Machine Learning Algorithms for Long Covid Effects Detection

Harit Ahuja

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY TORONTO, ONTARIO

NOVEMBER 2023

Abstract

In the realm of the Internet of Things (IoT) and Machine learning (ML), there is a growing demand for applications that can improve healthcare. By integrating sensors, cloud computing and ML we can create a powerful platform that enables insights into healthcare. Building upon these concepts, we propose a novel approach to address the widespread problem of long COVID. We utilize a wearable device to capture electroencephalogram (EEG) readings, which are then transformed through a set of processing steps into actionable decisions. We use a methodology that initiates data collection from a Cognitive-Motor Integration (CMI) task, followed by data preprocessing, feature engineering, and then the application of ML and advanced Deep Learning (DL) algorithms. To address challenges like data scarcity and privacy concerns, we generate synthetic data and train them using the same model as the original data for comparative analysis. Our method was tested on real cases and achieved prominent results: the CNN-LSTM model achieved 83% accuracy with original data and surged to 93% using synthetic data.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisors, Professor Marin Litoiu and Professor Lauren Sergio, for providing me support and guidance throughout my graduate studies. Under their guidance, I was able to get involved in the field of Machine Learning, Internet of Things and the intricacies of Neuroscience, which enhanced my research experience. Professor Litoiu consistently offered fresh viewpoints and essential insights that significantly contributed to my advancement in the Master's program. I am grateful for the support provided by Professor Sergio. Her extensive expertise in computational neuroscience, coupled with her vision, was crucial for this achievement. I deeply cherish the substantial research experience I acquired while working with her.

I would also like to extend my heartfelt thanks to Professor Heather Edgell and Dr. Smriti Badhwar for assistance in data collection for my thesis and for graciously providing access to their laboratory facilities. I would also like to thank all my friends and colleagues at the Center of Excellence for Research in Adaptive Systems (CERAS) lab who helped me throughout this thesis. Additionally, I am deeply appreciative of Professor Radu Campeanu and Professor Marios Fokaefs for dedicating their valuable time to serve as the committee members for my thesis.

Finally, I would like to thank my parents and my brother for their support and encouragement throughout my studies as a graduate student.

Table of Contents

List of Tables

List of Figures

Disclaimer

The content of this thesis is based on the paper titled “Machine Learning Algorithms for Detection of Visuomotor Neural Control Differences in Individuals with PASC and ME” authored by Harit Ahuja, Smriti Badhwar, Heather Edgell, Marin Litoiu, and Lauren E Sergio, currently under review by the neuroscience journal Frontiers.

As the first author of the paper, my contributions to the study were comprehensive, spanning the conception of the methodology through to the detailed analysis of results. Specifically, my contributions include:

1. **Data Curation:** Responsible for directly acquiring EEG data and overseeing the meticulous collection process from the study participants.

2. **Formal Analysis:** Engaged in comprehensive analytical procedures, utilizing statistical and machine learning techniques to extract meaningful insights from the EEG data.

3. **Investigation:** Conducted a thorough exploration of the fundamental research questions, employing a detailed and comparative study that differentiated between healthy individuals and those afflicted with long COVID, thereby highlighting key neurological contrasts critical to understanding the condition's prolonged impact.

4. **Methodology:** Devised the methodological framework for the study, incorporating the stages of data processing and the formulation of machine and deep learning models, along with the generation of synthetic data.

5. **Visualization:** Produced spectrograms of the EEG data, which played a vital role in augmenting the model's efficiency by providing a more precise visual representation of the complex data patterns.

Publications during Master Studies

1. Ahuja, H., Badhwar, S., Litoiu, M., Edgell, H. and Sergio, L. Cognitive-motor performance and associated brain activity shows differences in individuals with Post Acute Sequelae of SARS-CoV-2 (PASC) or Myalgic Encephalomyelitis (ME). Society for Neuroscience (SFN) 2022, San Diego, California.

2. K. Sarda, Z. Namrud, R. Rouf, H. Ahuja, M. Rasolroveicy, M. Litoiu, L. Shwartz, and I. Watts, "ADARMA: Auto-Detection and Auto-Remediation of Microservice Anomalies by Leveraging Large Language Models," in Proc. CASCON'23, Las Vegas, USA, Sep. 11–14, 2023.

3. H. Ahuja, K. Varadarajan, and M. Jammal, "Towards Smart Interaction: Hand Gesture Recognition Using Machine Learning in IoT Scenarios," in Proc. IEEE Global Conf. on Artificial Intelligence and Internet of Things, Dubai, UAE, 2023, to be presented on Dec. 10, 2023.

4. Ahuja, H., Badhwar, S., Litoiu, M., Edgell, H. and Sergio, L. Machine Learning Algorithms for Detection of Visuomotor Neural Control Differences in Individuals with PASC and ME. Frontiers in Neuroscience (under review).

5. Z. Namrud, Y. Rouf, R. Rouf, H. Ahuja, K. Sarda, M. Litoiu, I. Watts, A. De Magalhaes, C. Holliday, and S. Mostafa, "Automated Anomaly Remediation for Cloud-native Applications," presented at the CASCON, Toronto, Canada, 2022.

Chapter 1

1. Introduction

The COVID-19 pandemic has significantly impacted global health and economies [1] since its emergence in 2019. The virus primarily transmits through respiratory droplets and contaminated surfaces, manifesting symptoms such as fever, cough, and shortness of breath. A significant subset of patients experiences long-term complications called "long COVID" or post-acute sequelae of SARS-CoV-2 infection (PASC), characterized by cognitive dysfunction commonly known as "brain fog," among other symptoms [2]. These lingering effects pose substantial global challenges to healthcare systems, exhibiting symptomatic overlaps [3] with conditions such as Myalgic Encephalomyelitis (ME).

Consequently, emerging technologies like the Internet of Things (IoT) are assuming an increasingly pivotal role in the healthcare industry [4]. The pandemic has accelerated the digital transformation within healthcare, making it more amenable to integrating IoT frameworks. These frameworks offer exceptional capabilities for remote patient monitoring and granular medical data collection [5]. Wearable EEG devices, as part of this IoT ecosystem, hold promise in capturing real-time neurological data relevant to conditions like long COVID.

With patient privacy being a vital concern, there is a need for innovative strategies to augment the existing datasets while preserving the individual's anonymity. In line with this, our research also explores advanced data augmentation methods, including synthetic EEG data. We employ machine learning (ML) algorithms to process and analyze this data, thereby providing a standardized methodology to evaluate the performance of various ML models for early detection of cognitive impairments related to long COVID.

1.1 Problem and Motivation

The emergence of long COVID and its associated neurological symptoms has highlighted the necessity for efficient and consistent monitoring of affected individuals. One of the principal challenges is the scarcity of user-friendly, wearable EEG devices [6] for widespread monitoring of long COVID symptoms. This limitation presents a logistical gap, potentially causing many affected individuals to remain undiagnosed or not obtain prompt medical care. The availability of these devices has critical potential to enhance healthcare accessibility and cost-effectiveness. By leveraging wearable IoT devices, medical professionals can receive real-time data, enabling patients to be more proactive in their health management. This study aims to make detecting long COVID more accessible by introducing a novel method that employs a commercially available four-channel EEG headband.

As we explore the detection of neurological altercations tied to long COVID, a consistent and standardized methodology is yet to be established. The rising integration of machine learning in healthcare underscores the pressing need for a standardized approach. Venturing further into this domain, we find that the intricacies of EEG data interpretation require specialized techniques. In response to this, our study addresses this by transforming raw EEG data into a spectrogram-like matrix [7]. This transformation offers a comprehensive, two-fold representation encompassing spatial and temporal information, serving as an optimal input for machine learning models. By adopting this methodology, we aim to establish consistency and reproducibility in processing and interpreting EEG signals to ensure that results are comparable across various studies.

Obtaining large volumes of EEG data presents a significant challenge, especially when balancing capturing comprehensive neurological insights for model precision and preserving patient confidentiality. The necessity to understand the neurological symptoms of long COVID further emphasizes this challenge, with ethical considerations adding complexity to the data collection process. To bridge this gap, our approach includes generating synthetic EEG data using Wasserstein Generative Adversarial Networks (WGANs) [8]. This innovative method not only enriches our dataset but also solves the issues of utilizing individual EEG records, ensuring a more diverse and broader training set while maintaining the privacy of individuals. The aim is to provide a reliable and accessible method for early detection and monitoring of long COVID cognitive impairments, contributing to more efficient healthcare management of this complex condition.

1.2 Research Objectives and Questions

The goal of this research is to answer the following research questions (RQ):

RQ-1: Is it possible to distinguish EEG signals from the IoT device between healthy subjects and individuals suffering from long COVID?

RQ-2: Does the performance of models differ for machine learning techniques compared to deep learning algorithms?

RQ-3: Does the utilization of synthetically generated data result in superior performance compared to original data when applied to the same machine learning model?

The RQs are described below:

RQ-1: EEG signals demonstrate intricate patterns that could offer insights into various health conditions when analyzed. Leveraging IoT wearable technology, these signals can be recorded in real-time. Our research primarily focuses on determining if the analyzed EEG data, with the help of machine learning algorithms, can reliably distinguish healthy individuals and those affected by long COVID.

