,23.	Accuracies: - Myo Dataset (7 gestures) = 98.81 - NinaPro DB5: - 12 gestures = 69.62 - 17 gestures = 67.42 - 23 gestures = 61.63
[129]	Jester dataset	Wavelet packet decomposition (WPT)	R-CNN (Region-based CNN)	4	- (4 classes) wave-in/gun/fist/close-fingers: Accuracy = 96.5
[131]	Starting from the "Rest" position, a specific action was performed, returning to "Rest" for 1 second, collecting one set of 400 EMG data points from 50 samples across 8 channels within 1 second using Myo band.	Reshape Function	Hybrid Network: CRNN (CNN + RNN (GRU))	10	Accuracies: (10-classes) - Close/Open/Rest/Right/Left/Thumb-up/Thumb-down/Supination/Pronation/Fist. - On Training Subjects = 96.57, - On Fresh Subjects = 95.10
[132]	The source gesture set of 30 gestures, representing various states of elbow, wrist, and finger joints, was utilized. Two target sets, TG_BS (30 gestures mirroring the source set) and TG_New (10 custom gestures distinct from the source set), were collected from 28 new participants without muscular or joint disorders.	Sliding window segmentation and Reshaping	Source: CNN, Target: CNN, CNN + LSTM	Source dataset = 30, TG_BS = 30, TG_New = 10, DB-a = 8.	- Source: 92.4 - Target CNN+LSTM: TG_BS = 93.73±7.03, - TG_New = 97.34±3.79%, DB-a = 94.57±6.77
[133]	NinaPro DB 1	Du's Coloring Scheme	2-layer CNN	52	Accuracy = 82.54
[134]	NinaPro DB-1, Csl-hdemg, CapgMyo	HD-sEMG: direct, Sparse: time window	2-layer CNN (multistream CNN)	52, 27, 8	Accuracies: - NinaPro = 85, - Csl-hd EMG = 95.4, - CapgMyo = 99.8
[135]	NinaPro DB1	MVMD	SVM + Separable CNN (S-CNN)	52	Accuracies: NinaPro_DB1 - Exercise set-1 (12 classes) = 93.95, - Exercise set-2 (17 classes) = 92.9, - Exercise set-3 (23 classes) = 88.67

TABLE 7. *(Continued.)* **A summary of signal-to-image transformations and CNNs in gesture recognition applications.**

[55]	20 healthy participants repetitively performed 6 gestures, each 12 times consecutively over 2 days, with the gesture and rest periods lasting 1 second.	Welch's method	5-layer CNN	6	(6-classes) - Right/Left/Up/Down/Stop/None: Accuracy = 99
[136]	Ninapro DB5, New Dexterity dataset	Mapping raw sEMG signals into dimensions N × T, where N represents the number of electrodes and T signifies the time steps	TMC-ViT, ViT, CNN, DCNN, LSTM Linear Discriminant Analysis, Random Forest, SVM	18,5	Accuracies: -Ninapro DB5 (Raw signals) -TMC-ViT = 89.6 -ViT = 82.6 -CNN = 83.89 -DCNN = 80.64 -LSTM = 80.53 Average Accuracy: -New Dexterity dataset (5 classes) Pinch grasp/ tripod grasp/power grasp/co-contraction of all muscles/ rest state. -TMC-ViT = 99.93 (Raw signals) -Linear Discriminant Analysis = 93.49 -Random Forest = 97.66 -SVM = 94.33
[125]	- Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings," IEEE Trans. Neural Syst. Rehabilitation. "A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition," Nature Electron.	STFT	BSS-CViT (Blind source separation convolution Vision Transformer) (3 conv layers)	34,21	Accuracies: -Dataset 1 = 96.61 -Dataset 2 = 91.98

Vision Transformer to handle multi-channel temporal signals, the model incorporates convolutional neural network blocks to reduce dimensionality and extract embeddings. These embeddings undergo further processing in the Vision Transformer blocks with multi-head attention mechanisms for efficient feature extraction. Evaluation of the Ninapro DB5 dataset shows an impressive 89.60% accuracy with raw sEMG data, outperforming other DL models like ViT, CNN, DCNN, and LSTM. The TMC-ViT achieved an outstanding 99.93% accuracy on the New Dexterity dataset, surpassing classical machine learning methods like Linear Discriminant Analysis, Random Forest, and SVM. Despite having efficient prediction times of less than 0.56 ms per sample while competing with CNN and RF models, the TMC-ViT excels in accuracy, showcasing a performance edge over other deep learning models considered in the study.

B. CNNS IN OTHER EMG APPLICATIONS

Various other avenues were investigated and pursued in the realm of EMG. Reference [138]'s work centers on fusing EEG and EMG signals to determine the actions and intentions of the user by classifying task weight levels: 0 lbs, 3 lbs, and 5 lbs. using CNN. The acquired signals were processed and segmented to remove non-moving portions, which were converted into images by 2 methods: the signal images method involves stacking the time series signals into an array to form an image, while spectrogram images were generated by applying STFT using the Hann window. These images were then normalized and fused to input 2D CNN. Three fusion methods were tested, out of which the grouped spectrogram method had the highest mean accuracy of 80.51% compared to stacked (80.03%) and mixed (79.72%). The authors state that frequency information captured in the spectrograms is more relevant for classifying task weight during motion than solely time domain information. The results, therefore, show that 2D CNN models leveraging the TF information via spectrograms outperformed the 1D CNN models using the raw time series signal images.

Meanwhile, [55] collected a dataset comprising EMG signals from the forearm using the Myo armband device. The signals underwent processing using Welch's method, which involved applying the FFT to each signal segment and then averaging the resulting periodograms to calculate the PSD for generating feature maps. These feature maps were used as input for a CNN designed to classify the signals into six categories: Right, Left, Up, Down, Stop, and None. The system achieved an average accuracy of 99% in recognizing gestures. To estimate the percentage of muscle activity, the system compared the envelope of the EMG signals to a reference signal from the user. The estimation was computed by averaging the envelope values obtained from selected sensors, depending on the recognized gesture. The system demonstrated precise muscle activity estimation, with consistent results achieved using the Butterworth filter and the root mean square (RMS) method. The percentage estimation indicated 50% for stable force and 22% for incremental force.

On a different note, [139] introduced a unique approach leveraging Transfer Learning for EMG-based personal identification using CWT and CNN. The EMG signals were collected from the open-hand gesture of 21 volunteers using the Myo armband and transformed into 2D

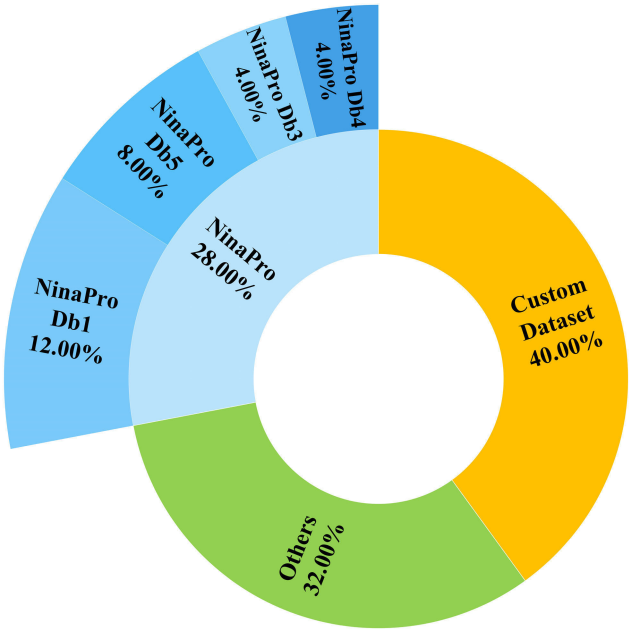


FIGURE 5. Distribution of datasets used in EMG classification tasks.

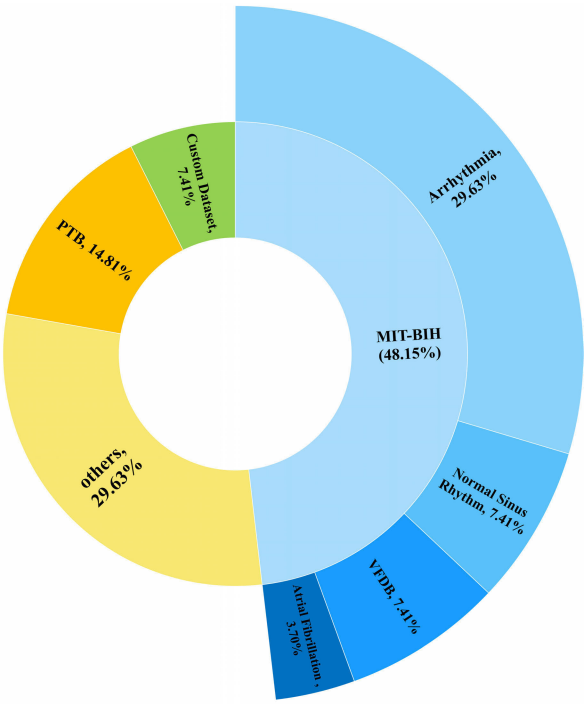


FIGURE 6. Distribution of datasets used in ECG classification tasks.

scalograms through CWT. The identification network utilized a 4-layer CNN trained with CWT scalograms, achieving a recognition accuracy of 99.203%, while the verification network employed a Siamese network, reaching an accuracy of 99.285%. When combined with a transfer learning algorithm, the personal identification network can efficiently retrain the model when new data is added.

In summary, the breakdown of EMG research datasets in Figure 5 offers noteworthy findings. The majority constitutes a custom dataset (40.00%), emphasizing tailored data collection. The NinaPro database significantly contributes 28.00%, with Db3 and Db4 each representing 4.00%, Db5 consisting of 8.00% and NinaPro Db1 comprising 12.00%. Other datasets comprise 32.00%, highlighting the significance of diverse and comprehensive data in understanding EMG signals.

IV. CNNS AND SIGNAL-TO-IMAGE CONVERSION IN ECG APPLICATION DOMAINS

The applications of CNNs in the realm of electrocardiography (ECG) represents a transformative approach that has brought new dimensions to the analysis and interpretation of cardiac data. With the advent of CNNs, ECG signals can be effectively converted into visual representations, harnessing the power of STI transformation techniques [19] indicates that, in comparative analyses, 2D-CNNs initialized with AlexNet weights demonstrate superior performance in contrast to 1D signal methods, even in the absence of extensive large-scale datasets.

Figure 6 illustrates the dataset distribution for ECG classification. The MIT-BIH dataset dominates, accounting for 48.15% of the total, while the smaller portions include PTB (14.81%), Custom Dataset (7.41%), and Others (29.63%).

A. ARRHYTHMIA DETECTION

Classifying arrhythmias is of significant importance in the realm of healthcare as it facilitates the early identification and accurate diagnosis of abnormal heart rhythms, thereby reducing the risk of critical cardiac incidents. Consequently, several DL approaches have been explored for arrhythmia classification. Starting with [65], a novel 2-stage classification approach was introduced, utilizing recurrence plots (RP) and CNN. The ECG data was obtained from several public databases to cover different arrhythmia types: MIT-BIH Arrhythmia Database, Creighton University Ventricular Tachyarrhythmia Database, MIT-BIH Atrial Fibrillation Database, and MIT-BIH Malignant Ventricular Ectopy Database. The data was segmented into 2-second ECG signal segments, converted into 2D RPs to visualize recurrence patterns, essentially avoiding extensive feature extraction and noise filtering steps, and then used for classifying between noise ventricular fibrillation (VF), and 'other' rhythms in the initial stage, while the later stage focused on classifying 'other' rhythms into atrial fibrillation (AF), normal, premature atrial contraction (PAC), and premature ventricular contraction (PVC) using beat-based RPs centered on detected R-peaks. The study compared AlexNet, VGG16, and VGG19 on the RP images and observed best performance and computational efficiency with AlexNet yielding an average testing accuracy of 91.83% for stage-1 and 98.44% for stage-2 classification of arrhythmias compared to accuracies of 89.92% for stage-1 and 91.52% for stage-2 by VGG-16 and 84.21% for stage-1 and 89.70% for stage-2 by VGG-19.

