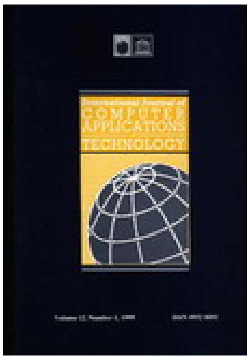


International Journal of Computer Applications in Technology

ISSN online: 1741-5047 - ISSN print: 0952-8091
https://www.inderscience.com/ijcat

A comprehensive review of electroencephalography data analytics

Marwa Saieed Khlief, Ali Kadhum Idrees

DOI: 10.1504/IJCAT.2023.10056441

Article History:

Received:	23 April 2022
Last revised:	27 May 2022
Accepted:	15 June 2022
Published online:	23 May 2023

A comprehensive review of electroencephalography data analytics

Marwa Saieed Khlief and Ali Kadhum Idrees*

Department of Computer Science,
University of Babylon,
Babylon, Iraq
Email: marwah.khlaif.gsci14@student.uobabylon.edu.iq
Email: ali.idrees@uobabylon.edu.iq
*Corresponding author

Abstract: This paper proposes a comprehensive review of Electroencephalography (EEG) data analytics. The EEG signal definition and the analysis process are presented. The public EEG data sets that were utilised by the researchers are explored. EEG data acquisition methods are investigated. This paper covers and summarises the work and techniques that have been done to compress EEG data. Significant approaches for feature extraction for EEG signal processing are illustrated. The collected features are then utilised to classify signals based on their properties. Machine learning techniques have become very important in this field in recent years because of their incredible ability to assess complicated volumes of data. Therefore, machine learning and deep learning for EEG data have been introduced. For researchers interested in EEG data analysis, this work can serve as a basic strategy and a roadmap.

Keywords: EEG; electroencephalography; EEG signal processing; data compression; machine learning; deep learning.

Reference to this paper should be made as follows: Khlief, M.S. and Idrees, A.K. (2023) 'A comprehensive review of electroencephalography data analytics', *Int. J. Computer Applications in Technology*, Vol. 71, No. 1, pp.78–88.

Biographical notes: Marwa Saieed Khlief received her BSc degree in Computer Science from the University of Babylon, Iraq in 2013. Her research interests include wireless sensor networks, EEG data compression and machine learning.

Ali Kadhum Idrees received his BSc and MSc degrees in Computer Science from the University of Babylon, Iraq in 2000 and 2003, respectively. He also received the PhD degree in Wireless Networks in 2015 from the University of Franche-Comte (UFC), France. Since 2018, he has been a Full Professor in Computer Science at the University of Babylon, Iraq. He has published over 45 research papers in international conferences, journals and book chapters. His research interests include wireless networks, WSNs, WBSNs, internet of things (IoT), IoMT, tactile internet, distributed computing, edge/fog computing, data mining, machine learning and optimisation in communication networks. He has actively served as the Technical Program Committee Member for various networking conferences and Reviewer for several international journals.

1 Introduction

The translation of vast amounts of data into usable information has become more crucial in different sectors in the era of 'big data'. It includes picture recognition, voice recognition and EEG signals, among other things. With the current exponential growth in the amount of data available, many studies have begun to focus on EEG research, taking into consideration the expectations created by the huge volume of data available, the large number of different file formats and the growing computational power of Artificial Intelligence (AI), which is now a powerful tool in the service of humans and companies. For example, IBM created a platform that analyses patient medical data to provide physicians with treatment alternatives. Most researchers prefer programming the method to extract meaningful information from EEG signals and computers since deep learning's prominence. This is owing to the machine learning algorithms' superior performance when dealing with a variety of real-world, complicated, and dynamic issues (using various methods, like those based on classification, regression and or unsupervised learning such as clustering) (Li et al., 2020).

EEG data compression has been shown to be an effective energy-saving strategy for reducing the quantity of data that must be transmitted across a network. Data compression reduces the amount of energy it takes to deliver compressed data by removing duplicated data (Ketshabetswe et al., 2021;

Titus and Sudhakar, 2020). Most research has so far focused on EEG processing, which can help researchers better comprehend the connections between brain activity and electrical signals. Using machine learning, we can evaluate, understand and extract sophisticated patterns from complicated inputs. People frequently convert the raw data into a wavelet or frequency before utilising it as input data since the signals collected are usually some forms of mixed noise and artefact combination. Raw EEG data, on the other hand, has been employed for anomaly classification concerns and brain activity decoding as Convolutional Neural Networks (CNN) have improved (Bidgoly et al., 2020).

In this regard, Figure 1 depicts the general process used in published research to extract, analyse and classify brain signals. This procedure is commonly broken down into four categories: signal acquisition, pre-processing, feature extraction and classification (Nicolas-Alonso and Gomez-Gil, 2012).

To begin, brain signals must be collected for future processing. As will be demonstrated, various techniques can be used by placing numerous electrodes on the surface of the head or inside the brain. Following that, the data collected must be pre-processed because it has certainly been altered by noise and exterior interferences. The noise in the signal is caused by the electrical power distribution system, nearby electrical devices, or by body functions such as sweating, breathing, eye blinking or body movements. Filtering techniques can be used to eliminate noise from a signal (Huster and Calhoun, 2018; Jiang et al., 2019). In addition, The EEG signal will be compressed before sending it to the destination across the network in the case of the remote patient monitoring. This can decrease the volume of transmitted data and increase the performance of network. Owing to the large amount of EEG signal data recorded, which essentially makes computation incredibly complex, feature extraction is achieved once the brain signal is free of interference and artefacts (Ai et al., 2019; Hira and Gillies, 2015). Feature extraction focuses on the reduction of data by generating new features from the initial collected data set that are non-redundant, contain the relevant information of the input data and allow for better classification by using a reduced representation acquired rather than the entire initial data set (Hira and Gillies, 2015). Li et al. (2020) introduced a review about using deep learning in the analysis of EEG. They are focused on using deep learning techniques like Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), Deep Belief Network (DBN) and Recurrent Neural Networks (RNN) to show their ability on analysing the non-fixed EEG data. They investigated some studies to show some applications in EEG signal processing and analysis such as human factors, neuromarketing, BCI and social interaction.

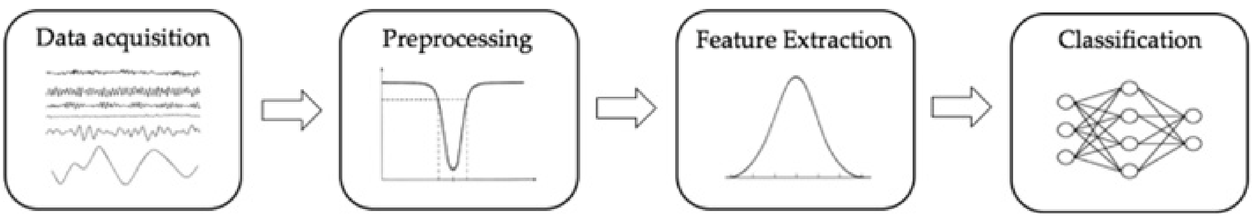
Bidgoly et al. (2020) proposed a survey about the authentication in the EEG and explain the main challenges that face these methods. This work is focused on some elements like required tasks by the user to achieve the authentication, EEG data sets and devices and the classification methods with their pre-processing approaches that are applied to EEG authentication (Bidgoly et al., 2020).