RQ-2: Machine learning (ML) and deep learning (DL) originate from the broader domain of artificial intelligence. However, they are employed in different scenarios depending on the type of data and problem at hand. While ML methods are suitable for structured data and explicitly featured engineering, DL, especially neural networks, excels at processing large volumes of unstructured data like images. Considering the complex nature of EEG readings, our study aims to evaluate if conventional ML algorithms can match the effectiveness of deep learning techniques in differentiating the data between our target groups.

RQ-3: The creation of synthetic data offers an innovative way to augment the existing datasets, particularly in fields where gathering data is constrained and challenging. This enhancement could improve the precision and adaptability of ML algorithms. Synthetic data becomes even more relevant given the usual concerns regarding EEG data's privacy and collection challenges. Our investigation aims to determine

whether models trained on synthetically generated data can outperform those trained on the original EEG dataset.

1.3 Thesis Contributions

The primary objective of this thesis is to design a methodology that seamlessly combines the capabilities of IoT wearables with advanced machine learning models. This specifically addresses the nuances, complexities and obstacles presented by long COVID. The notable contributions of this research encompass the following:

RC1: In this work, we introduce a methodology that holistically addresses the processing and analysis of EEG data acquired from IoT wearable devices. This method follows the systematic preprocessing of raw EEG signals, transforming them into a detailed spectrogram-like matrix. It ensures that the transformed data is a solid foundation for diverse machine learning techniques, meticulously designed to accurately discern variations in EEG readings between healthy individuals and those suffering from long COVD. Beyond refining the analysis, this methodology also holds the potential to serve as a benchmark in EEG-centric studies related to long COVID, ensuring uniformity and comparability in findings across varied research undertakings.

RC2: A crucial element of this thesis emphasizes a detailed comparison of traditional machine learning and evolving deep learning techniques. By analyzing how each method processes complex EEG data, we gained valuable insights useful for model selection in future research on similar data. Emerging from our empirical assessment, deep learning models, especially the CNN-LSTM approach, demonstrated remarkable proficiency, surpassing the traditional machine learning methods. The CNN-LSTM model leveraged the hierarchal feature extraction capabilities of Convolutional Neural Networks (CNN) [9] and the sequence learning strengths of Long Short-term Memory Networks (LSTM) [10]. This unique configuration adeptly captured both spatial and temporal aspects of EEG data, proving vital for early detection and understanding of long COVID's neurological effects. In contrast, among the machine learning models, the Random Forest attained the highest accuracy of 77%, highlighting the effectiveness of its ensemble-based approach.

RC3: One of the key aspects of our research is the exploration of synthetic data's potential in analyzing EEG readings. Our study aimed to discern if the models trained on artificial data could match or surpass

those trained on genuine EEG data, addressing privacy and data scarcity challenges. Our outcomes underscored the effectiveness of the synthetic approach, with models trained on this data attaining improved accuracy, at 93% over multiple runs, notably exceeding the 83% accuracy scored by those using the original data. Furthermore, statistical parallels between our original and synthetic datasets confirmed the consistency and reliability of our synthetic data generation approach, signalling its potential in enhancing deep learning models, particularly when acquiring vast volumes of data proves challenging.

1.4 Thesis Organization

This research is organized as follows: Chapter 2 offers a background on relevant concepts and technologies. Chapter 3 introduces the architecture and methodology. Chapter 4 presents the performance evaluation results. Chapter 5 concludes the study and suggests future directions.

Chapter 2

2. Background and Related Work

2.1 Background

In this section, we discuss foundational concepts crucial to the research. These include terms related to data collection like (EEG), data processing terminologies and applied methodologies like machine learning and deep learning.

2.1.1 Electroencephalogram (EEG)

Electroencephalogram (EEG) is a test used to record and track brain wave patterns using electrical signals present in the brain. The brain cells are active in constantly communicating with each other. This interaction between the brain cells is recorded as wavy lines on an EEG graph. EEG analysis is often helpful in diagnosing brain tumors, strokes, or sleep disorders. Two methods can obtain EEG: the non-invasive method, which does not require surgery as the electrodes are placed along the scalp. The other process is the invasive type, where the electrode is placed inside the brain but is known to damage the neuron when it finds its way.

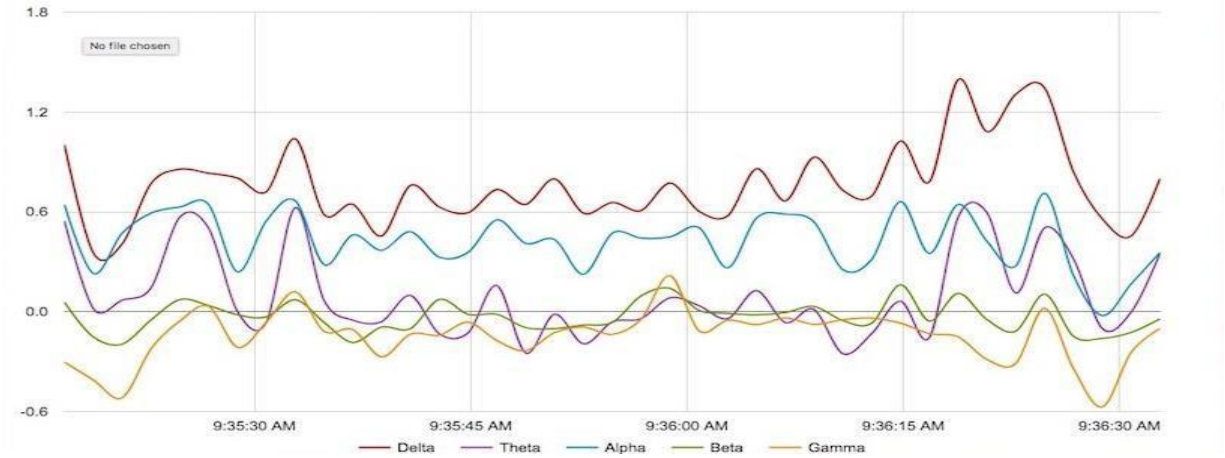

Figure 1 : EEG Waveforms Representing Various Brain Frequency Bands

2.1.2 Digital Signal Processing (DSP)

Digital signal processing (DSP) are real-world signals that help improve the accuracy and reliability of digital communication. It is used in numerous applications like image and video, telecommunication, and speech processing. The main goal of DSP is to manipulate digital signals to enhance their quality. Algorithms and mathematical models are used in the processing of a digital signal. Some standard algorithms used in DSP are signal quantization, filtering, and Fourier transform.

2.1.2.1 Time-Frequency Analysis

Time series analysis is a technique that uses statistical methods to notice the change in a set of given data points. Time series analysis helps to understand the trends and patterns that are important to understand to solve a problem. As an extension, time-frequency analysis delves deeper into the representation of these signals in both time and frequency domains. While various techniques facilitate this transformation, the Continuous Wavelet Transform (CWT) stands out for its adaptability in capturing intricate details across these domains.

2.1.2.2 Morlet Wavelet Transform

The Morlet wavelet transform is a time-frequency analysis technique based on the continuous wavelet transform (CWT). It is used to decompose a signal into a series of wavelets, each of which is characterized by a specific frequency and scale. The Morlet wavelet is defined by a sinusoidal oscillation modulated by a Gaussian function and is well-suited for analyzing signals that exhibit both frequency and temporal localization. The CWT and the Morlet wavelet transform are used in various applications, including signal denoising, image processing, and pattern recognition. One of the key advantages of the Morlet wavelet transform is that it provides a joint time-frequency representation of a signal, which can be used to identify and analyze transient events or changes in the frequency content of a signal over time.

2.1.3 Machine Learning

Machine learning is a field of artificial intelligence involving algorithms and statistical models to enable computers to learn and make predictions or decisions based on data. Machine learning algorithms are designed to improve their performance on a specific task through experience without being explicitly programmed to perform the job. There are several types of machine learning, including supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning involves training a model on labelled data, where the correct output is provided for each example in the training set. Unsupervised learning involves training a model on unlabeled data, allowing the model to discover patterns and relationships in the data. Semi-supervised learning involves training a model with labelled and unlabeled data. Reinforcement learning involves training a model to make decisions in an environment to maximize a reward. Machine learning is used in many applications, including image and speech recognition, natural language processing, and fraud detection.

2.1.3.1 Logistic Regression (LR)

Logistic Regression is a statistical technique commonly applied for binary classification. It quantifies the odds of an event occurring based on one or more parameters. LR models the relationship between the independent variables and a binary outcome using the logistic function.

2.1.3.2 Support Vector Machines (SVM)

Support Vector Machines aim to determine the hyperplane segregating datasets into distinct classes. They are practical and versatile in high-dimensional spaces, as they can achieve linear and non-linear classification by leveraging different kernel functions. The central principle is to maximize the margin between the closest data points of two classes, ensuring effective differentiation between classes.

2.1.3.3 Random Forest (RF)

Random Forest is an ensemble learning method that constructs an array of decision trees during its training phase. For classification tasks, it takes the mode of predictions from individual trees, and in the case of regression, it provides the mean prediction. This aggregation across multiple trees reduces the chance of overfitting and provides more accurate and stable predictions.

2.1.4 Deep Learning

Deep learning, an advanced branch of machine learning, harnesses multi-layered neural networks to analyze extensive datasets that enable complex pattern recognition and advanced forecasting capabilities. Its prominence in various applications, from computer vision to language processing, is attributed to its depth and computational power.