TABLE 8. A summary of signal-to-image transformations and CNNs in other EMG applications.

Ref.	Dataset	STI transform	Classifier	Output Classes	Results (%)
[138]	The dataset comprises fused EMG and EEG data with three classes obtained from 32 healthy subjects.	STFT	3-layer CNN	3	Grouped Spectrogram Method = 80.51±8.07 (Accuracy)
[55]	20 healthy participants repetitively performed 6 gestures, each 12 times consecutively over 2 days, with the gesture and rest periods lasting 1 second.	Welch's Method	5-layer CNN	6	Accuracies: Gesture recognition = 99 The percentage estimation: - Stable force = 50, - Incremental force = 22
[139]	21 diverse, healthy subjects contribute to reliable and extensive experimental data by repeating the same task 30 times, with each task including a 2-s pre-test period, a 1.5-s hand gesture period, and a 2-s post-test period.	CWT	- Personal identification = 4-layer CNN, - Personal Verification = Siamese 1-layer CNN	21	Accuracies: - Personal identification (4-layer CNN) = 99.203, - Personal Verification (Siamese 1-layer CNN) = 99.285

If the arrhythmias exhibit distinctive recurring patterns or extended dependencies, employing RPs may prove to be more effective in capturing and representing these recurring patterns. On the other hand, if arrhythmias present themselves as fleeting events characterized by fluctuating frequencies, CWTs can be instrumental in capturing and highlighting these dynamic features. Consequently, [140] generated 2D TF scalograms from individual heartbeats using CWT with the Mexican hat wavelet and fed them into a 3-layer CNN for feature extraction, incorporating RR interval-based features; namely, the Previous RR interval, Post RR interval, RR interval ratio, and Local average RR interval, and concatenated to the CNN features before passing through the dense layers. The proposed method was evaluated on the MIT-BIH arrhythmia database, where it successfully classified heartbeats into four main categories: Normal beat, Supraventricular ectopic beat (SVEB), Ventricular ectopic beat (VEB), and Fusion beat, achieving an overall accuracy of 98.74%.

Shifting the focus to [141], ECG time domain signals from the MIT-BIH database consisting of 5 types of arrhythmias namely: normal beat (NOR), left bundle branch block beat (LBB), right bundle branch block beat (RBB), premature ventricular contraction beat (PVC), and atrial premature contraction beat (APC) were transformed into spectrograms using STFT and fed to 2D CNN for classification, without the need for manual feature extraction. After conducting experiments with various learning rates and batch sizes, they settled on the optimal learning rate of 0.001 and a batch size of 2500. The proposed model was evaluated and compared with a 1D CNN model. The average accuracy of the proposed 2D-CNN model, at 99.00%, surpassed the 1D-CNN model, which achieved 90.03%. In a comprehensive study by [142], various representations of ECG spectrograms were compared alongside different CNN architectures for classifying 6 ECG heartbeat arrhythmias, including normal beat, left bundle branch block (LBBB), right bundle branch block (RBBB), premature ventricular contraction (PVC), namely Log-scale STFT, Mel-scale STFT, Bispectrum, and 3rd-order Cumulant. Additionally, four CNN architectures were compared, including AOCT-Net, MobileNet, SqueezeNet, and ShuffleNet. Among the CNN models, MobileNet achieved the highest accuracy, and the best approach was the combination of Mel-scale spectrogram representation and MobileNet, resulting in an accuracy of 94.6%. Moreover, the results indicate that SqueezeNet was fastest at 2.7s, followed by ShuffleNet, AOCT-Net, and MobileNet.

Reference [143] presented an automated method for detecting shockable ventricular cardiac arrhythmias (SVCA) on the analysis of ECG signals acquired from two public databases, the Creighton University database (CUDB) and MIT-BIH malignant ventricular arrhythmia database (VFDB) by leveraging the Fixed Frequency Range based on the Empirical Wavelet Transform (FFREWT) filter bank for multiscale analysis and segmentation, founded upon the detection of spectral boundary points. A 4-layered CNN yielded a classification accuracy of 81.25% for ventricular fibrillation vs ventricular tachycardia, 99.036% for shockable vs non-shockable classification tasks, and 99.8% for ventricular fibrillation vs normal.

Several researches were explored in segmenting the ECG signal and extracting meaningful features or representations from these segments for further analysis, albeit using different techniques and representations. Indeed few approaches share a common emphasis on analyzing segments of ECG signals. Certainly, [144]'s approach exemplifies this diverse landscape. They adopted a two-step preprocessing approach involving noise reduction and QRS segmentation. Subsequently, the feature extraction process unfolded in two stages: firstly, by evaluating QRS segments using spectral entropy derived by normalizing the power spectrum of the signal and then calculating its Shannon entropy, which results in TF maps, and then, by reducing the dimensionality of TFM using $(2D)^2$PCA. The performance of the proposed CNN with 3 Conv layers was assessed on the MIT-BIH dataset with 5 types of heartbeats for different time resolutions and frequency partitions, with the highest attained

TABLE 9. A summary of signal-to-image transformations and CNNs in Arrhythmia detection.

Ref	Dataset	STI transform	Classifier	Output Classes	Results (%)
[65]	MIT-BIH Arrhythmia Database, Creighton University Ventricular Tachyarrhythmia Database, MIT-BIH Atrial Fibrillation Database, and MIT-BIH Malignant Ventricular Ectopy Database (VFDB)	Recurrence Plot	- AlexNet, - VGG-16, - VGG-19	Stage-1 = 2 (Noise/VF), Stage-2 = 4 (Normal/ AF/PAC/PVC)	Accuracies: - VGG-19: - stage-1 = 84.21, - stage-2 = 89.70 - VGG-16: - stage-1 =89.92, - stage-2 = 91.52 - AlexNet: - stage-1 = 91.83, - stage-2 = 98.44
[140]	MIT-BIH Arrhythmia database	CWT (Mexican hat wavelet)	3-layer CNN	4	Accuracy - (4 classes) Normal/ SVEB/VEB/Fusion = 98.74
[141]	MIT-BIH Arrhythmia database	STFT	3-layer CNN	5	- (5 classes) Normal/LBB/ RBB/PVC/APC: Average Accuracy = 99.00
[142]	MIT-BIH Arrhythmia database.	- Log-scale STFT, - Mel-scale STFT, - Bispectrum, - 3rd order Cumulant	- AOCT-Net, - MobileNet, - SqueezeNet, - ShuffleNet	6	Accuracies: - (6 classes) - APB/LBB/ Normal/PVC/ RBB/aAP. - Bispectrum representation: - SqueezeNet = 93.7 - AOCT-Net = 93.5 - MobileNet = 93.4 - ShuffleNet = 92.8 - 3rd order cumulant representation: - MobileNet = 94.4 - AOCT-Net = 94.3 - SqueezeNet = 93.2 - ShuffleNet = 93 - log-scale spectrogram representation: - MobileNet = 92.7 - ShuffleNet = 91.6 - SqueezeNet = 91.2 - AOCT-Net = 90.5 - Mel-scale spectrogram representation: - MobileNet = 94.6 - SqueezeNet = 93.6 - AOCT-Net = 93.4 - ShuffleNet = 92.7
[143]	Creighton University database (CUDB), MIT-BIH malignant ventricular arrhythmia database (VFDB)	FFREWT	4-layer CNN	2, 2, 2	Accuracies: - (2 classes) Shockable/ non-shockable = 99.036, - (2 classes) ventricular-fibrillation/normal = 99.8 - (2 classes) ventricular-fibrillation/ ventricular-tachycardia = 81.25
[144]	MIT-BIH Arrhythmia dataset	Spectral Entropy	3-layer CNN	5	Accuracy - (5 classes) Normal/LBBB/ RBBB/PVC/APB = 98.33
[145]	European ST-T database (ESCDB)	Mapped amplitude values to pixel intensities across a 2D-pixel grid	4-layer CNN	3	Accuracy - (3 classes) Normal/ ventricular/ectopic ST-segment changes = 99.23

accuracy of 98.33% for a time resolution of 0.05 seconds and 18 frequency partitions. Meanwhile, [145]'s approach stands out by uniquely focusing on mapping R-R intervals to a 2D-pixel grid. ECG signal segment between two consecutive R-peaks, called the R-R interval, was extracted, and its amplitude values were mapped to pixel intensities across a 2D-pixel grid of 28×28 fixed image size. These ECG images were classified using the proposed 4-layer CNN architecture into 3 classes: normal, ST-change, and ventricular ectopic beat. They obtained the best accuracy of 99.23% with the intra-patient scheme on the European ST-T database recorded with single lead L3.

B. HEARTBEAT CLASSIFICATION

Several innovative approaches have emerged to enhance the accuracy and efficiency of heartbeat classification, which is paramount in healthcare for early detection and accurate diagnosis of cardiac conditions. Building upon this image-based approach, [146]'s research employed Faster R-CNN with 2D ECG images as input. After preprocessing ECG signals from the MIT-BIH database and patient recordings using EMD for denoising and detecting R-peaks via DWT, the transformation of 1D ECG signals into 2D ECG images was accomplished using the sliding window algorithm. The 2D ECG images were fed into a Faster R-CNN model, which consisted of a feature extractor and a region proposal network for classifying ECG beats into 5 classes, namely Normal(N), Supraventricular(S), Ventricular(V), Fusion(F), Unknown(Q). The model achieved an average test accuracy of 99.21% across classes, surpassing the performance of one versus rest SVM, which achieved an accuracy of 96.62%.

Another dimension of ECG signal transformation was explored by [147], who adopted CWT to convert 1D ECG signals into 2D scalogram images, which were then used to train an AlexNet deep CNN model with transfer learning, retraining only the fully connected layers. Experiments on the PhysioNet dataset with arrhythmia from MIT-BIH, congestive heart failure from BIDMC, and normal sinus rhythm (NSR) ECG data from subjects with no significant arrhythmias demonstrated that the proposed approach obtained 98.7% accuracy in classifying Arrhythmia (ARR)/Congestive heart failure (CHF)/Normal sinus rhythm (NSR). In a contrasting approach, [148] proposed efficient multimodal fusion frameworks, Multimodal Image Fusion (MIF) and Multimodal Feature Fusion (MFF), employing Gramian Angular Field (GAF), Recurrence Plot (RP), and Markov Transition Field (MTF) to create three different images for ECG heartbeat classification. This approach considers the fusion of multiple modalities for improved classification accuracy. In MIF, the images were fused into a 3-channel image input for a CNN classifier. In MFF, features extracted from each imaging modality using AlexNet were fused via a Gated Fusion Network and classified by SVM. Experiments were carried out using the PhysioNet MIT-BIH Arrhythmia dataset, which contained 5 arrhythmia types (N/S/V/F/Q), and the PTB Diagnostics (2 classes: MI/Normal) dataset for myocardial infarction (MI) detection. MFF achieved 99.7% and 99.2% accuracy in arrhythmia and MI classification, respectively, but MIF had a faster inference speed.