Luján et al. (2021) introduce a comprehensive survey about classification-based machine learning, classical techniques for signal processing, and neural recording for brain activity monitoring. Most of these studies did not introduce comprehensive review about the data analytics of EEG data.

The main contribution in this paper is to introduce a comprehensive review of Electroencephalography (EEG) data analytics. For researchers interested in EEG data analysis this work can serve as basic strategy and a roadmap. EEG signal processing is investigated via various studies about the compression techniques that have been done to compress EEG data. Several important approaches for feature extraction for EEG signal processing are illustrated. We survey some papers about machine and deep learning for EEG data classification. The fog computing is introduced as an important layer between the IoT devices and cloud for EEG data analytics.

This paper is organised as follows: The theoretical background of the EEG signals is introduced in the next section. Section 3 gives the public data sets on EEG. Section 4 explains the EEG acquisition. The EEG compression is presented in Section 5. In Section 6, the machine learning for EEG classification is introduced. The fog computing is introduced in Section 7. Further discussion is presented in Section 8. The conclusion and perspectives are given in Section 9.

Figure 1 EEG signal analysis process



2 Theoretical background of EEG

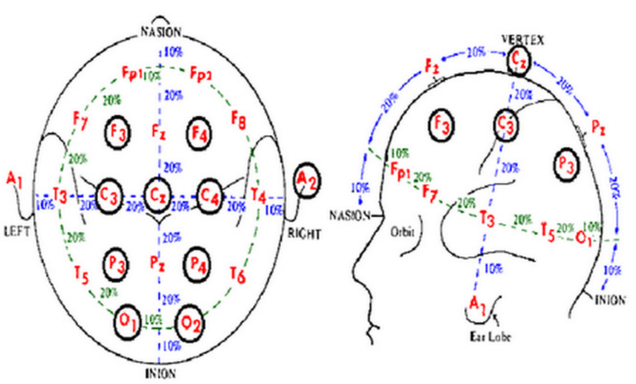
The Electroencephalogram (EEG) is a method of measuring electrical activity in the brain, commonly along the surface of the scalp. Ionic flows mediated by coordinated synaptic stimulation of brain neurons cause these electrical events (Schomer and Da Silva, 2012), and they manifest as rhythmic voltage variations with amplitudes ranging from 5 to 100 V and frequencies ranging between 0.5 Hz and 40 Hz. Examining the EEG's significant frequencies and amplitudes waves in different areas of the brain might reveal information about a person's physical or mental condition (Newson and Thiagarajan, 2018). Brain waves are categorised into 5 frequency ranges according to their frequency (Tangkraingkij, 2016; Blankertz et al., 2003): The delta waveform (14 Hz) is the slowest and has the highest amplitude in general. During deep sleep, the delta band may be visible in infants and adults.

- Children, fatigued adults and those reliving memories all have theta (4.88 Hz) frequency. Theta waves have an amplitude of less than 100 V in most cases.
- Active contemplation, concentration and focused attention are all connected to beta (12.25 Hz). Additionally, doing or watching others conduct physical movements boosts Beta power. Beta waves have an amplitude of less than 30 V in most cases.
- During multimodal sensory processing, gamma (above 25 Hz) is detected. The amplitude of gamma patterns is the smallest.

Because of this, it is essential to pay attention to not only the dominant frequency but also the recording from a specific part of the brain during the investigation of brain waves.

Electrodes with low impedance are put into operation in order to gather EEG signals. The electrodes can be applied using conductive gel, referred to as a 'wet electrode' or directly to the skin, referred to as a 'dry electrode'. The 10–20 standard (Fu et al., 2020) is one example of a set of rules for placing and identifying electrodes on the scalp. The electrodes in the (10–20) standard are along the latitude and longitude, respectively, and are categorised by the lobe to which they correspond. On the other hand, the poles in the right hemisphere may be assigned even numbers, while the poles in the left hemisphere may be assigned odd numbers (see Figure 2).

Figure 2 The placement of the electrode, the lobes of the brain are beneath each electrode in the 10–20 standard



3 Data sets types

This section presents the most often used public data sets in EEG data analytics.

3.1 Motor motion array/EEG image data from Physionet

One of the most prominent and widely available data sets created by the Brain-Computer Interface (BCI) 2000 (Schalk et al., 2004) technology is the Physionet EEG Motor Movement/Imagery data set. This data set contains EEG recordings from 109 healthy volunteers who completed a variety of motor and imaging activities. At a 160 Hz sampling rate for two 1-minute baseline runs (one with open eyes, the other closed), 64 channels of EEG waves were recorded for three 2-minute rounds for each of the four activities performed by each subject. Think about opening and shutting the left or right hand, as well as both fists and feet. Also, visualise opening and shutting both feet (DelPozo-Banos et al., 2015).

3.2 BCI Competition II

This large data set contains a complete record of real BCI performance from three trained individuals over the course of 10 sessions. The patient sat in a reclining chair in front of a video screen in each trial and was required to stay immobile throughout the performance. A total of 64 channels of EEG were recorded on the scalp, each referring to an electrode on the right ear (amplification 20,000; band-pass 0.1–60 Hz). At 160 Hz, all 64 channels were digitised and saved. Cursor movement was controlled online using just a minimal number of channels (LeCun et al., 2015).

3.2 BCI Competition III data set

The data sets comprise EEG recordings of one to five people engaged in a variety of activities. As part of one of these data sets, the EEG recordings of three people (identified as K3b, K6b and L1b) were made using the Neuroscan amplifier with 62 EEG channels (60 electrodes + 2 reference electrodes) and a sampling rate of 250 Hz. Participants in the research envisioned the movement of four body parts as a reaction to this. Triggers can be found on the left, right, foot and tongue. These data sets are used in a wide range of research projects (Alanazi, 2009).

3.3 BCI Competition IV data sets

The fourth BCI competition, featuring five distinct data sets, was held in 2008 to continue the BCI contests. More subjects and tasks are included in these data sets. As an illustration, in data set 3A (Altuwaijri and Muhammad, 2022), the EEG signals of nine people were recorded at a sample rate of 250 Hz on 22 EEG channels and 3 EOG channels. The participants were given four different tasks, including picturing left- and right-hand motions. Each session consists of six runs separated by brief rests. Several articles have also used these data sets (Lawhern et al., 2018).