2.1.4.2 Convolutional Neural Networks (CNN)

CNNs are a type of deep learning model designed explicitly for handling grid-like data, such as images or time series data, making them particularly suitable for our study as we have converted EEG data into spectral image-like data. Figure 2 illustrates a simple schematic representation of a basic CNN.

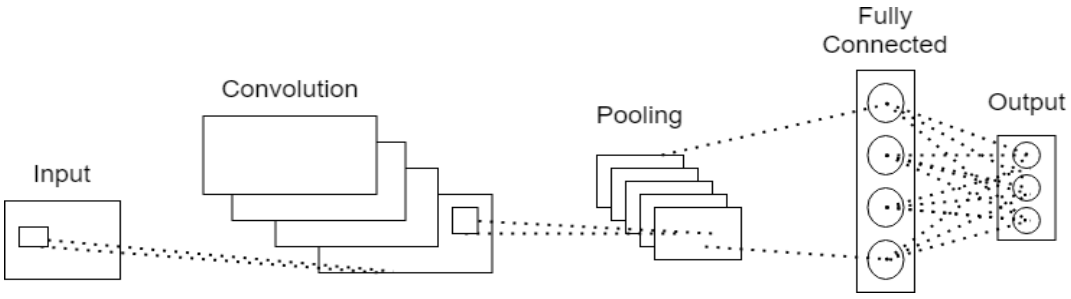

Figure 2 : Schematic Representation of a Basic Convolutional Neural Network (CNN)

This network consists of five primary components: an input layer, a convolutional layer, a pooling layer, a fully connected layer, and an output layer. The network components can be grouped into two principal parts: feature extraction and classification. The feature extraction component, which includes the input, convolutional, and pooling layers, identifies and isolates pertinent features from the input data. In the context of our study, this is particularly beneficial as convolutional layers are highly effective at processing image data recognizing spatial hierarchies and patterns that traditional methods might overlook. On the other hand, the classification component comprises the fully connected output layers, which categorize the data based on the features identified in the extraction stage. The activation layers introduce non-linearity into the model, which enables it to learn complex patterns. Pooling layers, playing a critical role in reducing the spatial dimensions of the data, help control overfitting and minimize computation time. Lastly, fully connected layers consolidate the learned features and produce the final classification output. In this study, we have integrated CNNs into the ConvLSTM [11] and CNN-LSTM models to efficiently extract spatial features from the transformed EEG data.

2.1.4.2 Long Short-Term Memory (LSTM)

LSTM networks, a type of recurrent neural network (RNN) architecture, excel at learning and retaining long input data sequences, making them particularly suited for time series and sequential data. This is mainly because of their unique architecture that can capture temporal dependencies across time, which is crucial for the EEG data in this study. Figure 3 provides a graphical representation of a typical LSTM unit comprising a memory cell and three essential gates: the input, forget, and output.

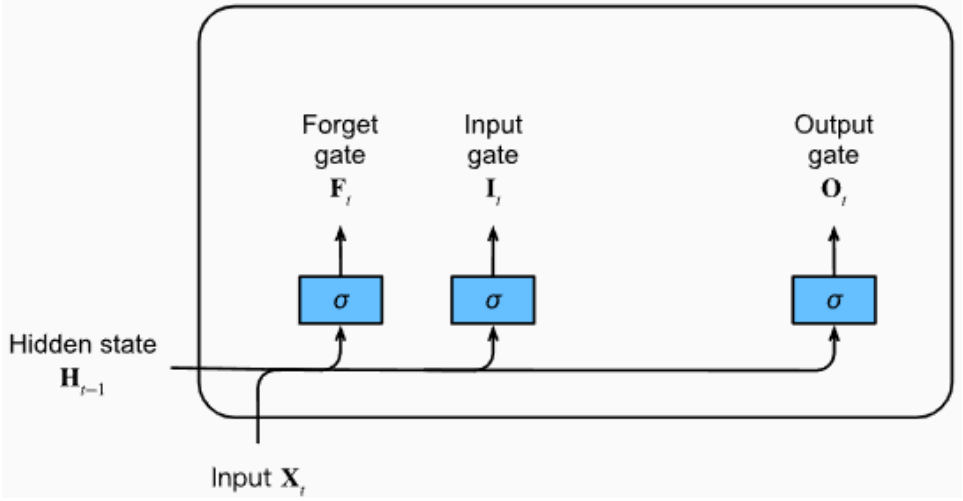

Figure 3: Diagram of a Long Short-Term Memory (LSTM) Network Highlighting the Input, Forget, and Output Gates

Each gate is represented by a fully connected layer with a sigmoid activation function, resulting in output values in the range of 0 to 1. Intuitively, the input gate determines how much the new input influences the current memory cell state. Conversely, the forget gate decides whether to retain or discard the current memory value. Lastly, the output gate controls the influence of the memory cell on the unit's output at the current time step. Alongside the gates, an input node computed with a tanh activation function is also present. This node provides a normalized version (-1 to 1) of the current input for the memory cell. The gating mechanisms combined with the memory cell enable LSTM networks to learn and recall long-term dependencies in the data, which are typically challenging for traditional RNNs. Given their ability to handle temporal patterns, this study incorporates LSTM networks into the ConvLSTM and CNN-LSTM models to efficiently learn temporal patterns from the EEG data.

2.1.4.3 Convolutional Long Short-Term Memory (ConvLSTM)

ConvLSTM, an abbreviation for Convolutional Long Short-Term Memory, represents a unique blend of convolutional neural network (CNN) with the long short-term memory (LSTM) framework. Unlike traditional architectures, it integrates convolutional operations directly in the LSTM cell structure. As a result, each LSTM cell is adept at navigating spatial dimensions and combines it with LSTM's abilities to understand temporal sequences like analyzing video streams. Notably, this model can process data directly, sidestepping the usual initial step of feature extraction.

2.1.4.4 CNN-LSTM:

The CNN-LSTM framework adopts a phased methodology to overcome data interpretation challenges. It begins by leveraging the convolutional layers intrinsic to CNNs to determine the spatial features within the dataset. Once the spatial details are extracted, they are routed to LSTM units, which focus on capturing the sequential dynamics of the data. This layered method provides the CNN-LSTM with a unique advantage, mainly where preliminary spatial analysis is before understanding the temporal sequence, a scenario commonly observed in detailed time series analysis. The methodology section will cover a deeper understanding and its specific application for this study.

2.1.4.5 Bi-LSTM:

Bi-LSTM serves as an evolved variant of the standard LSTM. It is designed to harness information from both past and forthcoming contexts. This dual-directional processing, spanning both forward and reverse paths, addresses challenges where data dependencies span in both temporal directions. The bi-directionality also enhances the model's conceptual understanding, positioning Bi-LSTMs as a preferred choice where in-depth understanding insights across time are particularly evident in tasks like sequence annotations.

2.1.4.6 Gated Recurrent Unit (GRU):

The Gated Recurrent Unit [12] is a streamlined alternative to the classic LSTM model. It was developed in response to the need for a recurrent architecture that mirrors the LSTM's capabilities in temporal sequence modelling; the GRU offers a more concise architecture. Utilizing its update and reset gates, it effectively regulates temporal information. By amalgamating several LSTM gates into a condensed set, the GRU often rivals the performance of its LSTM counterpart. This proves advantageous in events that prioritize computational efficiency, especially in environments constrained by processing resources.

$$\boldsymbol{Update\ gate}: \boldsymbol{z_t} = \boldsymbol{\sigma}(\boldsymbol{w_z} \times \boldsymbol{x_t} + \boldsymbol{U_z} \times \boldsymbol{h_{t-1}} + \boldsymbol{b_z}) \quad \textbf{(1)}$$

$$\boldsymbol{Reset\ gate}: \boldsymbol{r_t} = \boldsymbol{\sigma}(\boldsymbol{w_r} \times \boldsymbol{x_t} + \boldsymbol{U_r} \times \boldsymbol{h_{t-1}} + \boldsymbol{b_r}) \quad \textbf{(2)}$$

Where σ is the sigmoid activation function, W, U, and b are weight matrices and bias vectors.

2.1.5 Muse 2

The IoT device used in this work is the Muse 2 headset. This wearable headset is used for its two main features: it is cost-effective and a non-invasive portable device weighing 200 grams. It contains five dry electrodes of a single metal that plays the role of conductor between the skin and the electrode. Figure 4 displays the placement of the electrodes at TP9 (left ear), TP10 (right ear), AF7 (left forehead) and AF8 (right forehead). The names of electrode placement are derived from the 10-20 system, based on a standardized system used to compare the EEG results among different subjects. A third-party application called Mind Monitor collects and stores data from the IoT device. The data is recorded at a frequency rate of 256 data points per second. Given its emphasis on EEG data gathering, the Muse 2 headset is a pivotal asset in research environments. Its capability to provide real-time feedback on neural activity makes it essential for studies requiring instantaneous data interpretation.

Figure 4: Muse 2 Headset

2.1.6 Long COVID

Long COVID, also known as post-acute COVID-19 syndrome, is a term used to describe the lingering symptoms that some individuals experience after recovering from the initial infection caused by the SARS-CoV-2 virus. One of the reported symptoms of long COVID is brain fog, which refers to a feeling of mental confusion or difficulty with concentration and memory. This can significantly impact an individual's daily life and ability to perform tasks. The cause of brain fog in long COVID is not yet fully understood, but it is thought to be related to inflammation and other body changes due to the initial infection. More research is needed to understand the full range of symptoms and the long-term effects of long COVID. The range and intensity of symptoms linked to long COVID differ, spanning multiple physiological systems. This enduring aspect of the COVID-19 pandemic underscores the pressing need for and importance of global research, aiming to shed light on its origins and potential treatment strategies.