Departing from conventional ECG signal analysis, [149] utilizes an arterial blood pressure (ABP) dataset from the PhysioNet database, notably the MGH dataset from the multi-parameter databases (MIMIC) category, to classify heartbeats as normal or abnormal, offering a fresh perspective on data sources. The raw signals were preprocessed by denoising. DWT was used to extract details and approximation coefficients, which provide information about the different scales and details of the ABP signal and obtain a continuous signal. To gain more information about it in a 2D space, they used CWT to generate scalograms that were fed into CNN. The Semi-AlexNet model was used for classification and attained 89.03% accuracy.

C. BIOMETRIC AUTHENTICATION

The R-wave is a crucial feature in ECG analysis as it corresponds to ventricular depolarization. The ECG signal can be segmented as in [150] or analyzed around the R-peak as in [151]. The former approach is centered on short ECG signal segments, precisely extracting 0.5-second windows centered on R-peaks. To enhance information content, they employed CWT, yielding valuable TF representation images. A small CNN classifier was designed to learn from these CWT images with simpler decision boundaries. Additionally, pre-trained deep CNN models, namely GoogLeNet, ResNet, MobileNet, and EfficientNet, were evaluated for biometric recognition. Experiments were conducted on PTB single-session and ECG-ID multisession datasets, comprising 100 and 90 subjects, respectively. Identification and verification performance were also analyzed on the multisession data. The proposed model achieved an accuracy of 99.90% on PTB, 98.20% for ECG-ID mixed-session datasets, and 94.18% for ECG-ID multisession datasets. Regarding 0.5-second intervals surrounding the R-peaks in ECG-ID multisession datasets, the ResNet model obtained an accuracy of 97.28%. Verification performance was statistically significant for all models.

Meanwhile, the latter detected R-waves using the Pan-Tompkins algorithm from the filtered ECG signals of MIT-BIH NSR following segmentation to identify the P, QRS, and T waves. ECG images are obtained by projecting signals onto a 2D space using a linear equation after estimating the partial baseline using regression analysis. Finally, user recognition is processed through deep learning with automatic feature extraction and learning using an ensemble model consisting of 2 CNNs (one with 3 Conv layers and another with 2) to extract spatial features and one RNN using LSTM for temporal information. The best output features from all 3 models are fused and retrained to get the final classifier. This ensemble network model achieved the highest recognition rate of 98.9% compared to each single network.

Reference [152] described a personal recognition system based on the ECG signal's 2D coupling image. The

TABLE 10. A summary of signal-to-image transformations and CNNs in heartbeat classification.

Ref.	Dataset	STI transform	Classifier	Output Classes	Results (%)
[146]	MIT-BIH Arrhythmia database, Patient recordings	Sliding window algorithm	Faster Regions -CNN	5	(5-classes) - Normal/Supraventricular/Ventricular/Fusion/ Unknown: Average accuracy = 99.21
[147]	PhysioNet: MIT-BIH arrhythmia database, BIDMC CHF database, NSR from subjects with no significant arrhythmias	CWT	AlexNet	3	Accuracy - (3 classes): Arrhythmia (ARR)/Congestive heart failure (CHF)/Normal sinus rhythm (NSR) = 98.7
[148]	MIT-BIH Arrhythmia dataset, PTB Diagnostics	Image Fusion of-Gramian Angular Field (GAF), RP, and Markov Transition Field (MTF).	AlexNet+SVM	5, 2	Accuracies: MIT-BIH (5-classes) - N/S/V/F/Q = 99.7, PTB Diagnostics (2-classes) - MI/Normal = 99.2
[149]	MGH dataset	CWT	Semi AlexNet	2	Accuracy - (2 classes): Normal/Abnormal = 89.03

2D coupling image is created using three preprocessed and partitioned 1D ECG signal periods. These 2D coupling images were used to train a 12-layer CNN designed for image classification. For comparison, pre-trained networks like resnet-v2-152, inception-resnet-v2, inception-v4, and inception-v3 were tested on the MIT-BIH dataset. Experiments on the MIT-BIH and PTB datasets revealed that the pre-trained networks like Inception V3, Inception V4, Inception-ResNet V2, and ResNet V2-152 achieved accuracies of 98.71%, 99.12%, 100%, and 100%, respectively, across 18 classes whereas the proposed CNN achieved an accuracy of 99.2% on MIT-BIH. On PTB data, the accuracy of the 12-layer CNN was 98.45%. Meanwhile, [154] transformed the 1D ECG signal into a 224 × 224 pixel grayscale image, with pixel intensity determined by the amplitude of each ECG sample, and subsequently converted it into a binary image through thresholding. Furthermore, various feature extraction algorithms were applied to the resulting 2D binary image to effectively detect patterns, including peaks, noise, and baseline drifts, while also segmenting the 2D ECG images into beats centered around the R-peak. A custom CNN model with 6 Conv layers was designed and trained to classify an input ECG signal as belonging to the genuine user or an imposter. The proposed approach was evaluated on MWM-HIT, PTB, CYBHi and MIT-BIH databases, achieving accuracies of 96.20%, 99.27%, 90.20% and 99.96 respectively.

V. DISCUSSION

This review exclusively covers papers from the past five years. It underscores the intersection of signal processing and deep learning as a powerful strategy for automated feature learning and classification of complex physiological signals. The transformation of 1D time series data into 2D images enables the application of advanced 2-dimensional convolutional and other DL architectures that can effectively learn spatial, spectral, and temporal characteristics and discriminative features from the visual representations.

Figure 9 reveals that STFT dominates with 26.59% of the total classifiers, CWT and PSD techniques are also prevalent, representing 23.39%, within which, Morlet wavelet is the most frequently utilized, accounting for 6.38% in each. The Hanning window is the most popular choice within the STFT at 8.51%.

For EEG analysis, TF methods like STFT, CWT, and WPT, as well as feature calculation and projection methods like CSP and AEP, offer means to generate 2D images effectively capturing spectral, temporal, and spatial aspects. Though STFT remains widely used due to its computational efficiency, CWT provides variable resolution and has shown advantages in handling non-stationary dynamics compared to fixed-resolution STFT. Meanwhile, when used as complementary techniques, approaches like CSP and AEP can provide an essential dimension to the analysis and serve as potent tools to enhance the insights obtained through TF methods. Studies indicate that CNNs using integrated STI conversion approaches seeking to effectively capture the complex spectral, temporal, and spatial aspects of EEG can yield higher accuracies and represent a promising direction for future research in this field.

In EMG and ECG analysis, employing time-frequency transforms like STFT and CWT have been found effective in converting 1D signals into 2D spectrogram and scalogram images as input for CNN models. For EMG gesture recognition, CNNs achieve strong performance, with CWT providing better adaptation to signal non-stationarity, although STFT enables faster processing. However, direct signal reshaping of multi-channel EMG into 2D arrays also offers a simpler alternative enabling spatial pattern learning by CNNs. For ECG analysis, CNNs achieve high performance for detecting a range of arrhythmia types from ECG spectrograms and scalograms. In addition, when dealing with ECG signals, which exhibit specific recurrent patterns and long-term dependencies, Recurrence Plots (RPs) have proven to be effective in capturing these features. On the contrary, CWT was able to record ECG components manifesting as transient events with varying frequencies.

FIGURE 7. Distribution of various ML and DL techniques employed for physiological signal analysis.

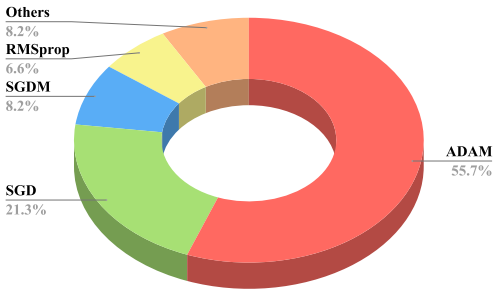


FIGURE 8. Distribution of optimizers used in CNNs for physiological signal classification.

Moreover, it has been reported that deeper networks and transfer learning models trained on natural images were able to mitigate overfitting on limited ECG data.

Figure 7 provides an insightful overview of the distribution of classifier choices in the context of physiological signal processing. Deep learning techniques are the dominant preference, standing alone, and constitute a significant 62.74% of the cases reviewed. Within DL, Custom CNNs emerge as the frontrunner, accounting for over half of DL approaches at 48.04%. Moreover, the application of pre-trained models, such as AlexNet, VGG, and ResNet, is notable and constitutes 14.7% of the classifier choices, emphasizing the significance of transfer learning. On the ML side, SVM stands out with 4.90%. Additionally, 19.6% of studies employed combinations of classifiers, reflecting the flexibility and innovation within the field. A key insight is that deeper CNN architectures like VGG, ResNet, and Inception pre-trained on natural images tend to extract more optimal features, even when data is limited. Training a DL model is often a

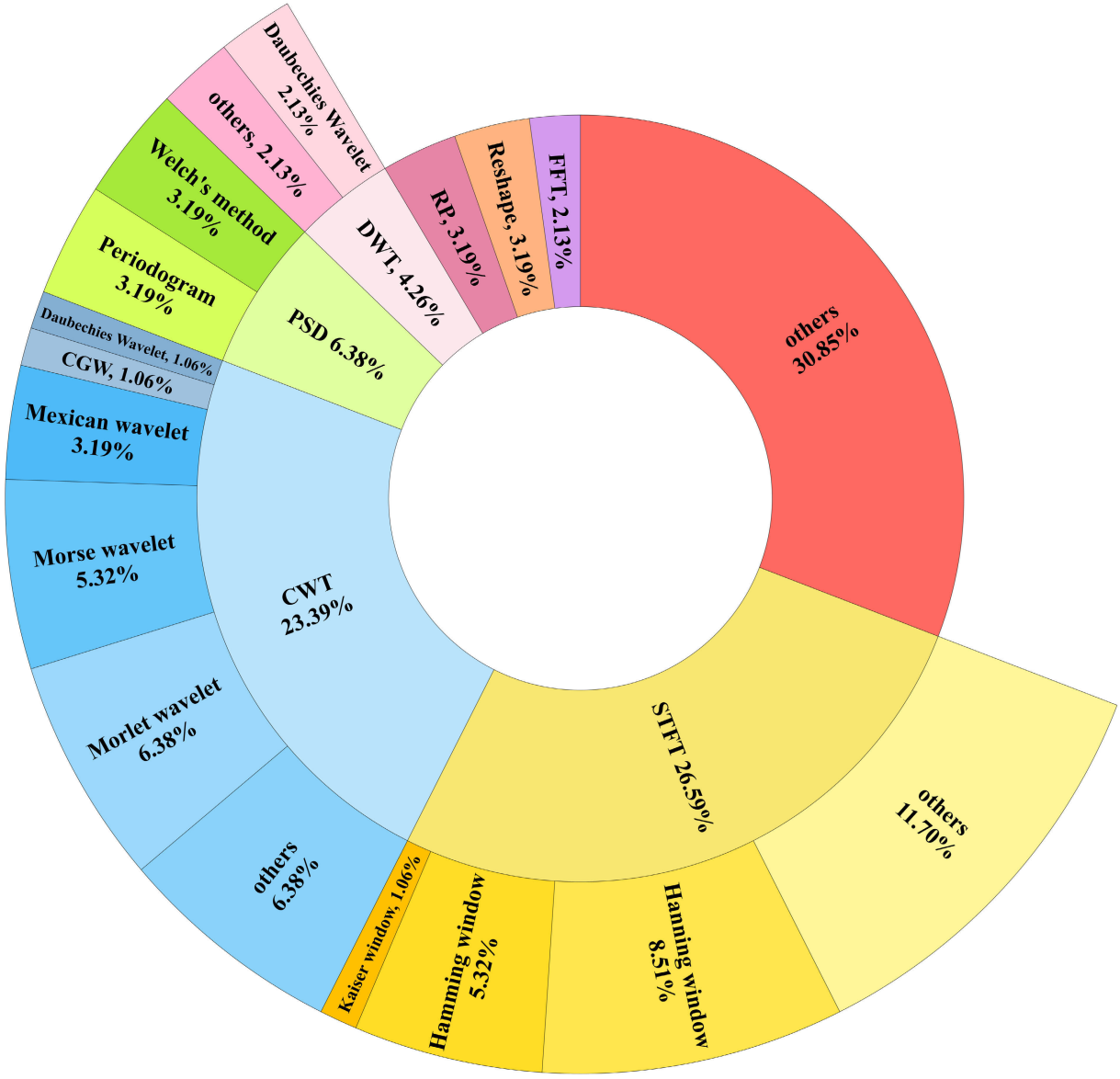


FIGURE 9. Distribution of various STI transformation techniques employed for physiological signal analysis.

multifaceted endeavor. The subtle interplay between parameter fine-tuning and architectural design choices influences the model's overall performance. Consequently, multiple combinations must be considered. In addition, the training phase involves intensive computations, resulting in lengthy training times. Transfer learning appears to be a potential fix to these problems. It accelerates the training process by utilizing pre-trained models and their learned features, reducing the requirement for intense parameter adjustments and training timeframes. Fine-tuning via transfer learning is thus a valuable strategy. However, shallower custom CNNs may suffice for some applications. Comparative analysis on larger benchmark datasets is needed to glean conclusive design guidelines.