3.4 EEG database at UCI KDD

This data set was collected using 64 electrodes and a 256 Hz sampling rate. The study enrolled a total of 122 males, both healthy and inebriated. The investigations, which were available to all participants, employed images from the Snodgrass and Vanderwart (1980) collections as visual stimuli. Two stimulants were used, one of which lasted 300 milliseconds and the other of which lasted 1.6 seconds. If S1 and S2 are identical, then they were indeed asked to respond to the question. Certain circumstances necessitated the usage of a single trigger (Bay et al., 2000).

3.5 The EEG database for Australia

During an 11-year research at J H Hospital (2002–1991), 40 patients' EEG signals were analysed (20 men and 20 females). Using 23 electrodes and a sample rate of 167 Hz, the EEG was recorded for around 20 minutes with both eyes open and closed. In a variety of research, this data gathering method has been employed (Hunter et al., 2005).

3.6 Bonn database

The database is divided into five sets, each containing 100 records. Pre-processing, feature extraction and machine learning classification are the three steps of the decision-making process. We employ single channel EEG recordings as input to our categorisation system. We chose channel T5-O1 from the multi-channel TUH data, which is generated from the TCP montage presented by Obeid et al. Based on the same source, we evaluate the first 60 seconds of each recording in the data set for categorisation. The findings may be directly compared by selecting identical areas (Obeid and Picone, 2018). Table 1 presents an overview of the Bonn data set.

Table 1 Overview of Bonn data set

Set	*Patients*	*Setup*	*Phase*
A	healthy	surface EEG	open eyes
B	healthy	surface EEG	closed eyes
C	epilepsy	intracranial EEG	interictal
D	epilepsy	intracranial EEG	interictal
E	epilepsy	intracranial EEG	seizure

3.7 DEAP

The DEAP data set is well-known in the field of emotion recognition, but it is also utilised for EEG authentication in some investigations. EEG data was collected from 32 healthy subjects using 32 channels and a sample rate of 512 Hz. Each one received 40 one-minute movies, each of which evokes a different feeling in pride, pleasure, fulfilment, hope, despair and fear, to name only a few examples (Schalk et al., 2004).

4 EEG acquisition protocols

Extrinsic stimulus activities include extrinsic activity states, mental tasks, and stimuli, which are the three types of protocols used in EEG recording in general. Authentication procedures or validity may be affected by the protocol chosen. Resting states and mental processes, for example. Tasks requiring external stimuli, on the other hand, necessitate the use of extra equipment beyond EEG recording devices. To provide the appropriate stimulation, several devices are required. Noise and disruptions in the environment may readily affect simple activities like resting states. The 'signal-to-noise ratio' of mental processes and actions that are followed by external stimuli is higher. Event-related potentials, which are electrical potentials in the brain triggered by any kind of sensory event, information from the outside, or mental effort, are used to achieve this goal. In this event, the ERP time is just one second long. As 'challenge-based response' technologies, ERP systems reverse the manner in which a non-biometric authentication command requires the user to reply (Idrus et al., 2013).

The most common methodology for obtaining EEG readings is resting states (Di et al., 2019). The individual is instructed to be fully calm, generally sitting The EEG signal is then collected in a chair in a tranquil environment. Although both closed and open eyes, referred to as 'REC' and 'REO', are employed in this technique, the dominating frequency range and the more effective EEG channels differ. The central region produces the highest outcomes in REO, whereas the parietal region produces the best results in REC. Furthermore, during REC, band is the dominating band. This protocol's popularity stems from its simplicity. Furthermore, there are no further prerequisites or instructions; nonetheless, the surroundings must be peaceful, and the participant must be free of other mental activity or the findings will be skewed (Barry et al., 2007). ERPs elicited by visual stimuli are referred to as 'Visual Evoked Potentials' (VEPs). The individual is asked to silently read some unconnected material. External equipment is required to provide the stimuli, as well as temporal limitations, which are a limitation of all VEPs. On the other hand, VEP components have been found to be highly stable over time and can meet the permanence criteria of biometric authentication. Furthermore, there is no need to synchronise recordings in SSVEP (Gasper et al., 2011). Acoustic stimuli is a type of ERP that happens when you hear music or a specific tone; however, it is less common than VEP (Keirn and Aunon, 1990). It presents participants with four distinct styles of music, each of which elicits various emotions and interests. Musical tastes were also requested as a form of self-identification.

The EEG is recorded using a variety of methods and stimuli. For example, people are instructed to watch short movies with music that elicits various emotional responses in them. The combination of EEG and EOG data is another multi-model strategy for enhancing classification accuracy (Bhateja et al., 2019).

5 EEG compression

The Electroencephalography (EEG) signal's relevance is that it is used to interpret brain activity in the form of electrical patterns. Head traumas, epilepsy, seizures, brain tumours, dizziness and sleep deprivation all need EEG data to identify brain irregularities. As a result, greater bandwidth and storage space are necessary in IoT networks for effective data transmission and storage. As a result, one of the major issues in the IoT network's medical health monitoring system is detecting and reducing the patient's large amount of EEG data in order to improve the network's performance while maintaining the accuracy of the data received at the end destination.

The potential of health data compression algorithms has been emphasised as a result of these issues. Several ways have been investigated to limit the amount of delivered health data while extending the battery life of the device. Lossless data compression is used in situations when information must stay 100% intact when compared to the original data. Lossless compression technologies, on the other hand, can yield a lower data reduction ratio. It makes use of data similarity; no more data is introduced, and the original data does not lose any information when certain bits of data are eliminated (Gravina et al., 2017).

Al-Nassrawy et al. (2020) proposed a high-performance fractal compression scheme for EEG health network traffic. A feasible EEG fractals compression model is presented in this work for lowering EEG traffic transported from the Patient Data Aggregator (PDA) to the destination (doctor, smart hospitals, emergency response, etc.). By lowering network traffic, the suggested approach allows EEG patient data transfer and improves the Wireless Body Sensor Network. The calculated fractal block size was discovered to play a critical function in creating greater Compression Ratio (CR) and driving the requisite Percentage Residual Difference (PRD). The suggested model has entirely surpassed existing strategies in terms of both outcomes and performance. The resulting CR can be as high as 160 while maintaining a PRD of less than one.

Al-Nassrawy et al. (2022) proposed a novel Lossless EEG Compression Model Using Fractal Combined with Fixed-Length Encoding Technique. This research presented a lossless fractals compression method for reducing the amount of EEG data transferred from the gateway cloud Patient Data Aggregator (PDA). By lowering the quantity of data traffic through the network, the proposed approach improves data communication in WBSNs. This method is tested and compared to various existing approaches, with the findings indicating that the proposed method outperforms the others.