2.1.7 Chronic Fatigue Syndrome

Chronic fatigue syndrome (CFS) is a medical condition characterized by persistent fatigue that is not relieved by rest and cannot be explained by any other underlying medical condition. CFS is also known as Myalgic Encephalomyelitis (ME). Brain fog is a common symptom of CFS/ME and can include difficulties with concentration, memory, and decision-making. It is not fully understood what causes CFS/ME, but it is believed to be the result of a combination of genetic, environmental, and infectious factors. CFS/ME can have a significant impact on an individual's quality of life and ability to perform daily tasks. The treatment often includes a combination of approaches, such as medication, therapy, and lifestyle changes. While the root causes are still underdetermined, a notable number of patients report the emergence of symptoms after experiencing a viral infection. The advocacy groups play a vital role in amplifying awareness of this condition, highlighting its implications for those affected.

2.1.8 Cognitive-Motor Integration Task

Cognitive-motor integration (CMI) tasks are generally used in studies related to neural activity. During the CMI task, the participants perform a job that requires hand and eye coordination movements at the same time. Such tasks are designed to delve into the synchronization of cognitive functions and motor responses. The simultaneous engagement of mental and physical abilities in CMI tasks sheds light on the complexity

of neural connections. Additionally, variations in task executions can highlight the influence of cognitive strain on motor functionalities.

2.2 Related Work

The repercussions of the COVID-19 pandemic on global health extend beyond immediate illness. The impact is evidenced by the persistent symptoms experienced by numerous individuals post-recovery, known as long COVID [13]. Expanding on this issue, Di Toro et al. [2] discuss the diverse clinical presentations and the involvement of multiple organs seen in patients, highlighting symptoms such as fatigue and breathlessness that continue for a considerable time. In reinforcing this understanding, a comprehensive review [14] observes that the impact of the virus in children is often less severe than in adults, suggesting a different impact of the virus across different age groups. In [15], it was reported that a significant fraction of COVID-19 hospitalizations involved patients suffering from neurological symptoms, illustrating an added dimension of complexity in understanding the long-term consequences of the disease. These insights elevate the discussion beyond the direct physiological health effects, acknowledging the intricate challenges faced in the aftermath of the virus. Investigations into the extended consequences of COVID-19 have revealed intricacies comparable to those found in pre-existing medical conditions. For instance, studies have identified a symptomatic relationship between long COVID individuals and those afflicted with conditions such as ME or CFS [3]. Notably, both groups displayed similar clinical features, which included cognitive issues such as problems with hand-eye coordination and memory loss. These symptoms [16] are not exclusive to post-viral conditions and resemble those observed in other cognitive disorders.

Interestingly, the cognitive symptoms associated with long COVID exceed the boundaries of post-viral repercussions, reflecting similarities between various cognitive disorders. These impairments echo the characteristics of Mild Cognitive Impairments (MCI). This transitional stage captures the cognitive decline associated with normal and more similar neurodegenerative stages, such as Alzheimer's disease or other forms of dementia [17]. This insight underscores the necessity for further inquiry into the neurological remnants of long COVID and its intersection with various cognitive impairments.

The global health crisis has significantly accelerated the integration of technology in healthcare infrastructures, with a specific emphasis on the advancements in the IoT [18]. In neuroscience, wearable EEG devices stand out due to their non-invasive nature and proficiency in conducting immediate health assessments, pivotal in cognitive health technology [19].

These devices are increasingly utilized with advanced machine learning techniques, improving their precision in analysis and interpretation. Bashivan et al. [20] have notably demonstrated the transformative role of deep learning in EEG data processing. Their pioneer research emphasizes how CNN can drastically refine EEG data classification, contributing to a more nuanced and practical approach to real-time health monitoring. IoT wearable devices play a crucial role in broadening our understanding of diverse neurological impairments, as seen by their use in research on chronic high-altitude hypoxia [21] and epilepsy [22]. However, a notable void remains in making these tools attainable for individuals suffering from long COVID, restricting their widespread application and hindering early symptom detection and intervention.

In tandem, machine learning has emerged as an effective tool in decoding the complex data procured by EEG studies. Its effectiveness in detecting neurological anomalies has been demonstrated across various conditions, including the automated diagnosis of mild cognitive impairment [23], highlighting its importance and versatility. This innovative approach, merging EEG data with machine learning techniques, represents significant advancements in e-health and IoT systems, especially in settings demanding rapid, accurate, preventive healthcare solutions. However, a significant gap is evident within the context of long COVID. There is an urgent need for a standardized methodology that employs machine learning for analyzing EEG findings. Filling this void would make the diagnostic procedure streamlined and facilitate a more nuanced understanding of long COVID and its neurological implications.

A distinct aspect of our approach involves generating synthetic data to enhance the machine learning models used in our study. Given the challenges in obtaining large volumes of EEG data and respecting patient privacy, we employ Wasserstein Generative Adversarial Networks (WGANs) to generate synthetic spectrograms. These advanced networks facilitate the generation of synthetic spectrograms, replicating the EEG patterns of healthy individuals and those grappling with the persistent complications of COVID-19. This strategy is mainly in the context of pressing demands for confidentiality in healthcare data handling, a concern prominently highlighted in recent studies [24]. Our approach offers a unique perspective on long COVID, building upon the established academic groundwork. The foundation of our methodology was devised upon a study that pioneered the idea of Generative Adversarial Networks (GANs) [25]. The versatility and effectiveness of GANs, evident in various research areas such as the generation of EEG data, have been underscored through subsequent detailed investigations [26], [27]. Drawing on these insights, our investigations harness the power of synthetic data, aiming to craft a reliable and innovative framework that ensures the privacy concerns and continuous evaluation of cognitive impairments associated with long COVID.

To conclude, examining existing literature highlights the emergent need to tackle the complexities of long COVID, especially its neurological consequences, often mirroring symptoms found in diverse cognitive conditions. The integration of IoT in healthcare, mainly through the adoption of wearable devices, has significantly revolutionized patient monitoring and addressing. Despite the critical role of EEG in tracking neurological health, there remains a noticeable void in its application combined with advanced machine learning for managing long COVID symptoms. Moreover, using synthetic data is an essential tool in contemporary research, providing solutions to the challenges associated with confidentiality issues and data insufficiency. Our research intends to bridge these gaps, contributing a unique perspective within the academic landscape and creating a foundation for managing the long-term repercussions of COVID-19.

2.3 Summary

In the background section, foundational concepts like EEG and digital signal processing were explored alongside machine and deep learning methodologies. Further discussions delved into utilizing Muse 2 for data collection and the relevance of conditions like long COVID and CFS. This base knowledge laid the groundwork for the detailed examination of related works.

Chapter 3

3. Architecture and Methodology

3.1 Overview

Our proposed methodology initially gathers data from participants performing a Cognitive-Motor Integration (CMI) task. Following this, we performed data preprocessing and feature engineering to prepare the data to feed the models. Both traditional machine learning and advanced deep learning algorithms were implemented. Once these models were trained, they were subsequently subjected to extensive testing to evaluate, validate, and compare their performances. An additional aspect of our methodology includes the use of synthetic data generation, which was incorporated as a strategic step to potentially improve the model's accuracy and address data scarcity challenges. The general flow and sequence of these steps can be seen in Figure 5. Each of these key steps will be further elaborated in the subsequent subsections.

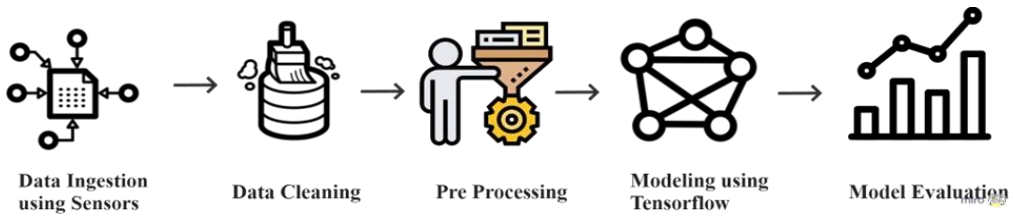

Figure 5: Methodology

3.2 Participants

The behavioural data and electroencephalogram were collected from 23 participants, as shown in Table 1. This included 10 healthy control participants and 13 participants in the disease group. The disease group comprised individuals with long COVID and ME/CFS, both of which have been observed to exhibit the symptom commonly referred to as "brain fog." The healthy participants had no history of brain injury or

neurological illness. All procedures were approved by York University's Human Participants Research Committee, and all participants provided informed consent to participate.