Another significant challenge arises when selecting hyperparameters, as the choice of optimizer and learning rate can profoundly influence training dynamics and convergence. Figure 8 shows that the most commonly adopted optimizer was the Adam optimizer (55.7%), with the SGD optimizer (21.3%) also being used but not that extensively. The optimizer choice can profoundly influence training dynamics and convergence. Furthermore, it was learned that learning rates in the order of 1e-4 yield optimal results, with 1e-3 also being effective for physiological signal classification in combination with Adam optimizer. Additionally, the neural network's architecture poses challenges, including the decision on the number of layers and the size of each layer. Our analysis showed that neural networks with 4, 5, or 6 layers were frequently employed for biomedical signal analysis. A poor choice may lead to underfitting or overfitting, hampering the model's generalization capability.

TABLE 11. A summary of signal-to-image transformations and CNNs in biometric authentication.

Ref	Dataset	STI transform	Classifier	Output Classes	Results (%)
[150]	PTB single session, ECG-ID multisession datasets	CWT	9-layer CNN, Pre-trained networks: GoogLeNet, ResNet, MobileNet, EfficientNet.	100, 90	Highest Accuracies: PTB (100-classes): - MobileNet = 100, - ResNet = 100 ECG-ID single-session dataset (90-classes): - GoogleNet = 98.2 ECG-ID mixed-session dataset (90-classes): - ResNet = 98.95 ECG-ID multi-session dataset (90-classes): - ResNet = 97.28
[151]	MIT-BIH Normal sinus rhythm	Mapping signals to a 2D space using linear equations.	The Ensemble network combines outputs from a 3-layer CNN, a 2-layer CNN and an LSTM.	18	98.9 Accuracy
[152]	MIT-BIH normal sinus rhythm database, PTB diagnostic database	Overlapping of 3 periods of 1-D ECG signal	- 12-layer CNN, - resnet-v2-152, - inception-resnet-v2, - inception-v4, and - inception-v3	18, 52	Accuracies: MIT-BIH (18-classes): - Inception V3 = 98.71 - Inception V4 = 99.12 - Inception-ResNet V2 = 100 - ResNet V2-152 = 100 - 12-layer CNN = 99.2 PTB diagnostic (52-classes): - 12-layer CNN = 98.45
[153]	MWM-HIT, MIT-BIH, PTB, CYBHi	Amplitude-based pixel intensity and thresholding.	6-layer CNN	2	Accuracies: - MWM-HIT = 96.20, - PTB = 99.27, - CYBHi = 90.20, - MIT-BIH = 99.96

In recent years, there has been a notable surge in research efforts to leverage advanced architectures combining convolutional structures with attention mechanisms. This approach offers a promising avenue for extracting both locally and globally dependent information. By integrating convolutional structures and attention mechanisms, these architectures facilitate the learning of spatial-temporal features [44].

Pooling operations in traditional methods like max or average pooling generate fixed-length representations, treating all positions equally and disregarding inter-position relationships. However, attention mechanisms dynamically weigh the significance of different positions based on context [83]. This allows for a more nuanced approach to sequence analysis. While CNNs struggle with capturing long-term dependencies, the CNN-ViT framework effectively extracts global context information, thereby improving long-term dependency modelling [125]. Lightweight Convolution Transformer (LCT) integrates convolution tokenization and attention-based pooling to extract spatial and temporal correlated information from multi-channel signals concurrently, addressing ViT's lack of translation equivariance and localization. Additionally, ViT's class token disregards the relationship between different time-step data, causing information loss [83]. Hybrid CNN-ViT architectures leverage the inherent bias of convolutional layers for uncertainty modelling in EEG data, enhancing local feature processing through self-attention mechanisms. Moreover, Data Uncertainty Learning integrates convolutions in transformers, mitigating Multi-Head attention's limitations in handling local features and ensuring stability in classification accuracy even with increased noise. Thus, CNN-ViT models not only reduce the computational complexity of transformer modules, which limits their deployment on resource-constrained signals through convolution down sampling but also enhance global context awareness and discrimination ability compared to pure ViT architectures [82].

Moreover, while CNNs have traditionally been limited in their capacity to capture long-term dependencies in raw physiological data, hybrid CNN-ViT architectures have emerged as a viable solution. By effectively extracting global context information, CNN-ViT frameworks improve the model's ability to model long-term dependencies, thereby enhancing the processing of local features through the self-attention mechanism in the transformer module [82], [125].

The inclusion of CNN layers in Transformers addresses their limited generalization on insufficient data by providing essential inductive biases like translation equivariance and localization [154], [155], leading to a more lightweight Transformer. CNN-ViT architectures alleviate the need for extensive training data required by vanilla ViT, with convolution blocks extracting spatial features and improving feature map resolution. Moreover, CNNs reduce input dimensionality while preserving important data, ensuring compatibility with ViT's input requirements. This integration allows CNNs to learn embeddings that serve as inputs for ViT, facilitating efficient feature extraction and embedding learning in hybrid architectures [136].

TABLE 12. Summary of various popular physiological signal datasets.

Dataset	Description
BONN University EEG [131]	The EEG signals were captured using a 128-channel device. It comprises five datasets organized into three classes, each with 100 single-channel EEG segments lasting 23.6 seconds. The samples were taken at a rate of 173.61Hz.
Bern-Barcelona EEG [132]	Five epilepsy patients' EEGs were recorded with 10-20 electrode placement at a sampling rate of 1024 Hz and then down sampled to 512 Hz. The dataset includes epileptic and non-epileptic states.
CHB-MIT [133]	The scalp EEG data from 23 pediatric patients and one adult patient were collected over 916 hours of continuous scalp EEG sampling at 256 Hz and classified into seizure and non-seizure states.
TUH EEG Seizure Corpus [134]	The dataset contains 16,986 EFG sessions from 10,874 patients recorded with the International 10-20 system electrode placement. The majority of EEG data was sampled at 250 Hz. The dataset includes 6 classes.
BCI Competition III [135]	The dataset consists of 8 datasets that address various challenges in Brain-Computer Interface (BCI) research. These challenges include session-to-session transfer, non-stationarity, limited training sets, subject-to-subject transfer, continuous unlabeled data, asynchronous protocols, and idle states. Subjects, sessions, and days vary across these datasets, encompassing healthy and impaired individuals. Furthermore, the result classes differ among these datasets.
BCI Competition IV [136]	Participants were challenged to develop optimized algorithms for different paradigms and data types. The competition comprised four datasets, with subjects ranging from 2 to 9, classes ranging from 2 to 5, channels varying from 10 to 64, and sampling frequencies spanning from 100 Hz to 1000 Hz.
DEAP [68]	monitored 32 individuals' electroencephalograms (EEG) and peripheral physiological data while watching 40 one-minute-long music video clips. Participants assigned ratings to each video based on its arousal, valence, like/dislike, dominance, and familiarity levels.32 active AgCl electrodes were used to record the EEG at a 512 Hz sampling rate and were positioned using the worldwide 10-20 system.
SEED [137]	This study comprises eye movement and EEG data from 12 subjects and 3 more subjects. Data was collected while they were watching film clips. The film clips were chosen carefully to evoke various emotions, including good, negative, and neutral ones. EEG data and eye movements were recorded using the 62-channel ESI NeuroScan System and SMI eye-tracking glasses. EEG was captured using the worldwide 10-20 system at a sampling rate of 1000 Hz.
NinaPro dB1 [138]	It comprises EMG data from 27 people executing 53 hand movements, recorded with 10 OttoBock electrodes and delivered at a frequency of 2000 Hz after preprocessing. Finger, grip, wrist, and daily activity movements were recorded during 8 hours of synchronized EMG, kinematics, and label data. It aids in the creation and testing of EMG-based prosthetic hand control algorithms.
MIT-BIH [139]	The 47 patients in the BIH Arrhythmia Laboratory's 1975–1979 study is represented by 48 half-hour portions of two-channel ambulatory ECG recordings in the MIT-BIH Arrhythmia Database. The recordings were digitized across a 10-mV range at 360 samples per second at a resolution of 11 bits.
PTB Diagnostic ECG [140]	The database has 549 recordings from 290 individuals (ranging in age from 17 to 87). One to five recordings are used to represent each subject. The traditional 12 leads (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5, v6) as well as the three Frank lead ECGs (vx, vy, vz) are all simultaneously assessed in each record, making 15 signals. Each signal is digitalized with a sampling rate of 1000 samples per second and a resolution of 16 bits over a range of 16.384 mV. There are 9 classes in the database.

Tuning CNN-ViT models is more challenging than pure CNN architectures, requiring exploration of superior techniques for fusing CNN and Transformer modules [82]. Despite promising outcomes, clinical biomarkers of irregular EEG signal alterations remain understudied, which can be analyzed through fine-grained discriminative features of EEG signals, particularly those with high frequency and long-term dependencies. However, Transformers struggle with low-pass filtering -and reduced accuracy with increased depth, which can be addressed by incorporating CNN to introduce high-frequency components to strengthen the model's capability [83].

From our vantage point in the dynamic landscape of research, the fusion of CNNs with ViTs emerges as a promising avenue for the analysis of physiological signals. These hybrid architectures elevate conventional methods by seamlessly integrating CNNs' spatial feature extraction with ViTs' global context awareness and attention mechanisms. Moreover, CNNs aid in feature embedding and contribute to reducing the complexity of the network compared to pure ViT architectures, making them more computationally efficient. This integration also leverages the inherent inductive bias provided by CNNs, such as translation equivariance and localization, which aids in improving the model's interpretability and performance. However, challenges persist in optimizing these models and exploring finer features of physiological signals. Nonetheless, this integration marks a notable advancement in deciphering complex physiological data, holding potential for various clinical applications.

Furthermore, in advancing the field of physiological signal processing through deep learning, it is imperative to advocate for a consistent and reproducible evaluation framework

TABLE 13. A comparison of signal-to-image conversion techniques for physiological signals.