Hejrati et al. (2017) introduced an efficient, lossless multi-channel EEG compression based on channel clustering. A new lossless compression approach has been proposed that is both efficient and easy. Correlation between channels and within channels is used. A pre-processing step is performed in the first stage to extract intra-channel correlation using the differential pulse code modulation approach. The centroid of each cluster is determined and encoded using arithmetic coding, and the channels are grouped into discrete clusters. In the second step, the difference between the centroid and other channels' data is determined and encoded using arithmetic coding in each cluster.

Karimu and Azadi (2016) proposed a lossless EEG compression method using DCT and Huffman coding. Based on the characteristics of the DCT frequency spectrum and Hoffmann coding, a lossless hybrid compression technique for EEG has been devised in this study. It generates DCT coefficients for EEG segments below 40 Hz (dominant components). The quantitative DCT parameters are then encoded using a Huffman encoder at the transmitter location. We add a zero set of DCT coefficients above 40 Hz at the receiver location and then rebuild EEG segments using inverted DCT. We used our technique on the University of Bone database's five groups (labelled A-E).

Maazouz et al. (2015) proposed a DCT-Based Algorithm for Multi-Channel Near-lossless EEG Compression. For compression purposes, the Electroencephalogram (EEG) signals are discussed in this article. In the temporal domain, the EEG signal is correlated and this fact is employed to compress the signal. Data compression is useful for lowering transmission speed, energy consumption and the amount of memory required for storage (reducing the cost accordingly). Lossy compression based on the Discrete Cosine Transform (DCT) is employed in this article. This is a conservative procedure that forms the foundation of the well-known JPEG still format. Quantisation causes the loss of information. The compression approach also employs entropy coding. The compression ratio and Percent Root mean-square Distortion (PRD) are used to assess the results obtained. Rajasekar and Pushpalatha (2020) propose a Huffman quantisation method for EEG compression. The discrete cosine transform and inverse discrete cosine transform are used to improve data privacy while reducing data complexity. A new lossless EEG data compression scheme is proposed in Idrees and Idrees (2022) for IoT networks. The agglomerative hierarchical clustering and Huffman encoding are the two steps of the EEG data compression technique.

Dao et al. (2015) presented a lossy compression technique for EEG signals. The Electroencephalogram (EEG) signal has been extensively utilised to evaluate brain processes and diagnose several brain-related disorders in this article. They're frequently recorded for a long period at good quality, which necessitates a lot of memory for storage and transmission. Signal compression is required to minimise signal size. Lossy compression techniques, as compared to lossless compression techniques, will yield a substantially greater Compression Ratio (CR) by making use of human cognition's limits. However, this comes at the expense of increased pressure distortion, which decreases the EEG signal's precision.

Titus and Sudhakar (2020) introduced a simple but efficient EEG data compression algorithm for neuromorphic applications. The research provides a simple and unique computational way for compressing MCEEG signals using Pseudo-spatial-standard Coding (n-SPC). After calibrating the signals, two processes, the spatial encoder and the pseudo encoder, work on the integer and fractional parts of the measured data, respectively. The technique delivers considerably improved signal quality. The uncompressed sleep spindle detection signal was extracted from an EEG recording and compared to two expert visual recordings available in the DREAMS Sleep Spindles database to determine efficacy. As a result, the suggested compression strategy may be applied to the recording, archiving, BCI systems and neural systems processes of MCEEG.

Birvinskas et al. (2015) proposed fast DCT algorithms for EEG data compression in embedded systems. The use of rapid Discrete Cosine Transform (DCT) algorithms for lossy EEG data reduction is discussed in this work. The signal is partitioned into eight samples using this method, and each set is DCT-transformed. Before transmission, the least-significant transform coefficients are eliminated and the inverse transform is filled with zeros. When high speed and minimal computing complexity are required, this technique can be applied in real-time embedded systems.

Alsenwi et al. (2016) suggested a performance analysis of hybrid lossy/lossless compression techniques for EEG data. Two lossless compression methods are utilised in this study to convert random EEG data into high-frequency data: Discrete Cosine Transformation (DCT) and Discrete Wave Transformation (DWT). As a result, using a lossless compression technique after lossy compression is a smart way to get a high-compression ratio while avoiding signal distortion. We employ two lossless compression methods: Run Length Encoding (RLE) and Arithmetic Encoding. To evaluate the effectiveness of the suggested system, the total compression and rebuilding decompression times, Compression Ratio (CR), Root Mean Square Error (RMSE) and Structural Similarity Index (SSIM) are examined.

Alsenwi et al. (2017) proposed a hybrid compression technique with data segmentation for electroencephalography data. Because of the extended recording duration, high-sampling rate and large number of electrodes used in medical applications, Electroencephalogram (EEG) data volumes are huge. As a result, more space and bandwidth are required to efficiently transport and retain data. As a result, EEG data compression is a critical issue for efficiently transmitting EEG data with less bandwidth and storing it in less space. An effective EEG compression technique is presented in this research. The EEG data is first separated into *N* segments and then processed using the Discrete Cosine Transform (DCT). The transformed parameters are subjected to a threshold procedure, with any values falling below the threshold being set to zero. Finally, the Run Length Encoding (RLE) technique is used to encode the obtained parameters. A reverse method can be used to recover the EEG signal. In this calculation, the total compression and reconstruction time (*T*), Compression Ratio (CR) and root mean error difference (PRD) ratio are calculated. In order to verify the effectiveness of the proposed algorithm, simulation results show a significant improvement in compression time using data segmentation.

Idrees et al. (2022) proposed a lossless data compression-based k-means clustering and Huffman encoding at the edge to decrease the volume of EEG data before sending it to the fog node in the Internet of Medical Things network. Campobello et al. (2021) presented a simple encoding compression method to reduce the EEG data at the limited resources IoT devices. For data compression, an optimal tensor truncation mechanism is required. In the proposed work, researchers first transform the multi-channel EEG signal as a tensor and then decide the compact tensor's optimal size. To depict the data embedded in the tensor, researchers accomplish tensor decomposition and acquire a core tensor of considerably smaller volume (Das and Kyal, 2021). Table 2 contains a summary about the compression techniques used for EEG signals.