Group	Number of Participants	Age (Mean)	Age (St. Dev)
Control	10	31.6	18.24
ME	7	45.71	10.33
long COVID	6	44.67	18.45
ME/Covid combined	13	45.23	14.57

Table 1: Distribution of Participants by Group, Average Age (Mean, Std Dev), and Health Condition

3.3 Experimental Task

In this study, we employed a between-subjects design. This design involves using two distinct groups of participants to compare the effect of health on performance. The between-subjects design prevents potential carry-over or learning effects that may influence the results when a single participant is exposed to multiple conditions. In our case, the two groups comprised healthy participants and long COVID/ME patients. The 'different conditions' refer to the health status of the participants, not the task itself. Thus, while both groups performed the same computer-based visuomotor skill evaluation tasks using the BrDI (Brain Dynamics Indicator, 3MotionAI Inc.) application, their responses (given their health condition) are expected to be different and thus comparable. Participants performed computer-based visuomotor skill evaluation tasks using the BrDI application. They sat at a desk with a 10.1-inch tablet (Samsung Galaxy Tab) within easy reach for these tasks. The task, as depicted in Figure 6, involved using the index finger of the participant's dominant hand to navigate a cursor (white dot, 5 mm diameter) from the screen to one of the four peripheral targets (up, down, left, or right relative to center).

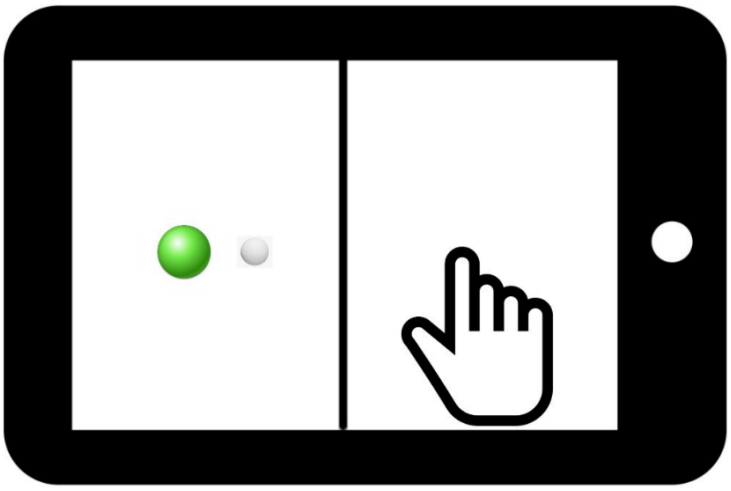

Figure 6: An illustration of the Tablet Screen, Divided into Two Halves by a Vertical Line. The White Dot Represents the Cursor that Participant Controls, and the Green Dot Denotes the Target

Beginning the trial, as detailed in Figure 7, involved guiding the cursor to a solid green circle. A green circle appeared at the periphery after holding the cursor there for 2 seconds. The trial concluded when the cursor remained in the final target for approximately 500 milliseconds. The task incorporated 20 trials, with five trials directed towards each of the four targets.

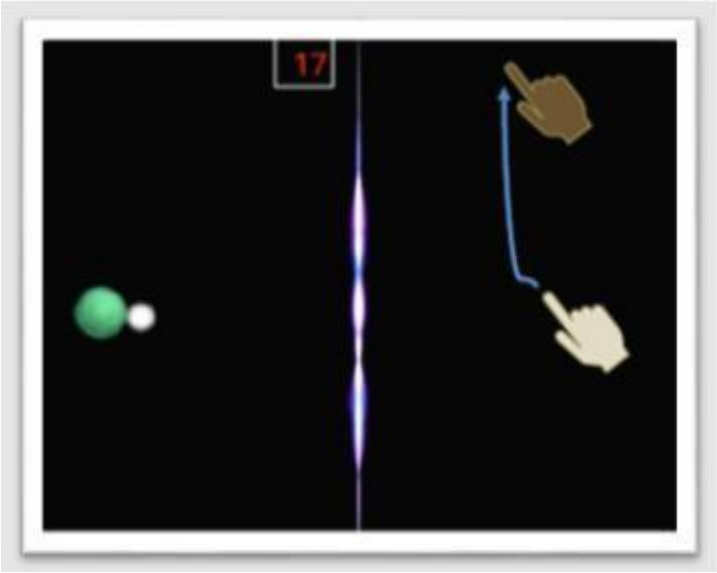

Figure 7 : CMI Task

In the trial, a vertical line split the display, with the targets and cursor shown on the left side of the screen. The right portion of the screen functioned as the active, interactive domain, where the individuals positioned and maneuvered their fingers. This design ensured that the hand's movements did not obstruct the view of the cursor and the targets. The task demanded integrating spatial and cognitive rules, requiring participants to perform a more complex task involving cognitive-motor integration (CMI), which combined spatial and cognitive rules to execute activities that demanded mental and physical coordination. The cursor feedback was inverted by 180 degrees, requiring the participant to slide their finger in one direction to move the cursor in the opposite direction towards the target. Before commencing the trial, participants were instructed to minimize unnecessary eye blinking and jaw clenching, as these actions could introduce noise into the EEG data and potentially affect the study's results.

3.4 Behavioural Data

In our study, we calculated a range of kinematic variables, including reaction time, movement duration, the mean path length from a central starting point to a peripheral end target, peak velocity, and measure of accuracy and precision, denoted as absolute and variable error, respectively. Furthermore, we also computed the number of error trials. A trial was deemed unsuccessful for several reasons: if the participant failed to initiate within 4,000 milliseconds of the start target's appearance, left the home target too quickly (under 150 milliseconds), had an excessively long reaction time exceeding 8000 milliseconds, or took more than 10000 milliseconds to arrive at the final target. A trial was also marked as an error if there was an initial 'directional deviation', where the primary movement exited the central target at an angle of more than 45 degrees from a direct line to the target. The movement onsets, initial stopping point, and if there were corrective movements, final stopping point were scored at 10% peak velocity. When calculating kinematic dependent measures, individual trials that exceeded two standard deviations from the participant's mean were eliminated before calculating that outcome.

Additionally, a previous analysis on the behavioural data demonstrated significant performance differences between participants with either Post-Acute Sequelae of SARS-CoV-2 infection (PASC) or Myalgic Encephalomyelitis (ME) (combined as one group), and healthy control participants [28]. This was determined using a one-way ANOVA with age as a covariate. Specifically, the ME/PASC group displayed significantly longer reaction times (483.7±48.5 ms) and movement times (2810.4±939.1 ms), relative to healthy controls (388.8±59.8 ms and 2012.0±1408.0 ms respectively, $p < 0.05$). They also demonstrated

lower peak velocities when performing the cognitive-motor integration task (301.7±134.8 mm/s vs. 486.5±273.2 mm/s, $p<0.01$), relative to controls. Due to the low number of male participants, we did not have adequate power to further examine the data for sex-related differences on these dependent measures.

Note: The analyses conducted utilized a specific subset of data gathered in the context of a broader PASC/ME study (Badhwar et al. in preparation).

3.5 Data Acquisition and Recording

While participants were performing the task, the Mind Monitor software gathered EEG data from a portable EEG headband called Muse 2™ (InteraXon Inc., Toronto, Canada). This software is a specialized tool designed for comprehensive EEG data capture and visualization. The EEG data from the Muse 2 device was extracted via four electrodes: TP9, TP10, AF7, and AF8, which align with the International 10–20 System for electrode placement, as shown in Figure 8. Data from Muse 2 was collected at a 256 Hz frequency.

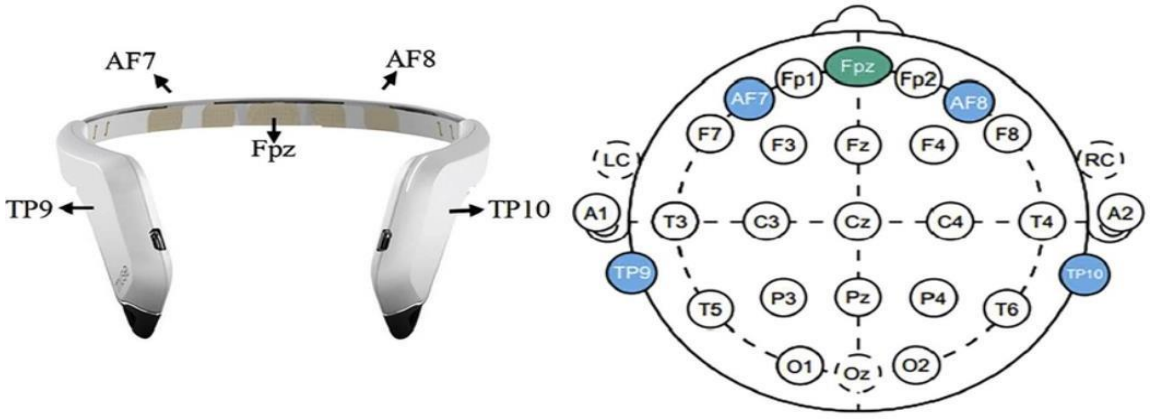

Figure 8: Muse 2 headband and the 10-20 System of Electrode Placement for Muse

3.6 Data Cleaning and Preprocessing

The data cleaning and preprocessing process involved a series of steps to develop a high-quality dataset appropriate for machine learning evaluation. Owing to the high sensitivity of the Muse 2 device, the EEG signals were subjected to noise interference, including disruptions caused by participants eye blinking or jaw clenching. Such instances introduced noise into the dataset and potentially resulted in blank values. Hence, noise elimination was an imperative step. This was accomplished using two Python libraries,

NumPy and Pandas. NumPy is a potent library that enables efficient manipulation of large numerical data arrays. Pandas, built on top of NumPy, provides tools for data cleaning, such as handling missing data. This process was critical in enhancing the dataset's efficiency by eliminating duplicate values and minimizing model bias.