STI Conversion Method	Advantages	Disadvantages
FFT	- Computationally efficient - Captures frequency information - Useful for extracting frequency components from physiological signals	- Does not retain time information - Assumes signal stationarity
PSD	- Provides spectral power distribution - Noise-resistant - Valuable for analyzing frequency domain features in physiological signals	- Averaging can obscure transient features - Assumes weak stationarity
Welch's method	- Reduces noise in PSD estimate - Does not require the entire signal - Beneficial for noise reduction in power spectral density estimation	- Averaging leads to loss of temporal resolution - Segment choice affects frequency resolution
STFT	- Computationally Efficient than other TF methods. - Suitable for real-time applications. - Easy to implement - Suitable for identifying time-frequency patterns in physiological signals	- TF Resolution Trade-off. - Lower Resolution than CWT. - Less flexibility due to fixed window width. - Windowing can cause spectral leakage
CWT	- Flexible time-frequency resolution - Handles non-stationary signals - Effective for time-frequency analysis of non-stationary physiological signals	- Computationally intensive - Wavelet selection affects performance
DWT	- Multi-resolution analysis - Sparse representation - Suitable for multi-scale feature extraction in physiological signal processing	- Lack of standard decomposition approach - Risk of losing information
RP	- Visualizes recurrent patterns - Valuable for detecting recurrent patterns and nonlinear dynamics in physiological signals	- Parameters like delay and threshold affect results - No direct frequency information
Reshaping	- Simplifies input representation into a 2D matrix - Easily integrates with CNN architectures.	- Loss of sequential information. - Fixed windowing may not suit all temporal dynamics. - May distort original signal characteristics. - Sensitivity to parameter choice and signal type

across studies. Currently, there exists a lack of standardized practices for evaluating the performance of deep learning models applied to physiological signal analysis. This variability in evaluation methodologies makes it challenging to compare results across studies and hinders the progress of the field as a whole. Therefore, there is a pressing need for the establishment of standardized benchmarks, datasets, evaluation metrics, and experimental protocols to ensure fair and meaningful comparisons between different algorithms and approaches.

While deep learning models have shown remarkable success in automated feature learning and classification of physiological signals, it is essential to acknowledge the limitations and potential biases inherent in these models. Deep learning models are often regarded as ''black boxes,'' making it challenging to interpret their decisions and understand the underlying mechanisms driving their predictions. As such, there is a growing interest in developing transparent and interpretable deep learning models that can provide insights into how they arrive at their decisions. Techniques such as attention mechanisms, layer-wise relevance propagation, and saliency maps can help elucidate the important features and patterns learned by deep learning models, enabling clinicians and researchers to trust and interpret their predictions more effectively.

This emerging field still presents multiple challenges regarding optimal STI techniques, neural network architecture, model interpretability, and computational efficiency. As methods mature, real-world clinical translation will necessitate multi-modal frameworks fusing physiological signals with clinical context. Ultimately, intelligent integration of signal processing and deep learning promises to unlock clinically relevant insights from complex physiological data.

VI. CONCLUSION

In summary, this paper reviewed innovative techniques and recent advances in converting physiological signals into 2D image representations to enable automated feature learning using CNNs. A systematic analysis was presented, spanning diverse applications in EEG, EMG, and ECG signal processing. The relative merits of employing different STI transformations, deep network architectures, and training

methodologies were discussed. The review synthesized key insights from current literature and highlighted challenges and promising research directions at the intersection of deep learning and physiological signal analysis. This comprehensive overview of state-of-the-art methods aims to catalyze continued innovations in designing effective end-to-end systems for extracting clinically valuable information from multidimensional physiological data using advanced machine learning.

REFERENCES

[1] L. Wang and Z. Meng, "Multichannel two-dimensional convolutional neural network based on interactive features and group strategy for Chinese sentiment analysis," *Sensors*, vol. 22, no. 3, p. 714, Jan. 2022, doi: 10.3390/s22030714.

[2] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Comput. Intell. Mag.*, vol. 8, no. 2, pp. 20–33, May 2013, doi: 10.1109/MCI.2013.2247823.

[3] V. Gupta, M. Mittal, V. Mittal, and N. K. Saxena, "A critical review of feature extraction techniques for ECG signal analysis," *J. Inst. Eng. India B*, vol. 102, no. 5, pp. 1049–1060, Oct. 2021, doi: 10.1007/s40031-021-00606-5.

[4] H. F. Posada-Quintero, J. P. Florian, Á. D. Orjuela-Cañón, and K. H. Chon, "Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity," *Amer. J. Physiol. Regul. Integr. Comparative Physiol.*, vol. 311, pp. 582–591, Sep. 2016, doi: 10.1152/ajpregu.00180.2016.

[5] T. Wang, C. Lu, G. Shen, and F. Hong, "Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," *PeerJ*, vol. 7, p. e7731, Sep. 2019, doi: 10.7717/peerj.7731.

[6] H. Alaskar, "Convolutional neural network application in biomedical signals," *J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 45–59, 2018, doi: 10.15640/jcsit.v6n2a5.

[7] D. Shah, K. G. Gopan, and N. Sinha, "An investigation of the multidimensional (1D vs. 2D vs. 3D) analyses of EEG signals using traditional methods and deep learning-based methods," *Frontiers Signal Process.*, vol. 2, pp. 1–15, Jul. 2022, doi: 10.3389/frsip.2022.936790.

[8] A. K. Singh and S. Krishnan, "ECG signal feature extraction trends in methods and applications," *Biomed. Eng. OnLine*, vol. 22, no. 1, pp. 1–36, Mar. 2023, doi: 10.1186/s12938-023-01075-1.

[9] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C. M. Hill, and P. R. White, "Signal processing techniques applied to human sleep EEG signals—A review," *Biomed. Signal Process. Control*, vol. 10, pp. 21–33, Mar. 2014, doi: 10.1016/j.bspc.2013.12.003.

[10] P. Boonyakitanont, A. Lek-uthai, K. Chomtho, and J. Songsiri, "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101702, doi: 10.1016/j.bspc.2019.101702.

[11] S. K. Pahuja and K. Veer, "Recent approaches on classification and feature extraction of EEG signal: A review," *Robotica*, vol. 40, no. 1, pp. 77–101, Jan. 2022, doi: 10.1017/s0263574721000382.

[12] K. P. Ayodele, W. O. Ikezogwo, M. A. Komolafe, and P. Ogunbona, "Supervised domain generalization for integration of disparate scalp EEG datasets for automatic epileptic seizure detection," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103757, doi: 10.1016/j.compbiomed.2020.103757.

[13] Y. Zhang, Y. Guo, P. Yang, W. Chen, and B. Lo, "Epilepsy seizure prediction on EEG using common spatial pattern and convolutional neural network," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 465–474, Feb. 2020, doi: 10.1109/JBHI.2019.2933046.

[14] J. Jin, Y. Miao, I. Daly, C. Zuo, D. Hu, and A. Cichocki, "Correlation-based channel selection and regularized feature optimization for MI-based BCI," *Neural Netw.*, vol. 118, pp. 262–270, Oct. 2019, doi: 10.1016/j.neunet.2019.07.008.

[15] M. D. Basar, A. D. Duru, and A. Akan, "Emotional state detection based on common spatial patterns of EEG," *Signal, Image Video Process.*, vol. 14, no. 3, pp. 473–481, Apr. 2020, doi: 10.1007/s11760-019-01580-8.

[16] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398, doi: 10.1016/j.ymssp.2020.107398.

[17] F. E. Aswad, G. V. T. Djogdom, M. J.-D. Otis, J. C. Ayena, and R. Meziane, "Image generation for 2D-CNN using time-series signal features from foot gesture applied to select cobot operating mode," *Sensors*, vol. 21, no. 17, p. 5743, Aug. 2021, doi: 10.3390/s21175743.

[18] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.

[19] Y. Wu, F. Yang, Y. Liu, X. Zha, and S. Yuan, "A comparison of 1-D and 2-D deep convolutional neural networks in ECG classification," Oct. 2018. Accessed: Feb. 16, 2024. [Online]. Available: https://arxiv.org/abs/1810.07088v1

[20] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.

[21] E. Izci, M. Değirmenci, M. Akif Özdemir, and A. Akan, "ECG arrhythmia detection with deep learning," in *Proc. 28th Signal Process. Commun. Appl. Conf. (SIU)*, Oct. 2020, pp. 1–4, doi: 10.1109/SIU49456.2020.9302219.

[22] T. Joon Jun, H. M. Nguyen, D. Kang, D. Kim, D. Kim, and Y.-H. Kim, "ECG arrhythmia classification using a 2-D convolutional neural network," 2018, *arXiv:1804.06812*.

[23] H. S. Nogay and H. Adeli, "Detection of epileptic seizure using pretrained deep convolutional neural network and transfer learning," *Eur. Neurol.*, vol. 83, no. 6, pp. 602–614, Jan. 2021, doi: 10.1159/000512985.

[24] D. Borra, V. Mondini, E. Magosso, and G. R. Müller-Putz, "Decoding movement kinematics from EEG using an interpretable convolutional neural network," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107323, doi: 10.1016/j.compbiomed.2023.107323.

[25] M. Rashid, N. Sulaiman, A. P. P. A. Majeed, R. M. Musa, A. F. A. Nasir, B. S. Bari, and S. Khatun, "Current status, challenges, and possible solutions of EEG-based brain–computer interface: A comprehensive review," *Frontiers Neurorobotics*, vol. 14, pp. 1–35, Jun. 2020, doi: 10.3389/fnbot.2020.00025.

[26] A. Herrel, V. Schaerlaeken, C. Ross, J. Meyers, K. Nishikawa, V. Abdala, A. Manzano, and P. Aerts, "Electromyography and the evolution of motor control: Limitations and insights," *Integrative Comparative Biol.*, vol. 48, no. 2, pp. 261–271, Aug. 2008, doi: 10.1093/icb/icn025.

[27] G. R. Naik, S. E. Selvan, and H. T. Nguyen, "Single-channel EMG classification with ensemble-empirical-mode-decomposition-based ICA for diagnosing neuromuscular disorders," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 7, pp. 734–743, Jul. 2016, doi: 10.1109/TNSRE.2015.2454503.

[28] J. Qi, G. Jiang, G. Li, Y. Sun, and B. Tao, "Surface EMG hand gesture recognition system based on PCA and GRNN," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 6343–6351, May 2020, doi: 10.1007/s00521-019-04142-8.

[29] X. Luo, L. Yang, H. Cai, R. Tang, Y. Chen, and W. Li, "Multi-classification of arrhythmias using a HCRNet on imbalanced ECG datasets," *Comput. Methods Programs Biomed.*, vol. 208, Sep. 2021, Art. no. 106258, doi: 10.1016/j.cmpb.2021.106258.

[30] U. R. Acharya, Y. Hagiwara, J. E. W. Koh, S. L. Oh, J. H. Tan, M. Adam, and R. S. Tan, "Entropies for automated detection of coronary artery disease using ECG signals: A review," *Biocybern. Biomed. Eng.*, vol. 38, no. 2, pp. 373–384, Jan. 2018, doi: 10.1016/j.bbe.2018.03.001.

[31] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[32] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[33] Z. Khan, N. Yahya, K. Alsaih, and F. Meriaudeau, "Segmentation of prostate in MRI images using depth separable convolution operations," in *Intelligent Human Computer Interaction* (Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Deutschland, Germany: Springer, 2021, pp. 132–141, doi: 10.1007/978-3-030-68449-5_14.

[34] M. Sun, Z. Song, X. Jiang, J. Pan, and Y. Pang, "Learning pooling for convolutional neural network," *Neurocomputing*, vol. 224, pp. 96–104, Feb. 2017, doi: 10.1016/j.neucom.2016.10.049.