Table 2 Summary of the compression techniques

Name (Year)	*Lossless*	*Lossy*	*Dataset*		*Technique*
			Bonn	*Others*	
Al-nassrawy et al. (2019)		✓	✓		Fractal compression
Al-nassrawy et al. (2022)	✓		✓		Fractal compression with fixed length encoding
Hejrati and Fathi (2017)	✓			✓	Arithmetic coding
Ykarimu and Azadi (2016)	✓		✓		DCT + Huffman coding
Maazouz et al. (2015)	✓			✓	DCT algorithm
Rajasekar and Pushpalatha (2020)	✓			✓	Huffman-based discrete cosine transforms
Idrees and Idrees (2021)	✓		✓		Hieratical Clustering + Huffman coding
Dao and Li (2015)		✓			Wavelet-based, filter-based, predictor and other non-wavelet compression
Sudhakar and Titus (2018)	✓	✓		✓	Normalised spatial pseudo code(n-SPC)
Birvinskas and Jusas (2015)		✓		✓	Fast DCT Algorithm
Alsenwi et al. (2016)	✓	✓		✓	DCT + DWT (lossy) RLE + arithmetic encoding(lossless)
Saeed et al. (2017)	✓	✓			DCT+RLE
Idrees et al. (2022)	✓		✓		K-means Clustering + Huffman coding
Campobello et al. (2021)		✓		✓	Simple and efficient encoding scheme
Das and Kyal (2021)		✓		✓	Tensor truncation method

6 Machine learning for EEG data classification

Most people preferred to train data using machine learning techniques such as Naïve Bayes or support machine learning due to a lack of data and low-performance computers. Early techniques sought to explicitly program the needed knowledge for certain tasks; nevertheless, they ran into problems when coping with complicated real-world situations since defining all of the information requested for an AI system to get excellent outcomes by hand is such a tough effort (Hameed and Idrees, 2022). Several methods are used to extract features of EEG data. In this section, some methods can be presented (Luján et al., 2021): Temporal analysis allows for the identification of normal and abnormal wave patterns in EEG signals, as well as the existence or absence of brain rhythms. EEG signals are frequently linked in time. As a result, time samples can be predicted by using autoregressive for stationary signals or adaptive autoregressive for nonstationary signals, such as Linear Prediction (LP) and Independent Component Analysis (ICA).

Temporal approaches are sometimes unable to give substantial characteristics. This can occur when a captured signal has weak temporal or spatial resolution, allowing only oscillatory activity to be preserved. These oscillations can potentially be represented by fundamental sinusoid functions using the Fourier Transform (FT). In these circumstances, the spectral analysis provides extra information by clarifying the EGG signal's prominent frequencies. Fast Fourier Transform (FFT), Short-Time Fourier Transform (STFT), spectrogram, Autoregressive Method (ARM) and Eigenvectors are the key approaches used in this situation, with band power and high-order spectra being the main characteristics recovered (HOS). In most cases, 'Short-Time Fourier Transform' (STFT) or 'spectrograms' provide superior results for neural signals since short-time windows are used to divide the data into tiny time intervals.

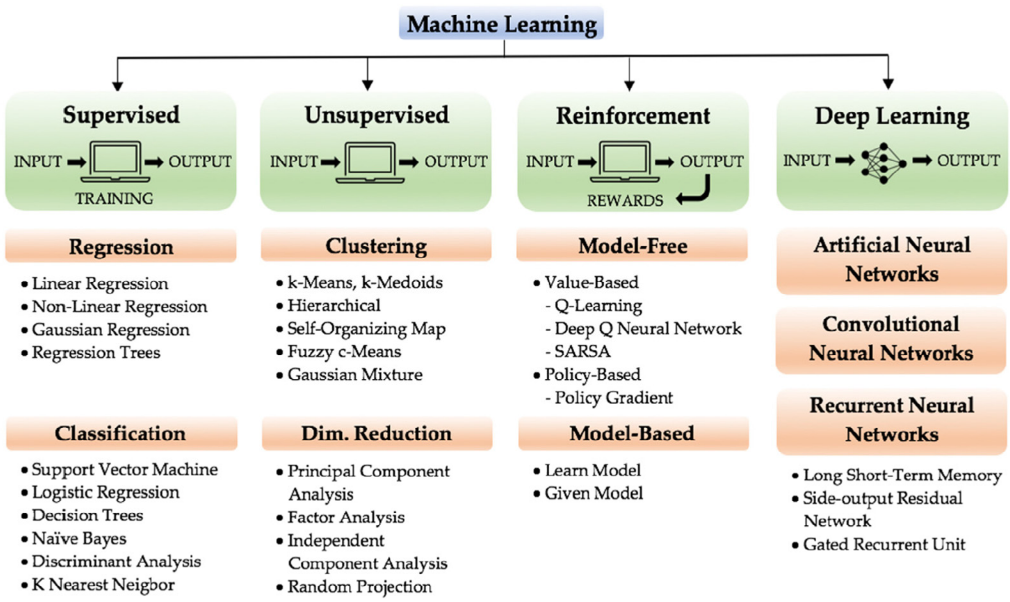
To enhance EEG studies, the benefits of temporal and frequency investigations can be combined. The signal breakdown for nonperiodic signals is commonly done using time-frequency transformations, which provide a lot of useful information. Various techniques, such as the Wigner–Ville distribution, scalogram, Hilbert–Huang spectrum or discrete wavelet transform, can be used to divide signals into single instantaneous frequencies throughout time (DWT). This approach has the benefit of converting non-linear, non-stationary brain input into linear and stationary components. Furthermore, this approach enables such a compromise between temporal and frequency resolutions, resulting in the best possible representation of the signals.

Machine learning, as we all know, offers a lot of potential for teaching AI systems cognitive capabilities via experience gained through training enormous volumes of data. However, feature extraction from raw data has limits. It may be necessary to devote a significant amount of effort to coding a complicated algorithm capable of manually extracting characteristics of raw data. Despite the limitations of these techniques, machine learning has been utilised to evaluate EEG data in a number of studies. Figure 3 shows the machine learning classification for EEG data (Jiang, 2021).

This section gives a quick overview of the machine learning techniques used in EEG categorisation. After successful feature extraction, the treated EEG data is ultimately ready for categorisation using machine learning techniques. Machine learning is described as computer models and algorithms that can learn and adapt without explicit instructions or human involvement from data and experience. Machine learning techniques are divided into four groups, as shown in Figure 3, deep learning, reinforcement learning, unsupervised learning and supervised learning (Hull, 2021).

Supervised learning methods use predetermined inputs and predefined outputs to build a model utilising data points with known outcomes. Following that, the system may use the obtained training and expertise to make judgments or predictions (i.e., predict future outputs using fresh inputs). Regression and classification are the two types of supervised learning available. Linear regression, non-linear regression, Gaussian process regression, and regression trees are examples of regression learning techniques that are commonly used to forecast statistical parameters. For classification output variables, classification algorithms such as 'Support Vector Machines' (SVM), 'logistic regression', 'decision trees', 'Naïve Bayes', 'discriminant analysis' and 'K-Nearest Neighbour (KNN)' are used. Only two classes of values are feasible, such as male–female, yes–no, true–false, and so on (Jo, 2021).

Figure 3 Machine learning for EEG data classification



Source: Luján et al. (2021).