Further refining the data preparation stage, a technique known as data windowing was employed to discern the temporal features of EEG signals more accurately. This involved segmenting the continuous EEG data into smaller, overlapping frames or 'windows'. Each segment was carefully constituted to ensure that it contained sufficient contextual data to be depicted as a representative of a broader signal environment. Undertaking this step was essential in preserving the authenticity and completeness of the patterns captured within the EEG data, explicitly considering the noise and fluctuations inherent in these signals. By dividing the data into these windows, it was possible to create a structured format suitable for time-series analysis, thereby equipping machine learning models with the capability to detect and extrapolate the complex temporal variations present in the data.

After cleaning the dataset, it was partitioned into training and testing sets, following a 70-30 ratio for the train-test split. The data points were standardized using the StandardScaler algorithm from the sklearn library. This algorithm automatically adjusts each feature's mean (μ) and standard deviation (σ) to zero and one, respectively, ensuring uniform scaling across all features. This process is crucial for maintaining data compatibility with machine learning algorithms, as it aligns all features on a common scale without the need for manually setting specific parameters. This is achieved by applying the following formula:

$$x' = \frac{(x-\mu)}{\sigma} \quad \textbf{(3)}$$

In this formula, x' represents the standardized value, x is the original data point, μ is the mean of the feature vector, and σ is the standard deviation of the feature vector. Using this formula, we ensure that every standardized data point reflects how many standard deviations it stands from the mean of the original data. This standardization step is critical because it enables the machine learning model to converge faster and achieve higher accuracy.

Next, the Continuous Wavelet Transform (CWT) was applied to the signals to extract time-frequency information. This transformation enabled the extraction of time and frequency features in the data, providing valuable insights for long COVID effects detection. To further elucidate the process of converting the EEG signals into spectrogram-like using CWT, we provide a pseudo-algorithm in Algorithm 1. This algorithm outlines the critical steps undertaken to achieve this transformation.

Algorithm 1: Algorithm to generate spectograms using EEG data

Result: 64x64 Spectrogram Matrix
Initialize the CWT parameters
Apply CWT on the raw EEG
while *Frequency* ≤ 64 **do**
-Generate a Morlet wavelet by the convolution of sine wave and gaussian.
-Compute the CWT of the Morlet wavelet
-Find the convolution of CWT of wavelet and CWT of signal using the pointwise multiplication.
-Convert the convoluted signal back to time domain using inverse fourier transform
-Find the magnitude of the complex signal and square it to obtain the absolute power component of the signal.
end

The resulting transformed data, as shown in Figure 9 and 10, were used as input for machine learning and deep learning models. Spectrograms offer a visual representation of the frequency content of the EEG signals over time, allowing the models to learn spatial and temporal patterns effectively.

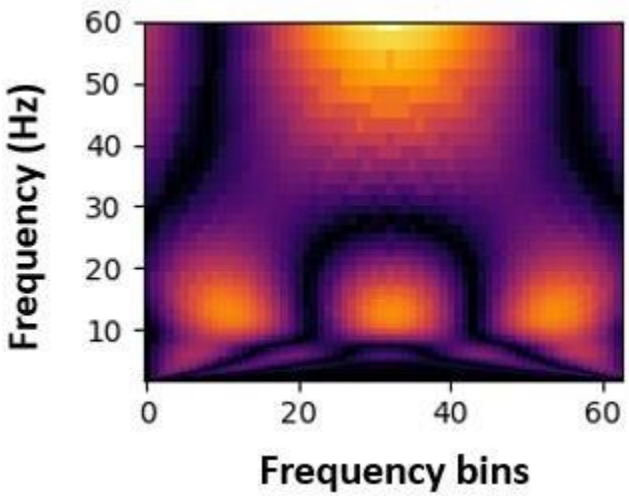

Figure 9 : Sample Spectrogram for Healthy Participants

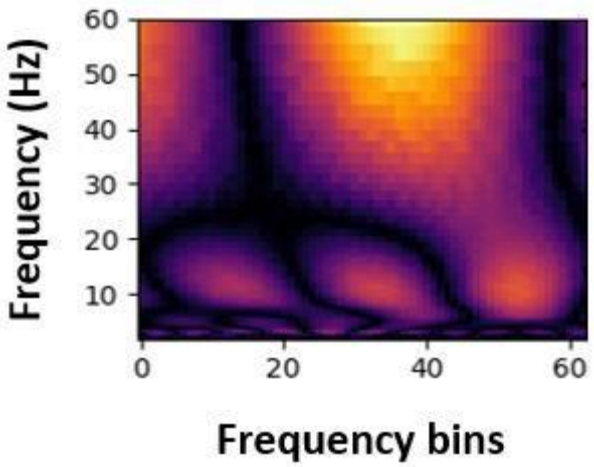

Figure 10: Sample Spectrogram for Long COVID Participants

Finally, the spectrogram-like matrices were converted to the ‘float32’ data type to ensure compatibility with machine learning and deep learning algorithms. The selection of 'float32' was driven by its ability to accommodate real-valued coefficients generated by the Continuous Wavelet Transform (CWT). Moreover, this data type is a requisite for our deep learning model, as it effectively manages fractional values emerging from the model's computations while optimizing memory usage. Converting raw EEG signals into spectrograms and preparing the data for analysis is critical for accurately detecting long COVID effects, enabling early intervention and better management of this illness, ultimately improving the quality of life for affected participants.

3.7 Modeling

In this study, EEG data were utilized to train machine learning (ML) and deep learning (DL) algorithms with the objective to distinguish between healthy participants and those affected by long COVID.

3.7.1 Machine Learning (ML)

Our work uses ML methodologies such as SVM, RF, and LR to categorize participants based on the features extracted from the processed EEG data. These algorithms showcase their flexibility and efficacy in various classification tasks related to long COVID detection.

3.7.1.1 Logistic Regression (LR)

In our study's framework, utilizing Logistic Regression (LR) was a strategic choice, especially for its proficiency in handling binary outcomes, making it suitable for analyzing long COVID symptoms. This model fundamentally uses log odds to represent class likelihoods. To adapt to the intricacies of our EEG data, we tuned the LR model for a higher number of iterations than usual. This enhanced iteration allowed the model to effectively capture and analyze the complex patterns present in the dataset, ensuring a thorough extraction of valuable insights.

3.7.1.2 Support Vector Machines (SVM)

The incorporation of SVM emphasizes its efficacy in determining the best hyperplane for distinct class divisions. The model's inherent quality to maximize the margin between the data points ensures accurate categorization, a trait essential when working with EEG data known for its complex and possible non-linear connections. Using kernel functions, SVM demonstrates flexibility and makes it proficient in navigating through the multi-dimensional spaces of our EEG dataset. This intrinsic flexibility augments its utility in extrapolating insights regarding long COVID symptoms from the data.

3.7.1.3 Random Forest (RF)

Employing Random Forest enhances the resilience of our research framework. As an ensemble method, it is characterized by its multitude of trees, each tree providing a unique vantage point. RF's foundational strategy involves training each tree on separate segments, allowing individualized interpretations. These standpoints are then harmonized by RF, providing a unified decision. The RF model was meticulously tuned to enlist an optimal count of trees in alignment with our research's requirements. This balance ensures an amalgamation of multifaceted insights while judiciously managing computational efficiency. The primary objective remains to achieve a holistic and nuanced understanding of individuals' likely expressions of long COVID symptoms.

3.7.2 Deep Learning

This section offers a detailed overview of the deep learning models utilized in this research: ConvLSTM, CNN-LSTM, GRU and Bi-LSTM. These techniques have been adopted based on their proven effectiveness in tasks related to EEG data analysis.

3.7.2.1 ConvLSTM

The ConvLSTM architecture adeptly captures spatial and temporal dependencies within the EEG data. Initially, the model reshapes the input data to align with its expected format. The structure involves a sequence of ConvLSTM2D layers, which is followed by a batch normalization layer to normalize the activation of the neurons. A distinct feature of this model is the integration of a self-attention mechanism [29]. This component helps to evaluate the significance of every sequence element contextually. After restructuring the data to match subsequent layers, we introduce dropout layers to minimize overfitting. Furthermore, the data undergoes flattening processes, then utilizing dense layers to predict the outcome. This model is then compiled using a binary-cross entropy loss function paired with the Adam [30] optimization technique.

3.7.2.2 CNN-LSTM

The CNN-LSTM model is a fusion of CNN's expertise in feature extraction, and LSTM's are effective in sequence-based learning. In terms of the code structure, the model starts with a sequence initialization. It is followed by a sequence of 2D convolutional layers characterized by specific filters, kernel sizes, and activation functions. In addition, these layers incorporated regularization methods. To down-sample the spatial dimensions, the architecture employed MaxPooling layers. Overfitting, a common problem in machine learning, is addressed using dropout layers. After these initial layers, the model's multi-dimensional output was transformed into a singular dimension form using a flattening mechanism. This restructured output was introduced to the dense layers, leading to class prediction. Once the whole architecture was in place, the model compilation was undertaken. Binary cross-entropy was selected as the loss function along with the Adam optimizer.