[35] Z. Khan, N. Yahya, K. Alsaih, S. S. A. Ali, and F. Meriaudeau, "Evaluation of deep neural networks for semantic segmentation of prostate in T2W MRI," *Sensors*, vol. 20, no. 11, p. 3183, Jun. 2020, doi: 10.3390/s20113183.

[36] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights into Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[37] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[38] Z. Khan, N. Yahya, K. Alsaih, M. I. Al-Hiyali, and F. Meriaudeau, "Recent automatic segmentation algorithms of MRI prostate regions: A review," *IEEE Access*, vol. 9, pp. 97878–97905, 2021, doi: 10.1109/ACCESS.2021.3090825.

[39] M. Buscema, "Back propagation neural networks," in *Substance Use & Misuse*. Oxford, U.K.: Taylor & Francis, 1998, doi: 10.3109/10826089809115863.

[40] M. A. Ozdemir, D. H. Kisa, O. Guren, and A. Akan, "Hand gesture classification using time–frequency images and transfer learning based on CNN," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103787, doi: 10.1016/j.bspc.2022.103787.

[41] M. Naz, J. H. Shah, M. A. Khan, M. Sharif, M. Raza, and R. Damaševič ius, "From ECG signals to images: A transformation based approach for deep learning," *PeerJ Comput. Sci.*, vol. 7, p. e386, Feb. 2021, doi: 10.7717/peerj-cs.386.

[42] M. Aslan, "CNN based efficient approach for emotion recognition," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7335–7346, Oct. 2022, doi: 10.1016/j.jksuci.2021.08.021.

[43] Z. Khademi, F. Ebrahimi, and H. M. Kordy, "A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105288, doi: 10.1016/j.compbiomed.2022.105288.

[44] B. Li, J. Wang, Z. Guo, and Y. Li, "Automatic detection of schizophrenia based on spatial–temporal feature mapping and LeViT with EEG signals," *Expert Syst. Appl.*, vol. 224, Aug. 2023, Art. no. 119969, doi: 10.1016/j.eswa.2023.119969.

[45] J. J. Bird, D. R. Faria, L. J. Manso, P. P. S. Ayrosa, and A. Ekárt, "A study on CNN image classification of EEG signals represented in 2D and 3D," *J. Neural Eng.*, vol. 18, no. 2, Apr. 2021, Art. no. 026005, doi: 10.1088/1741-2552/abda0c.

[46] B. R. Nayana and P. Geethanjali, "Analysis of statistical time-domain features effectiveness in identification of bearing faults from vibration signal," *IEEE Sensors J.*, vol. 17, no. 17, pp. 5618–5625, Sep. 2017, doi: 10.1109/JSEN.2017.2727638.

[47] A. Skoura, "Detection of lead-lag relationships using both time domain and time-frequency domain; An application to wealth-to-income ratio," *Economies*, vol. 7, no. 2, p. 28, Apr. 2019, doi: 10.3390/economies7020028.

[48] H. Witte and M. Wacker, "Time-frequency techniques in biomedical signal analysis: A tutorial review of similarities and differences," *Methods Inf. Med.*, vol. 52, no. 4, pp. 279–296, 2013, doi: 10.3414/me12-01-0083.

[49] S. Krishnan and Y. Athavale, "Trends in biomedical signal feature extraction," *Biomed. Signal Process. Control*, vol. 43, pp. 41–63, May 2018, doi: 10.1016/j.bspc.2018.02.008.

[50] J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "The fast Fourier transform and its applications," *IEEE Trans. Educ.*, vol. E-12, no. 1, pp. 27–34, Mar. 1969.

[51] *FourierAnalysisUno*. Accessed: Oct. 24, 2023. [Online]. Available: https://books.google.co.in/books?hl=en&lr=&id=ix2iCQ-o9x4C&oi=fnd&pg=PA1&dq=fourier+analysis+and+its+applications+gerald+b.folland&ots=iZzRHsG0ZB&sig=m6dBqJUvTWAoA8XmF6nvgxO2LeE&redir_esc=y#v=onepage&q=fourier%20analysis%20and%20its%20applications%20gerald%20b.folland&f=false

[52] S. Krishnan, "Advanced analysis of biomedical signals," in *Biomedical Signal Analysis for Connected Healthcare*. Cambridge, MA, USA: Academic, 2021, pp. 157–222, doi: 10.1016/B978-0-12-813086-5.00003-7.

[53] R. N. Youngworth, B. B. Gallagher, and B. L. Stamper, "An overview of power spectral density (PSD) calculations," *Proc. SPIE*, vol. 5869, Aug. 2005, Art. no. 58690U, doi: 10.1117/12.618478.

[54] J. Greenberg and B. Delgutte, "Course materials for HST.582J/6.555J/16.456J, biomedical signal and image processing, spring 2007," *MIT OpenCourseWare*. Accessed: Feb. 16, 2024. [Online]. Available: https://ocw.mit.edu/courses/hst-582j-biomedical-signal-and-image-processing-spring-2007/65956b1f5c262e9118b195b077811d70_ch4_dft.pdf

[55] J. O. Pinzón-Arenas, R. Jiménez-Moreno, and A. Rubiano, "Percentage estimation of muscular activity of the forearm by means of EMG signals based on the gesture recognized using CNN," *Sens. Bio-Sensing Res.*, vol. 29, Aug. 2020, Art. no. 100353, doi: 10.1016/j.sbsr.2020.100353.

[56] S. Nikkonen, H. Korkalainen, S. Kainulainen, S. Myllymaa, A. Leino, L. Kalevo, A. Oksenberg, T. Leppänen, and J. Töyräs, "Estimating daytime sleepiness with previous night electroencephalography, electrooculography, and electromyography spectrograms in patients with suspected sleep apnea using a convolutional neural network," *Sleep*, vol. 43, no. 12, pp. 1–7, Dec. 2020, doi: 10.1093/sleep/zsaa106.

[57] V. Gupta, M. Mittal, V. Mittal, and A. Gupta, "ECG signal analysis using CWT, spectrogram and autoregressive technique," *Iran J. Comput. Sci.*, vol. 4, no. 4, pp. 265–280, Dec. 2021, doi: 10.1007/s42044-021-00080-8.

[58] A. J. R. Simpson, "Time-frequency trade-offs for audio source separation with binary masks," 2015, *arXiv:1504.07372*.

[59] M. K. Kıymık, I. Güler, A. Dizibüyük, and M. Akın, "Comparison of STFT and wavelet transform methods in determining epileptic seizure activity in EEG signals for real-time application," *Comput. Biol. Med.*, vol. 35, no. 7, pp. 603–616, Oct. 2005, doi: 10.1016/j.compbiomed.2004.05.001.

[60] S. Behbahani, H. Ahmadieh, and S. Rajan, "Feature extraction methods for electroretinogram signal analysis: A review," *IEEE Access*, vol. 9, pp. 116879–116897, 2021, doi: 10.1109/ACCESS.2021.3103848.

[61] N. Qu, Z. Li, J. Zuo, and J. Chen, "Fault detection on insulated overhead conductors based on DWT-LSTM and partial discharge," *IEEE Access*, vol. 8, pp. 87060–87070, 2020, doi: 10.1109/ACCESS.2020.2992790.

[62] S. Chaudhary, S. Taran, V. Bajaj, and A. Sengur, "Convolutional neural network based approach towards motor imagery tasks EEG signals classification," *IEEE Sensors J.*, vol. 19, no. 12, pp. 4494–4500, Jun. 2019, doi: 10.1109/JSEN.2019.2899645.

[63] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[64] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhysics Lett.*, vol. 4, no. 9, p. 973, 1987.

[65] B. M. Mathunjwa, Y.-T. Lin, C.-H. Lin, M. F. Abbod, and J.-S. Shieh, "ECG arrhythmia classification by using a recurrence plot and convolutional neural network," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102262, doi: 10.1016/j.bspc.2020.102262.

[66] A. Stallone, A. Cicone, and M. Materassi, "New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, Sep. 2020, doi: 10.1038/s41598-020-72193-2.

[67] M. Barbosh, P. Singh, and A. Sadhu, "Empirical mode decomposition and its variants: A review with applications in structural health monitoring," *Smart Mater. Struct.*, vol. 29, no. 9, Sep. 2020, Art. no. 093001, doi: 10.1088/1361-665x/aba539.

[68] N. I. Hasan and A. Bhattacharjee, "Deep learning approach to cardiovascular disease classification employing modified ECG signal from empirical mode decomposition," *Biomed. Signal Process. Control*, vol. 52, pp. 128–140, Jul. 2019, doi: 10.1016/j.bspc.2019.04.005.

[69] Q. Liu, J. Cai, S.-Z. Fan, M. F. Abbod, J.-S. Shieh, Y. Kung, and L. Lin, "Spectrum analysis of EEG signals using CNN to model patient's consciousness level based on anesthesiologists' experience," *IEEE Access*, vol. 7, pp. 53731–53742, 2019, doi: 10.1109/ACCESS.2019.2912273.

[70] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, Jun. 2019, Art. no. 031001, doi: 10.1088/1741-2552/ab0ab5.

[71] A. A. Ein Shoka, M. M. Dessouky, A. El-Sayed, and E. E.-D. Hemdan, "EEG seizure detection: Concepts, techniques, challenges, and future trends," *Multimedia Tools Appl.*, vol. 82, no. 27, pp. 42021–42051, Nov. 2023, doi: 10.1007/s11042-023-15052-2.

[72] S. Mekruksavanich and A. Jitpattanakul, "Effective detection of epileptic seizures through EEG signals using deep learning approaches," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 4, pp. 1937–1952, Dec. 2023, doi: 10.3390/make5040094.