Unsupervised learning, on the other hand, uses just input data sets and no outputs, inferring result patterns without any reference, therefore, the computer may learn from the data without any external supervision. Clustering and dimensionality reduction are two types of this technology. Clustering algorithms like *K*-means, *k*-medoids, hierarchical clustering, self-organising map, fuzzy c-means and Gaussian mixture match input data, which show similarities in clusters and classify them depending on the existence or missing of such similarities. 'Principal component analysis', 'factor analysis', 'independent component analysis' and 'random projection' are examples of dimensionality reduction techniques that reduce the number of input variables in a data set to optimise classification and great fit a predictor whilst still lowering risk of data loss. Unsupervised learning is capable of doing more complicated tasks than supervised learning, but it is also more difficult to implement into practice (Yan et al., 2022).

The algorithms used in reinforcement learning learn from their own expertise, whereas an agent learns from the environment in which they are placed to attain a goal. In other words, the algorithms are capable of determining the optimum decision based on a system that rewards good selections. The agent, the environment, the action, the policy and the reward are the essential components of a reinforcement learning system. The agent is software that has been programmed to complete a certain task. The environment is the physical or virtual world in which the agent operates. An action is a change made by an agent that alters the environment's condition, and a policy is the collection of rules that the agent uses to make decisions. A reward is a monetary value assigned to a positive or bad activity (Yan et al., 2022).

Model-free and model-based strategies are two types of reinforcement learning algorithms. Model-free algorithms, like trial-and-error algorithms, do not build an explicit model of the environment, but instead conduct operations on the environment to directly extract the best policy. Value-based techniques (like 'Q-learning', 'Deep Q Neural Network' (DQN) and 'State-Action-Reward-State-Action' (SARSA)) reflect an optimal strategy as a result of precisely guessing the value function in every state, whereas policy-based algorithms (like 'Q-learning', 'Deep Q Neural Network' (DQN) and 'State-Action-Reward-State-Action' (SARSA)) reflect an optimal strategy as a result of precisely guessing the value function in every state. Without modelling the value function, policy-based algorithms (such as policy gradient) evaluate the best policy. Model-based algorithms (such as learn and given model) build a concrete model of the world, which the agent then explores to learn. The model calculates the predicted reward and future state for each state and action (Yan et al., 2022).

Deep learning is primarily based on multi-layered (input, hidden and output layers) neural networks that train from massive quantities of data, imitating the functions and working system of the brain by calculating data with millions of neurons. The data set's input characteristics are provided to the neural network's input layer. The hidden layers are positioned between the input and output layers, and each layer provides an output from a set of weighted inputs. The number of hidden layers varies depending on the complexity of the issue being solved. For given inputs, the output layer generates the final results. Deep learning algorithms execute computations and forecasts periodically in each layer in this fashion, progressively learning and increasing the accuracy of the findings over time. 'Artificial Neural Networks' (ANN), 'Convolutional Neural Networks' (CNN) and 'recurrent neural networks' are the three main types of deep learning techniques (RNN). 'Artificial Neural Networks' (ANNs), also known as 'Simulated Neural Networks' (SNNs), are at the heart of deep learning algorithms and their structure is inspired by the human brain, replicating organic neuron function (Jiang, 2021).

Based on an activation function that aids the network in learning any complicated relationship between inputs and outputs, ANNs are capable of learning any non-linear function. Universal function approximators are another name for these networks. CNNs (or ConvNets) are neural networks that are primarily used to handle auto-correlated input. CNNs are built around filters, sometimes known as kernels, that use convolutional processes to extract meaningful characteristics from the input. According to CNNs, the initial hidden layers can detect basic patterns, the following levels can identify patterns, and the last hidden layers are specialised and can recognise complicated patterns. 'Recurrent Neural Networks' (RNNs) are neural networks with recurrent connections between hidden layers. This recurrent feedback guarantees that the incoming data contains sequential information. As a result, RNNs are commonly utilised to solve issues involving time series data. 'Long Short-Term Memory' (LSTM), Side-output Residual Network (SRN), and gated recurrent unit are the most often used techniques for developing RNNs (GRU) (Hull, 2021).

The data processing nature and the way of learning the classifiers from it represent the principal difference between deep learning and machine learning. The machine learning techniques use only the structured and labelled data for learning, while the deep learning can use structured and labelled data and unstructured and unlabelled data for learning. The deep learning techniques are able to use several layers to achieve feature extraction from data and improve classification. Table 3 summarises the key differences between the two classification approaches, focusing on several properties such as training, data format, algorithm, database size and applications. Finally, deep learning techniques can strengthen learning, which is a highly developed unsupervised learning process during which the classifier 'learns' to be more precise based on strong feedback from previous estimates (Luján et al., 2021).

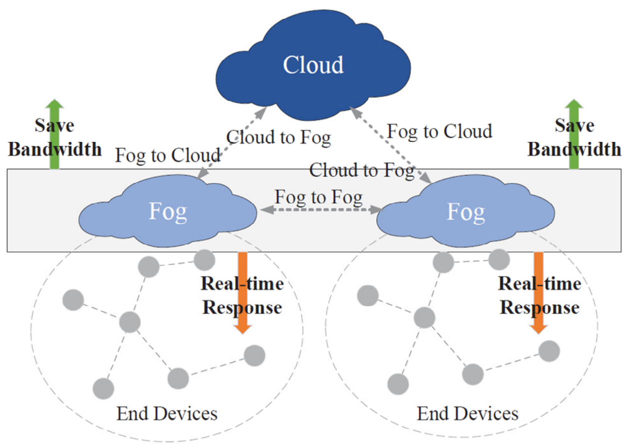
Table 3 Main differences between machine learning and deep learning

	Deep learning	*Machine learning*
Application	Complex tasks	Simple tasks
Algorithm	Neural network of algorithms	Variable algorithm
Data format	Unstructured data	Structured data
Training	System is self-learning.	System requires human for training
Database size	More than 10^6 od data points	Data set volume is manageable

7 Fog computing

The high bandwidth, high security and privacy and fast response in the case of high risk to the patient represent the main requirements for every smart medical system. These requirements cannot be provided by the current architecture of the IoT because it suffers from long delay response, limited bandwidth and increased cost for uploading data (Ferreira et al., 2021; Singh et al., 2021). Therefore, Cisco introduced fog computing as an addition to deal with these challenges. The cloud computing is extended by using the fog computing. This extension starts from the network core to the network edge. It provides several services like high powerful data processing and saving and analysing the collected IoT data in the middle point between the smart end IoT devices and the cloud platform. Figure 4 shows a fog computing based the architecture of IoT network (Yu et al., 2017). The fog computing layer is located closer to the smart IoT devices (data generators), providing several advantages like a fast response to the IoT sensor devices and a decrease in the volume of transferred data to the cloud, thus saving the bandwidth of the network. Hence, the huge volume of EEG data sensed from the patients can utilise this service to improve the performance of the network and provide a fast decision in the case of increasing the level of risk for the patients (Yu et al., 2017).