3.7.2.3 Bi-LSTM

The Bi-LSTM architecture [31] distinguishes itself through a bidirectional mechanism. This design allows the model to capture and process information from past and upcoming data sequences simultaneously. Examining the technical details, the initial steps involve reshaping the input data multiple times to align with the model's dimensional requirements. Following this, the model was then initialized using the

sequential method. At the core of the design are the bidirectional LSTM layers. Some of these layers are employed to return sequences, resulting in a layered effect, while other layers are structured to return only once. A self-attention mechanism brought further depth to this model by highlighting the relative significance of each sequence component. At the end of the architecture, the dense layers were added to generate class predictions. For the compilation stage, binary cross-entropy was paired with the Adam optimization technique. The final action involved meticulously reviewing the model's efficiency using numerous performance metrics.

3.7.2.4 Gated Recurrent Unit (GRU)

The GRU model's initial stages consisted of convolutional layers that processed the input data. This was followed by batch normalization stages, which standardize the activation function and layers that transform the output. Dropout layers were integrated to avoid overfitting risks. Central to the architecture, the GRU layers employed update and reset gates to manage information flow efficiently. Dense layers were used to generate the final model output. After its detailed assembly, the model was methodically compiled. It underwent training, and in-depth analysis was conducted to evaluate the model's ability.

3.8 Synthetic Data Generation

3.8.1 Overview

Synthetic data generation is an essential technique for augmenting datasets, especially when obtaining additional authentic data is challenging or raises privacy issues. Using Wasserstein Generative Networks (WGANs), we create synthetic spectrogram data to enhance the understanding of EEG data. These synthetic spectrograms are incorporated into the training phase. After the model is trained, it is evaluated on the original, unaltered test dataset. The results, measured through evaluation metrics, highlight the benefits of integrating synthetic data into the training process. This strategy amplifies the volume of training data and addresses privacy issues since synthetic data does not directly tie back to individual participants, ensuring their anonymity.

3.8.2 Methodology

3.8.2.1 Generating Original Spectrograms

While our approach to generating spectrograms mirrors the original methodology outlined previously, one significant distinction exists. In the original methodology, datasets from healthy participants and Long Covid patients were combined and processed collectively. However, our synthetic approach processes and handles the datasets separately, as shown in Figure 11. This ensures we maintain distinct spectrogram and label data for each class, allowing for more focused analysis for both categories. Finally, each class's spectrogram and label data are saved for subsequent analyses.

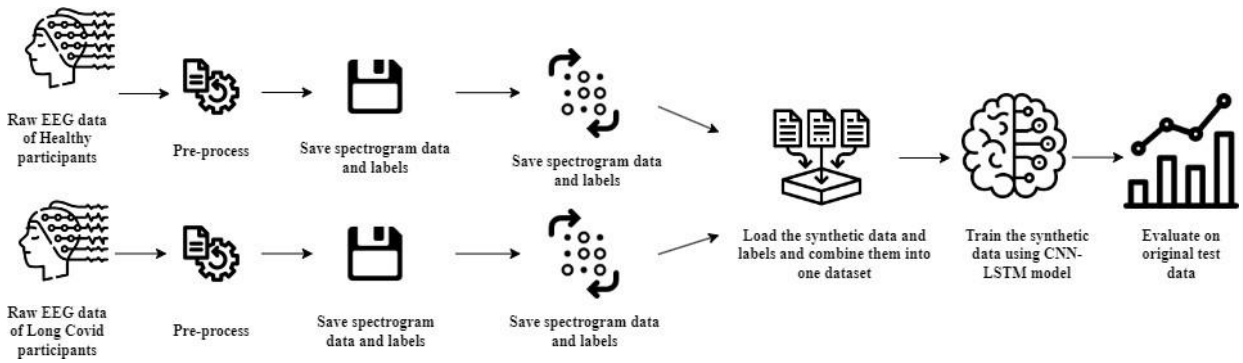

Figure 11 : Methodology for Generating and Evaluating Synthetic Spectrograms

3.8.2.2 Generating Synthetic Spectrograms

The process of generating synthetic spectrograms involves training a Wasserstein Generative Adversarial Network (WGAN) on the original spectrogram data to produce synthetic results. Here is a step-by-step breakdown:

- **Loading Data:** The dataset was initially loaded from the original spectrograms, which were stored in the '.npy' format. This format is a file format used by NumPy, a library in Python for numerical computing.

- **Defining the Neural Network Architectures:** The **Generator** creates synthetic data built upon three dense layers, each supplemented with batch normalization and leaky ReLU activation functions. On the other hand, the **Critic** replaces traditional discriminators in WGANs and is designed with dense layers enriched by dropout and leaky ReLU functionalities.

- **Training Process:** Our approach used the RMSprop optimizer to train the generator and the critic. This training employed a specific loss function derived from WGANs. We initiated the training process by generating synthetic images from random inputs. Following this, we proceeded to train the critic and the generator sequentially. A crucial aspect of our method required maintaining the critic's weights within the range of -0.01 to 0.01. This constraint is essential to ensure the Wasserstein distance remains valid and the training stabilizes, providing valuable gradients for the generator.

- **Generation of Synthetic Data:** Throughout the training cycle, every set of 100 epochs generates the synthesis of new data samples that are then appended to the synthetic dataset.

- **Post-Processing and Saving:** Post-training, the synthetic data undergoes denormalization to revert to its initial scale. Subsequently, the synthetic labels are allocated, and the data and labels are archived for further model training.

3.9 Summary

This study offers an in-depth analysis of EEG data to differentiate between individuals, aiming to distinguish between healthy individuals and those affected by long COVID. Processing techniques, including noise removal, were employed to transform the EEG data into a structured format. The feature extraction using Continuous Wavelet Transform (CWT) further prepared the data for robust evaluation. By leveraging ML algorithms, specifically SVM, RF and LR, the study identified specific patterns within the EEG data. Complementing this, advanced DL architectures like GRU, CNN-LSTM, BiLSTM, and ConvLSTM further highlighted patterns of long COVID. An innovative addition to the research involved generating synthetic spectrograms data using WGANs. This approach not only augmented the available data but also addressed privacy concerns.

Chapter 4

4. Performance Evaluation

4.1 Model Evaluation

In this study, we evaluated the performance of various machine learning (ML) and deep learning (DL) models for detecting long COVID effects using EEG data. The evaluation metrics employed included accuracy, precision, recall, and F1-score, with macro averaging applied for precision, recall, and F1-score.

The formulas for Accuracy, Precision, Recall, and F1-score are as follows:

Let **TP** = True Positives, **FP** = False Positives, **TN** = True Negatives, **FN** = False Negatives

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

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})} \quad \textbf{(5)}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})} \quad \textbf{(6)}$$

$$\boldsymbol{F1 - score} = \frac{\boldsymbol{2 \times (Precision \times Recall)}}{(\boldsymbol{Precision} + \boldsymbol{Recall})} \quad \textbf{(7)}$$

Upon conducting the performance evaluation, it was observed that deep learning models surpassed their machine learning counterparts in terms of accuracy. Detailed performance metrics for each of these models are provided in Table 2 and 3. Table 2 specifically presents the metrics for the ML models. Among these models, the Random Forest model achieved the best performance with an accuracy of 77%, indicating that its ensemble learning approach provided a more reliable and stable classification model for detecting long COVID effects. The Logistic Regression model attained a 63% accuracy, accompanied by average accuracy of 63% for precision, recall, and F1-score. In comparison, the Support Vector Machine model demonstrated a 51% accuracy and achieved accuracy of 55% for precision, 51% for recall, and 41% for F1-score.

Model	Accuracy %	Precision %	Recall %	F1-score %
LR	63	62	64	63
RF	77	72	75	74
SVM	51	55	51	42

Table 2: Performance Metrics Comparison for Machine Learning Models

As shown in Table 3, where the performance metrics for DL models are presented, ConvLSTM and CNN-LSTM models showed notable results, with the CNN-LSTM model demonstrating the highest performance, exhibiting an accuracy of 83%.

Model	Accuracy %	Precision %	Recall %	F1-score %
CNN-LSTM	83	85	82	83
ConvLSTM	77	80	77	77
GRU	63	64	63	62
Bi-LSTM	70	71	70	70

Table 3: Evaluation of Deep Learning Models Based on Key Performance Metrics

The superior performance of the CNN-LSTM model can be attributed to its unique combination and order of CNN and LSTM layers. Unlike other models, the CNN-LSTM architecture starts with CNN layers, which are excellent at extracting spatial features from the data. It then feeds these features into LSTM layers, which excel at capturing temporal dependencies in sequences. This model capitalizes on the

strengths of both components: CNN layers efficiently process spatial information and reduce the complexity of the input. Once simplified, this input is passed to the LSTM layers, which can then focus on extracting the time-based patterns without being overwhelmed by high-dimensional raw data. Therefore, the CNN-LSTM model can analyze both spatial and temporal aspects of the data more efficiently while considering the contextual importance of each input sequence element. These findings suggest that deep learning models, particularly the CNN-LSTM model, show promise as effective tools for early detection and management of long COVID effects using EEG data, which is crucial for the neuroscience field. Timely intervention could improve patient outcomes and contribute to a better understanding of the long-term effects of COVID-19 on the brain.