[73] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Netw.*, vol. 124, pp. 202–212, Apr. 2020, doi: 10.1016/j.neunet.2020.01.017.
[74] T. S. Cleatus and M. Thungamani, "Epileptic seizure detection using spectral transformation and convolutional neural networks," *J. Inst. Eng. India B*, vol. 103, no. 4, pp. 1115–1125, Aug. 2022, doi: 10.1007/s40031-021-00693-4.
[75] B. Mandhouj, M. A. Cherni, and M. Sayadi, "An automated classification of EEG signals based on spectrogram and CNN for epilepsy diagnosis," *Anal. Integr. Circuits Signal Process.*, vol. 108, no. 1, pp. 101–110, Jul. 2021, doi: 10.1007/s10470-021-01805-2.
[76] M. Rashed-Al-Mahfuz, M. A. Moni, S. Uddin, S. A. Alyami, M. A. Summers, and V. Eapen, "A deep convolutional neural network method to detect seizures and characteristic frequencies using epileptic electroencephalogram (EEG) data," *IEEE J. Transl. Eng. Health Med.*, vol. 9, pp. 1–12, 2021, doi: 10.1109/JTEHM.2021.3050925.
[77] Ö. Türk and M. S. Özerdem, "Epilepsy detection by using scalogram based convolutional neural network from EEG signals," *Brain Sci.*, vol. 9, no. 5, p. 115, May 2019, doi: 10.3390/brainsci9050115.
[78] R. Hussein, S. Lee, and R. Ward, "Multi-channel vision transformer for epileptic seizure prediction," *Biomedicines*, vol. 10, no. 7, p. 1551, Jun. 2022, doi: 10.3390/biomedicines10071551.
[79] Q. Xin, S. Hu, S. Liu, L. Zhao, and Y.-D. Zhang, "An attention-based wavelet convolution neural network for epilepsy EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 957–966, 2022, doi: 10.1109/TNSRE.2022.3166181.
[80] Y. Gao, B. Gao, Q. Chen, J. Liu, and Y. Zhang, "Deep convolutional neural network-based epileptic electroencephalogram (EEG) signal classification," *Frontiers Neurol.*, vol. 11, pp. 1–11, May 2020, doi: 10.3389/fneur.2020.00375.
[81] G. Wang, D. Wang, C. Du, K. Li, J. Zhang, Z. Liu, Y. Tao, M. Wang, Z. Cao, and X. Yan, "Seizure prediction using directed transfer function and convolution neural network on intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2711–2720, Dec. 2020, doi: 10.1109/TNSRE.2020.3035836.
[82] Z. Deng, C. Li, R. Song, X. Liu, R. Qian, and X. Chen, "EEG-based seizure prediction via hybrid vision transformer and data uncertainty learning," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106401, doi: 10.1016/j.engappai.2023.106401.
[83] S. Rukhsar and A. K. Tiwari, "Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals," *Comput. Methods Programs Biomed.*, vol. 242, Dec. 2023, Art. no. 107856, doi: 10.1016/j.cmpb.2023.107856.
[84] R. V. Godoy, T. J. S. Reis, P. H. Polegato, G. J. G. Lahr, R. L. Saute, F. N. Nakano, H. R. Machado, A. C. Sakamoto, M. Becker, and G. A. P. Caurin, "EEG-based epileptic seizure prediction using temporal multi-channel transformers," 2022, *arXiv:2209.11172*.
[85] A. Taebi and H. Mansy, "Time-frequency distribution of seismocardiographic signals: A comparative study," *Bioengineering*, vol. 4, no. 4, p. 32, Apr. 2017, doi: 10.3390/bioengineering4020032.
[86] G. Xu, X. Shen, S. Chen, Y. Zong, C. Zhang, H. Yue, M. Liu, F. Chen, and W. Che, "A deep transfer convolutional neural network framework for EEG signal classification," *IEEE Access*, vol. 7, pp. 112767–112776, 2019, doi: 10.1109/ACCESS.2019.2930958.
[87] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019, doi: 10.1109/TNSRE.2019.2915621.
[88] M.-A. Li, J.-F. Han, and J.-F. Yang, "Automatic feature extraction and fusion recognition of motor imagery EEG using multilevel multiscale CNN," *Med. Biol. Eng. Comput.*, vol. 59, pp. 2037–2050, Aug. 2021, doi: 10.1007/s11517-021-02396-w.
[89] Y. Hou, L. Zhou, S. Jia, and X. Lun, "A novel approach of decoding EEG four-class motor imagery tasks via scout ESI and CNN," *J. Neural Eng.*, vol. 17, no. 1, Feb. 2020, Art. no. 016048, doi: 10.1088/1741-2552/ab4af6.
[90] M. T. Sadiq, M. Z. Aziz, A. Almogren, A. Yousaf, S. Siuly, and A. U. Rehman, "Exploiting pretrained CNN models for the development of an EEG-based robust BCI framework," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105242, doi: 10.1016/j.compbiomed.2022.105242.
[91] N. Mammone, C. Ieracitano, and F. C. Morabito, "A deep CNN approach to decode motor preparation of upper limbs from time–frequency maps of EEG signals at source level," *Neural Netw.*, vol. 124, pp. 357–372, Apr. 2020, doi: 10.1016/j.neunet.2020.01.027.
[92] M. S. Al-Quraishi, I. Elamvazuthi, T. B. Tang, M. S. Al-Qurishi, S. H. Adil, M. Ebrahim, and A. Borboni, "Decoding the user's movements preparation from EEG signals using vision transformer architecture," *IEEE Access*, vol. 10, pp. 109446–109459, 2022, doi: 10.1109/ACCESS.2022.3213996.
[93] A. Keutayeva and B. Abibullaev, "Exploring the potential of attention mechanism-based deep learning for robust subject-independent motor-imagery based BCIs," *IEEE Access*, vol. 11, pp. 107562–107580, 2023, doi: 10.1109/ACCESS.2023.3320561.
[94] M. X. Cohen, "A better way to define and describe Morlet wavelets for time-frequency analysis," *NeuroImage*, vol. 199, pp. 81–86, Oct. 2019, doi: 10.1016/j.neuroimage.2019.05.048.
[95] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.
[96] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44317–44328, 2019, doi: 10.1109/ACCESS.2019.2908285.
[97] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, "EEG-based emotion recognition with deep convolutional neural networks," *Biomed. Eng./Biomedizinische Technik*, vol. 66, no. 1, pp. 43–57, Feb. 2021, doi: 10.1515/bmt-2019-0306.
[98] T. Song, W. Zheng, S. Liu, Y. Zong, Z. Cui, and Y. Li, "Graph-embedded convolutional neural network for image-based EEG emotion recognition," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 3, pp. 1399–1413, Jul. 2022, doi: 10.1109/TETC.2021.3087174.
[99] S. Liu, X. Wang, L. Zhao, J. Zhao, Q. Xin, and S.-H. Wang, "Subject-independent emotion recognition of EEG signals based on dynamic empirical convolutional neural network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 5, pp. 1710–1721, Sep. 2021, doi: 10.1109/TCBB.2020.3018137.
[100] S. Hwang, K. Hong, G. Son, and H. Byun, "Learning CNN features from DE features for EEG-based emotion recognition," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1323–1335, Aug. 2020, doi: 10.1007/s10044-019-00860-w.
[101] F. Wang, S. Wu, W. Zhang, Z. Xu, Y. Zhang, C. Wu, and S. Coleman, "Emotion recognition with convolutional neural network and EEG-based EFDMs," *Neuropsychologia*, vol. 146, Sep. 2020, Art. no. 107506, doi: 10.1016/j.neuropsychologia.2020.107506.
[102] L. Farokhah, R. Sarno, and C. Fatichah, "Simplified 2D CNN architecture with channel selection for emotion recognition using EEG spectrogram," *IEEE Access*, vol. 11, pp. 46330–46343, 2023, doi: 10.1109/ACCESS.2023.3275565.
[103] P. Pandey and K. R. Seeja, "Subject independent emotion recognition system for people with facial deformity: An EEG based approach," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 2311–2320, Feb. 2021, doi: 10.1007/s12652-020-02338-8.
[104] J. Wang and M. Wang, "Review of the emotional feature extraction and classification using EEG signals," *Cognit. Robot.*, vol. 1, pp. 29–40, Jan. 2021, doi: 10.1016/j.cogr.2021.04.001.
[105] S. K. Khare and V. Bajaj, "Time–frequency representation and convolutional neural network-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021, doi: 10.1109/TNNLS.2020.3008938.
[106] P. Saparia, A. Patel, H. Shah, K. Solanki, A. Patel, and M. Sahayata, "Schizophrenia: A systematic review," *J. Clin. Experim. Psychol.*, vol. 9, no. 1, pp. 65–70, 2022. Accessed: Feb. 16, 2024. [Online]. Available: https://www.iomcworld.org/open-access/schizophrenia-a-systematic-review-94323.html
[107] S. K. Khare, V. Bajaj, and U. R. Acharya, "SPWVD-CNN for automated detection of schizophrenia patients using EEG signals," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2021.3070608.
[108] A. Shalbaf, S. Bagherzadeh, and A. Maghsoudi, "Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals," *Phys. Eng. Sci. Med.*, vol. 43, no. 4, pp. 1229–1239, Dec. 2020, doi: 10.1007/s13246-020-00925-9.
[109] U. Budak, V. Bajaj, Y. Akbulut, O. Atila, and A. Sengur, "An effective hybrid model for EEG-based drowsiness detection," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7624–7631, Sep. 2019, doi: 10.1109/JSEN.2019.2917850.
[110] S. V. Faraone et al., "The world federation of ADHD international consensus statement: 208 evidence-based conclusions about the disorder," *Neurosci. Biobehavioral Rev.*, vol. 128, pp. 789–818, Sep. 2021, doi: 10.1016/j.neubiorev.2021.01.022.

[111] M. Moghaddari, M. Z. Lighvan, and S. Danishvar, "Diagnose ADHD disorder in children using convolutional neural network based on continuous mental task EEG," *Comput. Methods Programs Biomed.*, vol. 197, Dec. 2020, Art. no. 105738, doi: 10.1016/j.cmpb.2020.105738.

[112] L. Dubreuil-Vall, G. Ruffini, and J. A. Camprodon, "Deep learning convolutional neural networks discriminate adult ADHD from healthy individuals on the basis of event-related spectral EEG," *Frontiers Neurosci.*, vol. 14, pp. 1–12, Apr. 2020, doi: 10.3389/fnins.2020.00251.

[113] J. Zhang, R. Yao, W. Ge, and J. Gao, "Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG," *Comput. Methods Programs Biomed.*, vol. 183, Jan. 2020, Art. no. 105089, doi: 10.1016/j.cmpb.2019.105089.

[114] S. Datta and N. V. Boulgouris, "Recognition of grammatical class of imagined words from EEG signals using convolutional neural network," *Neurocomputing*, vol. 465, pp. 301–309, Nov. 2021, doi: 10.1016/j.neucom.2021.08.035.

[115] Y. Sun, C. Yang, Z. Xu, and Y. Lu, "Recurrence plot-assisted detection of focal/non-focal EEG signals using ensemble deep features," *J. Med. Biol. Eng.*, vol. 43, no. 2, pp. 176–184, Apr. 2023, doi: 10.1007/s40846-023-00785-0.

[116] C. Ieracitano, N. Mammone, A. Bramanti, A. Hussain, and F. C. Morabito, "A convolutional neural network approach for classification of dementia stages based on 2D-spectral representation of EEG recordings," *Neurocomputing*, vol. 323, pp. 96–107, Jan. 2019, doi: 10.1016/j.neucom.2018.09.071.

[117] Z. Aslan and M. Akin, "Automatic detection of schizophrenia by applying deep learning over spectrogram images of EEG signals," *Traitement du Signal*, vol. 37, no. 2, pp. 235–244, Apr. 2020, doi: 10.18280/ts.370209.

[118] M. N. A. Tawhid, S. Siuly, H. Wang, F. Whittaker, K. Wang, and Y. Zhang, "A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from EEG," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0253094, doi: 10.1371/journal.pone.0253094.

[119] A. Miltiadous, E. Gionanidis, K. D. Tzimourta, N. Giannakeas, and A. T. Tzallas, "DICE-Net: A novel convolution-transformer architecture for Alzheimer detection in EEG signals," *IEEE Access*, vol. 11, pp. 71840–71858, 2023, doi: 10.1109/ACCESS.2023.3294618.

[120] A. Qayyum, I. Razzak, M. Tanveer, M. Mazher, and B. Alhaqbani, "High-density electroencephalography and speech signal based deep framework for clinical depression diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 4, pp. 2587–2597, Jul. 2023, doi: 10.1109/TCBB.2023.3257175.

[121] M. A. Mulkey, H. Huang, T. Albanese, S. Kim, and B. Yang, "Supervised deep learning with vision transformer predicts delirium using limited lead EEG," *Sci. Rep.*, vol. 13, no. 1, pp. 1–8, May 2023, doi: 10.1038/s41598-023-35004-y.

[122] T. J. Wei, A. R. B. Abdullah, N. B. M. Saad, N. B. M. Ali, and T. N. S. B. T. Zawawi, "Featureless EMG pattern recognition based on convolutional neural network," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, p. 1291, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1291-1297.

[123] D.-C. Oh and Y.-U. Jo, "Classification of hand gestures based on multi-channel EMG by scale average wavelet transform and convolutional neural network," *Int. J. Control, Autom. Syst.*, vol. 19, no. 3, pp. 1443–1450, Mar. 2021, doi: 10.1007/s12555-019-0802-1.