Figure 4 Fog computing based the architecture of IoT network



8 Further discussion

EEG is an additional sensitive objective measure, which could be utilised not only in the neuroscience research, but also in practical clinical applications that are closely related to human life and health. The EEG data analysis is made by using machine learning and deep learning to predict the epileptic seizure and the Alzheimer's disease. EEG data analytics provides a significant diagnostic benefit for different intracranial lesions like encephalitis, cerebral apoplexy, metabolic encephalopathy and brain tumours. The research direction and open issues for EEG data analytics include:

- Providing a new deep learning architecture for EEG data processing to provide high accuracy for predicting the level of the risk of the patient remains one open issue that requires extensive study and research.
- A decision-making model based on big EEG data at the fog layer is an open issue that needs to be investigated.
- EEG data authentication and security represent a big challenge in this field.

9 Conclusion and perspectives

A comprehensive review of EEG signal processing and analysis is given in this paper. EEG data collection and pre-processing are studied. Various studies about the techniques used to compress EEG data are being investigated. The feature extraction methods for EEG signal processing are introduced. Machine and deep learning have become more important in this sector in recent years due to their incredible ability to analyse large amounts of data. As a result, EEG data has been subjected to machine learning and deep learning. The main differences between machine learning and deep learning are presented. Fog computing is introduced as an important layer between the IoT devices and cloud for EEG data analytics. It can reduce the required bandwidth for sending data from IoT devices to the cloud and decrease the latency. Some open research directions are introduced. In the future, we plan to extend this study to include big data analytic techniques and their role in future EEG data analytics. The impact of EEG data analytics on the use of Tactile Internet architecture and 6G networks will also be investigated.

References

Ai, Q., Chen, A., Chen, K., Liu, Q., Zhou, T., Xin, S. and Ji, Z. (2019) 'Feature extraction of four-class motor imagery EEG signals based on functional brain network', *Journal of Neural Engineering*, Vol. 16, No. 2. Doi: 10.1088/1741-2552/ab0328.

Alanazi, M.N. (2009) 'Basic rules to build correct UML diagrams', *Proceedings of the International Conference on New Trends in Information and Service Science*, IEEE, pp.72–76.

Al-Nassrawy, K.K., Al-Shammary, D. and Idrees, A.K. (2020) 'High performance fractal compression for EEG health network traffic', *Procedia Computer Science*, Vol. 167, pp.1240–1249.

Al-Nassrawy, K.K., Idrees, A.K. and Al-Shammary, D. (2022) 'A novel lossless EEG compression model using fractal combined with fixed-length encoding technique', *AI and IoT for Sustainable Development in Emerging Countries: Challenges and Opportunities*, Springer International Publishing, Cham, pp.439–454.

Alsenwi, M., Ismail, T. and Mostafa, H. (2016) 'Performance analysis of hybrid lossy/lossless compression techniques for EEG data', *Proceedings of the 28th International Conference on Microelectronics (ICM)*, IEEE, pp.1–4.

Alsenwi, M., Saeed, M., Ismail, T., Mostafa, H. and Gabran, S. (2017) 'Hybrid compression technique with data segmentation for electroencephalography data', *Proceedings of the 29th International Conference on Microelectronics (ICM)*, IEEE, pp.1–4.

Altuwaijri, G.A. and Muhammad, G. (2022) 'A multibranch of convolutional neural network models for electroencephalogram-based motor imagery classification', *Biosensors*, Vol. 12, No. 1. Doi: 10.3390/bios12010022.

Barry, R.J., Clarke, A.R., Johnstone, S.J., Magee, C.A. and Rushby, J.A. (2007) 'EEG differences between eyes-closed and eyes-open resting conditions', *Clinical Neurophysiology*, Vol. 118, No. 12, pp.2765–2773.

Bay, S.D., Kibler, D., Pazzani, M.J. and Smyth, P. (2000) 'The UCI KDD archive of large data sets for data mining research and experimentation', *ACM SIGKDD Explorations Newsletter*, Vol. 2, No. 2, pp.81–85.

Bhateja, V., Gupta, A., Mishra, A. and Mishra, A. (2019) 'Artificial neural networks based fusion and classification of EEG/EOG signals', *Information Systems Design and Intelligent Applications: Proceedings of 5th International Conference INDIA,* Springer, Singapore, pp.141–148.

Bidgoly, A.J., Bidgoly, H.J. and Arezoumand, Z. (2020) 'A survey on methods and challenges in EEG based authentication', *Computers and Security*, Vol. 93. Doi: 10.1016/j.cose.2020.101788.

Birvinskas, D., Jusas, V., Martisius, I. and Damasevicius, R. (2015) 'Fast DCT algorithms for EEG data compression in embedded systems', *Computer Science and Information Systems*, Vol. 12, No. 1, pp.49–62.

Blankertz, B., Muller, K.R., Curio, G., Vaughan, T.M., Schalk, G., Wolpaw, J.R., Schlogl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T. and Schroder, M. (2004) 'The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials', *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 6, pp.1044–1051.

Campobello, G., Gugliandolo, G. and Donato, N. (2021) 'A simple and efficient near-lossless compression algorithm for multichannel EEG systems', *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 1150–1154.

Dao, P.T., Li, X.J. and Do, H.N. (2015) 'Lossy compression techniques for EEG signals', *International Conference on Advanced Technologies for Communications (ATC)*, IEEE, pp.154–159.

Das, S. and Kyal, C. (2021) 'Efficient multichannel EEG compression by optimal tensor truncation', *Biomedical Signal Processing and Control*, Vol. 68. Doi: 10.1016/j.bspc.2021.102749.

DelPozo-Banos, M., Travieso, C.M., Weidemann, C.T. and Alonso, J.B. (2015) 'EEG biometric identification: a thorough exploration of the time-frequency domain', *Journal of Neural Engineering*, Vol. 12, No. 5. Doi: 10.1088/1741-2560/12/5/056019.

Di, Y., An, X., He, F., Liu, S., Ke, Y. and Ming, D. (2019) 'Robustness analysis of identification using resting-state EEG signals', *IEEE Access*, Vol. 7, pp.42113–42122.

Ferreira, J., David, J.M.N., Braga, R., Campos, F., Ströele, V. and Aguiar, L.D. (2021) 'Ontology-based data integration for the Internet of Things in a scientific software ecosystem', *International Journal of Computer Applications in Technology*, Vol. 67, Nos. 2/3, pp.252–262.

Fu, Y., Zhao, J., Dong, Y. and Wang, X. (2020) 'Dry electrodes for human bioelectrical signal monitoring', *Sensors*, Vol. 20, No. 13. Doi: 10.3390/s20133651.

Gaspar, C.M., Rousselet, G.A. and Pernet, C.R. (2011) 'Reliability of ERP and single-trial analyses', *NeuroImage*, Vol. 58, No. 2, pp.620–629.

Gravina, R., Alinia, P., Ghasemzadeh, H. and Fortino, G. (2017) 'Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges', *Information Fusion*, Vol. 35, pp.68–80.

Hameed, M.K. and Idrees, A.K. (2022) 'Sensor device scheduling-based cuckoo algorithm for enhancing lifetime of cluster-based wireless sensor networks', *International Journal of Computer Applications in Technology*, Vol. 68, No. 1, pp.58–69.

Hejrati, B., Fathi, A. and Abdali-Mohammadi, F. (2017) 'Efficient lossless multi-channel EEG compression based on channel clustering', *Biomedical Signal Processing and Control*, Vol. 31, pp.295–300.

Hira, Z.M. and Gillies, D.F. (2015) 'A review of feature selection and feature extraction methods applied on microarray data', *Advances in Bioinformatics*. Doi: 10.1155/2015/198363.

Hull, I. (2021) *Machine Learning for Economics and Finance in TensorFlow 2*, Berkeley, CA, A Press.

Hunter, M.R.L.L., Smith, R.L., Hyslop, W., Rosso, O.A., Gerlach, R., Rostas, J.A.P., Williams, D.B. and Henskens, F. (2005) 'The Australian EEG database', *Clinical EEG and Neuroscience*, Vol. 36, No. 2, pp.76–81.

Huster, R.J. and Calhoun, V.D. (2018) 'Progress in EEG: multi-subject decomposition and other advanced signal processing approaches', *Brain Topography*, Vol. 31, pp.1–2.

Idrees, A.K., Idrees, S.K., Couturier, R. and Ali-Yahiya, T. (2022) 'An edge-fog computing-enabled lossless EEG data compression with epileptic seizure detection in IoMT networks', *IEEE Internet of Things Journal*, Vol. 9, No. 15, pp.13327–13337.

Idrees, S.K. and Idrees, A.K. (2022) 'New fog computing enabled lossless EEG data compression scheme in IoT networks', *Journal of Ambient Intelligence and Humanized Computing*, pp.1–14.

Idrus, S.Z.S., Cherrier, E., Rosenberger, C. and Schwartzmann, J.J. (2013) 'A review on authentication methods', *Australian Journal of Basic and Applied Sciences*, Vol. 7, No. 5, pp.95–107.

Jiang, H. (2021) *Machine Learning Fundamentals: A Concise Introduction*, Cambridge University Press.

Jiang, X., Bian, G.B. and Tian, Z. (2019) 'Removal of artifacts from EEG signals: a review', *Sensors*, Vol. 19, No. 5. Doi: 10.3390/s19050987.

Jo, T. (2021) *Machine Learning Foundations*: *Supervised, Unsupervised, and Advanced Learning*, Springer International Publishing, Cham.

Karimu, R.Y. and Azadi, S. (2016) 'Lossless EEG compression using the DCT and the Huffman coding', *Journal of Scientific and Industrial Research,* Vol. 75, No. 10, pp.615–620.

Keirn, Z.A. and Aunon, J.I. (1990) 'A new mode of communication between man and his surroundings', *IEEE Transactions on Biomedical Engineering*, Vol. 37, No. 12, pp.1209–1214.

Ketshabetswe, K.L., Zungeru, A.M., Mtengi, B., Lebekwe, C.K. and Prabaharan, S.R.S. (2021) 'Data compression algorithms for wireless sensor networks: a review and comparison', *IEEE Access*, Vol. 9, pp.136872–136891.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P. and Lance, B.J. (2018) 'EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces', *Journal of Neural Engineering*, Vol. 15, No. 5. Doi: 10.1088/1741-2552/aace8c.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, Vol. 521, pp.436–444.

Li, G., Lee, C.H., Jung, J.J., Youn, Y.C. and Camacho, D. (2020) 'Deep learning for EEG data analytics: a survey', *Concurrency and Computation: Practice and Experience*, Vol. 32, No. 18. Doi: 10.1002/cpe.5199.

Luján, M.Á., Jimeno, M.V., Mateo Sotos, J., Ricarte, J.J. and Borja, A.L. (2021) 'A survey on eeg signal processing techniques and machine learning: applications to the neurofeedback of autobiographical memory deficits in schizophrenia', *Electronics*, Vol. 10, No. 23. Doi: 10.3390/electronics10233037.

Maazouz, M., Kebir, S.T., Bengherbia, B., Toubal, A., Batel, N. and Bahri, N. (2015) 'A DCT-based algorithm for multi-channel near-lossless EEG compression', *Proceedings of the 4th International Conference on Electrical Engineering (ICEE)*, IEEE, pp.1–5.

Newson, J.J. and Thiagarajan, T.C. (2018) *EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. Front Hum Neurosci*. Doi: 10.3389/fnhum.2018.00521.

Nicolas-Alonso, L.F. and Gomez-Gil, J. (2012) 'Brain computer interfaces, a review', *Sensors*, Vol. 12, No. 2, pp.1211–1279.

Obeid, I. and Picone, J. (2018) 'Machine learning approaches to automatic interpretation of EEGs', *Signal Processing and Machine Learning for Biomedical Big Data*, CRC Press, pp.271–300.

Rajasekar, P. and Pushpalatha, M. (2020) 'Huffman quantization approach for optimized EEG signal compression with transformation technique', *Soft Computing*, Vol. 24, pp.14545–14559.

Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N. and Wolpaw, J.R. (2004) 'BCI2000: a general-purpose brain-computer interface (BCI) system', *IEEE Transactions on Biomedical Engineering*, Vol. 51, No. 6, pp.1034–1043.

Schomer, D.L. and Da Silva, F.L. (2012) *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, Lippincott Williams & Wilkins.

Singh, D., Thakur, A., Singh, M. and Sandhu, A. (2021) 'IoT implementation strategies amid COVID-19 pandemic', *International Journal of Computer Applications in Technology*, Vol. 65, No. 4, pp.389–398.

Snodgrass, J.G. and Vanderwart, M. (1980) 'A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity', *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 6, No. 2, pp.174–215.

Tangkraingkij, P. (2016) 'Significant frequency range of brain wave signals for authentication', *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Springer International Publishing, pp.103–113.

Titus, G. and Sudhakar, M.S. (2020) 'A simple but efficient EEG data compression algorithm for neuromorphic applications', *IETE Journal of Research*, Vol. 66, No. 3, pp.303–314.

Yan, D., Weng, J., Huang, S., Li, C., Zhou, Y., Su, H. and Zhu, J. (2022) 'Deep reinforcement learning with credit assignment for combinatorial optimization', *Pattern Recognition*, Vol. 124. Doi: 10.1016/j.patcog.2021.108466.

Yu, T., Wang, X. and Shami, A. (2017) 'A novel fog computing enabled temporal data reduction scheme in IoT systems', *Proceedings of the IEEE Global Communications Conference*, IEEE, pp.1–5.