[124] Y. Yamanoi, Y. Ogiri, and R. Kato, "EMG-based posture classification using a convolutional neural network for a myoelectric hand," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101574, doi: 10.1016/j.bspc.2019.101574.

[125] M. D. Dere and B. Lee, "A novel approach to surface EMG-based gesture classification using a vision transformer integrated with convolutive blind source separation," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 1, pp. 181–192, Jan. 2024, doi: 10.1109/JBHI.2023.3330289.

[126] K.-T. Kim, S. Park, T.-H. Lim, and S. J. Lee, "Upper-limb electromyogram classification of reaching-to-grasping tasks based on convolutional neural networks for control of a prosthetic hand," *Frontiers Neurosci.*, vol. 15, pp. 1–10, Oct. 2021, doi: 10.3389/fnins.2021.733359.

[127] N. Duan, L.-Z. Liu, X.-J. Yu, Q. Li, and S.-C. Yeh, "Classification of multichannel surface-electromyography signals based on convolutional neural networks," *J. Ind. Inf. Integr.*, vol. 15, pp. 201–206, Sep. 2019, doi: 10.1016/j.jii.2018.09.001.

[128] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng, "Hand gesture recognition using compact CNN via surface electromyography signals," *Sensors*, vol. 20, no. 3, p. 672, Jan. 2020, doi: 10.3390/s20030672.

[129] V. Shanmuganathan, H. R. Yesudhas, M. S. Khan, M. Khari, and A. H. Gandomi, "R-CNN and wavelet feature extraction for hand gesture recognition with EMG signals," *Neural Comput. Appl.*, vol. 32, no. 21, pp. 16723–16736, Nov. 2020, doi: 10.1007/s00521-020-05349-w.

[130] R. X. Gao and R. Yan, "Wavelet packet transform," in *Wavelets*. Boston, MA, USA: Springer, 2011, pp. 69–81, doi: 10.1007/978-1-4419-1545-0_5.

[131] E. Kim, J. Shin, Y. Kwon, and B. Park, "EMG-based dynamic hand gesture recognition using edge AI for human–robot interaction," *Electronics*, vol. 12, no. 7, p. 1541, Mar. 2023, doi: 10.3390/electronics12071541.

[132] X. Chen, Y. Li, R. Hu, X. Zhang, and X. Chen, "Hand gesture recognition based on surface electromyography using convolutional neural network with transfer learning method," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 4, pp. 1292–1304, Apr. 2021, doi: 10.1109/JBHI.2020.3009383.

[133] Y. Cheng, G. Li, M. Yu, D. Jiang, J. Yun, Y. Liu, Y. Liu, and D. Chen, "Gesture recognition based on surface electromyography-feature image," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 6, pp. 1–13, Mar. 2021, doi: 10.1002/cpe.6051.

[134] W. Wei, Y. Wong, Y. Du, Y. Hu, M. Kankanhalli, and W. Geng, "A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface," *Pattern Recognit. Lett.*, vol. 119, pp. 131–138, Mar. 2019, doi: 10.1016/j.patrec.2017.12.005.

[135] K. Yang, M. Xu, X. Yang, R. Yang, and Y. Chen, "A novel EMG-based hand gesture recognition framework based on multivariate variational mode decomposition," *Sensors*, vol. 21, no. 21, p. 7002, Oct. 2021, doi: 10.3390/s21217002.

[136] R. V. Godoy, G. J. G. Lahr, A. Dwivedi, T. J. S. Reis, P. H. Polegato, M. Becker, G. A. P. Caurin, and M. Liarokapis, "Electromyography-based, robust hand motion classification employing temporal multi-channel vision transformers," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10200–10207, Oct. 2022, doi: 10.1109/LRA.2022.3192623.

[137] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014, doi: 10.1109/TSP.2013.2288675.

[138] J. Tryon and A. L. Trejos, "Evaluating convolutional neural networks as a method of EEG-EMG fusion," *Frontiers Neurorobotics*, vol. 15, pp. 1–20, Nov. 2021, doi: 10.3389/fnbot.2021.692183.

[139] L. Lu, J. Mao, W. Wang, G. Ding, and Z. Zhang, "A study of personal recognition method based on EMG signal," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 4, pp. 681–691, Aug. 2020, doi: 10.1109/TBCAS.2020.3005148.

[140] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, and C. Ou, "Automatic ECG classification using continuous wavelet transform and convolutional neural network," *Entropy*, vol. 23, no. 1, p. 119, Jan. 2021, doi: 10.3390/e23010119.

[141] J. Huang, B. Chen, B. Yao, and W. He, "ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network," *IEEE Access*, vol. 7, pp. 92871–92880, 2019, doi: 10.1109/ACCESS.2019.2928017.

[142] A. M. Alqudah, S. Qazan, L. Al-Ebbini, H. Alquran, and I. A. Qasmieh, "ECG heartbeat arrhythmias classification: A comparison study between different types of spectrum representation and convolutional neural networks architectures," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 10, pp. 4877–4907, Oct. 2022, doi: 10.1007/s12652-021-03247-0.

[143] R. Panda, S. Jain, R. Tripathy, and U. R. Acharya, "Detection of shockable ventricular cardiac arrhythmias from ECG signals using FFREWT filter-bank and deep convolutional neural network," *Comput. Biol. Med.*, vol. 124, Sep. 2020, Art. no. 103939, doi: 10.1016/j.compbiomed.2020.103939.

[144] A. Asgharzadeh-Bonab, M. C. Amirani, and A. Mehri, "Spectral entropy and deep convolutional neural network for ECG beat classification," *Biocybern. Biomed. Eng.*, vol. 40, no. 2, pp. 691–700, Apr. 2020, doi: 10.1016/j.bbe.2020.02.004.

[145] M. Wasimuddin, K. Elleithy, A. Abuzneid, M. Faezipour, and O. Abuzaghleh, "Multiclass ECG signal analysis using global average-based 2-D convolutional neural network modeling," *Electronics*, vol. 10, no. 2, p. 170, Jan. 2021, doi: 10.3390/electronics10020170.

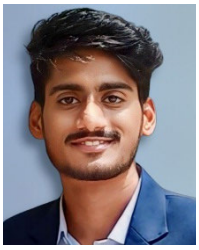
[146] Y. Ji, S. Zhang, and W. Xiao, "Electrocardiogram classification based on faster regions with convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2558, Jun. 2019, doi: 10.3390/s19112558.

[147] Z. K. Senturk, "From signal to image: An effective preprocessing to enable deep learning-based classification of ECG," *Mater. Today, Proc.*, vol. 81, pp. 1–9, Jan. 2023, doi: 10.1016/j.matpr.2022.10.223.

[148] Z. Ahmad, A. Tabassum, L. Guan, and N. M. Khan, "ECG heartbeat classification using multimodal fusion," *IEEE Access*, vol. 9, pp. 100615–100626, 2021, doi: 10.1109/ACCESS.2021.3097614.
[149] R. Arvanaghi, S. Danishvar, and M. Danishvar, "Classification cardiac beats using arterial blood pressure signal based on discrete wavelet transform and deep convolutional neural network," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103131, doi: 10.1016/j.bspc.2021.103131.
[150] D. A. AlDuwaile and M. S. Islam, "Using convolutional neural network and a single heartbeat for ECG biometric recognition," *Entropy*, vol. 23, no. 6, p. 733, Jun. 2021, doi: 10.3390/e23060733.
[151] M.-G. Kim, H. Ko, and S. B. Pan, "A study on user recognition using 2D ECG based on ensemble of deep convolutional neural networks," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 5, pp. 1859–1867, May 2020, doi: 10.1007/s12652-019-01195-4.
[152] J. S. Kim, S. H. Kim, and S. B. Pan, "Personal recognition using convolutional neural network with ECG coupling image," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 5, pp. 1923–1932, May 2020, doi: 10.1007/s12652-019-01401-3.
[153] M. Hammad, S. Zhang, and K. Wang, "A novel two-dimensional ECG feature extraction and classification algorithm based on convolution neural network for human authentication," *Future Gener. Comput. Syst.*, vol. 101, pp. 180–196, Dec. 2019, doi: 10.1016/j.future.2019.06.008.
[154] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.
[155] Y. Sun, W. Jin, X. Si, X. Zhang, J. Cao, L. Wang, S. Yin, and D. Ming, "Continuous seizure detection based on transformer and long-term iEEG," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5418–5427, Nov. 2022, doi: 10.1109/JBHI.2022.3199206.

K. E. CH VIDYASAGAR received the B.Tech. degree in biomedical engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2009, the M.Tech. degree in biomedical engineering from VIT University, Vellore, India, in 2011, and the Ph.D. degree from the Center for Biomedical Engineering, Indian Institute of Technology Delhi, India. Since 2014, he has been contributing as an Assistant Professor with the Department of Biomedical Engineering, University College of Engineering, Osmania University, Hyderabad, India. Before joining Osmania University, he was a Research Assistant with the Department of Electrical Engineering, Faculty of Engineering, University of Malaya, from 2013 to 2014. His Ph.D. research with IIT Delhi was conducted under the quality improvement program sponsored by AICTE, the Government of India. His research interests include biomedical devices, signal processing, machine learning, rehabilitation engineering, and the tribo-corrosion of biomaterials. His dedication to academic and research pursuits underscores his commitment to advancing knowledge and innovation in the field of biomedical engineering.

K. REVANTH KUMAR (Student Member, IEEE) was born in Hyderabad, Telangana, India, in 2001. He received the Diploma degree in biomedical engineering from the Government Institute of Electronics, Secunderabad, India, in 2021. He is currently pursuing the bachelor's degree in biomedical engineering with the University College of Engineering, Osmania University, Hyderabad. His research interests include machine learning, deep learning, biomedical instrumentation, medical imaging, and biomedical signal processing.



G. N. K. ANANTHA SAI (Student Member, IEEE) was born in Hyderabad, Telangana, India, in 2002. He received the Diploma degree in biomedical engineering from the Government Institute of Electronics, Secunderabad, India, in 2021. He is currently pursuing the bachelor's degree in biomedical engineering with the University College of Engineering, Osmania University, Hyderabad. His research interests include machine learning, deep learning, medical imaging, and biomedical signal processing.

MUNAGALA RUCHITA (Student Member, IEEE) was born in Hyderabad, Telangana, India, in 2003. She is currently pursuing the bachelor's in biomedical engineering with the University College of Engineering, Osmania University, Hyderabad. Her research interests include machine learning, deep learning, medical imaging, and biomedical signal processing.

MANOB JYOTI SAIKIA (Member, IEEE) received the B.E. degree in electronics and communication engineering from Visvesvaraya Technological University, Belgaum, India, in 2009, the M.Tech. degree in bioelectronics from Tezpur University, Tezpur, Assam, India, in 2013, and the Ph.D. degree in electrical engineering from the University of Rhode Island, Kingston, RI, USA, in 2019. He is currently an Assistant Professor with the Electrical Engineering Department, University of North Florida. He was a Research Associate with the School of Engineering, Tufts University, from September 2019 to August 2022. He was also a Senior Research Associate with the Department of Engineering, Boston College, from May 2022 to August 2022. From January 2016 to July 2019, he was a Research Assistant in a project funded by the National Science Foundation, USA. From 2012 to 2015, he was awarded a Senior Research Fellowship from the Ministry of Science and Technology, India, working with Indian Institute of Science, Bengaluru, India. His research interests include biomedical instrumentation, sensors, neuroimaging (fNIRS and EEG), signal processing, machine learning, and the Internet of Things.