4.2 Evaluation of Synthetic Spectrograms

Following the application of our synthetic data generation methodology, the CNN-LSTM model was subjected to extensive classification metric assessments. The results, tabulated over five separate runs, consistently emphasized the model's proficiency in differentiating EEG patterns between healthy participants and those with long COVID. Averaging the metrics over these runs, the model attained a precision, recall, and F1-score of approximately 93% for both classes. Notably, the average accuracy over all runs was 93%, which displays the model's robustness and the efficacy of using synthetic data in the training process. Detailed results of these assessments are presented in Table 4.

Run	Class	Metrics (%)			Accuracy (%)
		Precision	Recall	F1-score	
1	0	0.90	0.97	0.94	0.94
	1	0.97	0.90	0.94	
2	0	0.87	0.97	0.92	0.92
	1	0.97	0.87	0.92	
3	0	0.92	0.98	0.95	0.95
	1	0.98	0.92	0.95	
4	0	0.92	0.97	0.94	0.94
	1	0.97	0.92	0.94	
5	0	0.85	0.97	0.91	0.90
	1	0.97	0.84	0.90	
				Avg Accuracy	**0.93**

Table 4: Classification Report for Synthetic Spectrograms Over 5 Runs: Distinguishing Between Healthy Participants (Class 0) and Long COVID Participants (Class 1).

4.3 Comparison of Original vs Synthetic Spectrograms

In this segment, a comparison was made between the performance metrics of the CNN-LSTM model trained on original and synthetic data. While the original data-trained CNN-LSTM achieved an accuracy of 83%, the synthetic data-trained model registered a commendable average accuracy of 93% over multiple runs. Additionally, we compared the mean and standard deviation differences between the original and synthetic datasets, as shown in Table 5. It is worth noting that the values in both datasets were normalized, accounting for the low magnitudes observed in the mean and standard deviation. The standard deviation of the original dataset is approximately 42.74, which closely aligns with the 44.54 observed in the synthetic dataset, with a minor difference of 1.70, reflecting a percentage difference of approximately 4%. On the other hand, the original dataset features a mean of approximately 810.74, similar to the synthetic data's mean of 851.28 with a minor difference of 40.53, which translates to a percentage difference of approximately 5%. These results suggest that the synthetic data generation methodology is reliable and offers an enhanced performance potential when training the model. Moreover, it indicates the potential of synthetic data to

augment existing datasets and significantly enhance the predictive capabilities of deep learning models, especially in areas where the collection of extensive real-world data might be challenging or time-consuming.

Metric	Original Dataset	Synthetic Dataset	Difference
Mean(μV)	810.74	851.28	40.54
Std Dev	42.74	44.54	1.80

Table 5: Summary Statistics of Original and Synthetic Spectrogram Datasets.

4.4 Summary

This research meticulously evaluated the cognitive-motor deficits in individuals afflicted with long COVID. It was established that these individuals notably demonstrated the increased reaction time and movements durations, and diminished peak velocities compared to healthy individuals. The study further proceeded with an analytical examination between the machine and deep learning models in identifying long COVID effects using EEG data. It was observed that deep learning models, particularly the CNN-LSTM model, outperformed the traditional machine learning models, showcasing enhanced ability in processing and integrating EEG data's spatial and temporal features.

An Innovative stride was made by incorporating the synthetic data in the training phase of the CNN-LSTM model, yielding a notable accuracy of 93%. This improvement in accuracy, when compared with the model using original datasets, emphasizes the significant augmentation provided by the synthetic data in the model's training process. Furthermore, the congruence in the statistical parameters between the original and synthetic datasets solidifies the credibility of synthetic data approaches. This development is pivotal, indicating that the usage of synthetic data has the potential to amplify the analytical depth and predictive accuracy of deep learning models, particularly in complex conditions like understanding the neurological implications of long COVID.

Chapter 5

5. Conclusion and Future Work

5.1 Conclusion

Our study demonstrates the potential of using EEG data combined with advanced computational models in discerning healthy individuals and those affected by long COVID. Notably, the CNN-LSTM model exhibited superior performance, emphasizing its potential for early detection and intervention. A significant advancement in our research is the introduction of synthetic spectrograms generated through Wasserstein Generative Adversarial Networks (WGANs). These synthetic spectrograms were used in a separate training phase and notably achieved an average accuracy of 93%, enhancing performance metrics. This success underlines the utility of synthetic data in training models where real-world data is limited or raises privacy concerns.

These insights have significant implications for the neuroscience field, as early detection and intervention could lead to improved patient outcomes and a better understanding of the long-term effects of COVID-19 on the brain. The study also highlights the importance of data preprocessing and feature engineering in developing high-quality datasets for training and testing machine learning and deep learning models. The utilization of Continuous Wavelet Transform and spectrogram-inspired matrices facilitated the efficient extraction of time and frequency characteristics from the EEG data, yielding crucial insights for identifying the effects of long COVID.

5.2 Summary of Contributions

This study introduces several pivotal contributions to understanding EEG data, the effects of long COVID, and the application of machine and deep learning models, focusing mainly on the data obtained from IoT wearables. These innovative measures contribute to the more advanced, nuanced research and practical, real-world application in addressing the complexities associated with the effects of long COVID. The principal contributions include:

Contribution 1: Formulation of a comprehensive methodology for processing and analyzing EEG data from an IoT wearable device called Muse 2. This approach is specifically tailored to studies concerning the effects of long COVID. It includes a detailed transformation of raw EEG signals into spectrogram-like matrices, enhancing their suitability for analysis through machine learning techniques.

Contribution 2: An extensive comparison and analysis of traditional machine learning methods against deep learning techniques for differentiating between healthy individuals and those affected by long COVID. While traditional machine learning models exhibited competency, with specific models achieving accuracies of 77%, they were notably surpassed by the more advanced deep learning techniques. The CNN-LSTM model emerged as particularly effective, achieving an accuracy of 83%. The model's ability to capture and decode EEG data's spatial and temporal features suggests it could be crucial in identifying the neurological effects of long COVID.

Contribution 3: Utilization of synthetic data in EEG analysis related to effects of long COVID, particularly in tackling the pressing concerns of privacy and scarcity of data. The method of synthetic data generation was not only successful but also led to positive outcomes. The CNN-LSTM model trained on synthetic data exhibited an average accuracy of 93%. These results display the potential of synthetic data to enhance the model's abilities to evaluate the data.

5.3 Future Work

Future research could focus on increasing the sample size and exploring additional features to enhance the models' performance further. Moreover, investigating the generalizability of the findings to other neurological conditions may expand the applications of these methodologies beyond long COVID detection. Ultimately, this study provides a valuable foundation for developing non-invasive, efficient, and accurate diagnostic tools to detect and manage the long-term effects of COVID-19 on the brain, improving patient care and contributing to a deeper understanding of this complex condition.

Expanding on this foundation, an essential direction for future work includes the development of a real-time platform for autonomous classification, potentially utilizing cloud computing's extensive capabilities

for prompt and large-scale data analysis. This advancement would require sophisticated solutions for data privacy and integral considerations due to the confidential nature of health-related information. Strategies like Federated Learning and Differential Privacy could be pivotal, allowing the framework to learn from encrypted data. This approach eliminates the need for a centralized repository, thus preserving individual privacy. Through these advanced methods, future investigations can aspire to refine the approach in long COVID diagnosis and management, ensuring security and regulatory compliance in the handling of patient data.

Bibliography

[1] W. H. Organization, "Coronavirus disease (COVID-19): weekly epidemiological update," 2020.

[2] A. Di Toro, A. Bozzani, G. Tavazzi, M. Urtis, L. Giuliani, R. Pizzoccheri, F. Aliberti, V. Fergnani and E. Arbustini, "Long covid: long-term effects?," *European Heart Journal Supplements,* vol. 23, 2021.

[3] A. L. Komaroff and L. Bateman, "Will covid-19 lead to myalgic encephalomyelitis/chronic fatigue syndrome?," *Frontiers in Medicine,* 2021.

[4] S. Junaid, "Recent Advancements in Emerging Technologies for Healthcare Management Systems: A Survey," *Healthcare (Basel, Switzerland),* vol. 10, no. 10, 2022.

[5] H. T. Yew, M. F. Ng, S. Z. Ping, S. K. Chung, A. Chekima and J. A. Dargham, "IoT Based Real-Time Remote Patient Monitoring System," 2020.

[6] S. Debener, F. Minow, R. Emkes, K. Gandras and M. d. Vos, "How about taking a low-cost, small, and wireless EEG for a walk?," *Psychophysiology,* vol. 49, no. 11, 2012.

[7] M. Chaudhary, M. S. Adams, S. Mukhopadhyay, M. Litoiu and L. E. Sergio, "Sabotage Detection Using DL Models on EEG Data From a Cognitive-Motor Integration Task," *Frontiers in Human Neuroscience,* vol. 15, 2021.

[8] M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875,* 2017.

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, 1997.

[11] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-C. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," *Proceedings of the 28th International Conference on Neural Information Processing Systems,* p. 802–810, 2015.

[12] K. Cho, B. v. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[13] A. Raveendran, R. Jayadevan and S. Sashidharan, "Long covid: An overview," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 15, p. 869–875, 2021.

[14] J. F. Ludvigsson, "Systematic review of COVID-19 in children shows milder cases and a better prognosis than adults," *Acta Paediatrica,* 2020.

[15] L. Mao, H. Jin and M. Wang, "Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China," *JAMA Neurology,* 2020.

[16] M. H.-B. Lam, Y.-K. Wing, M. W.-M. Yu, C.-M. Leung, R. C. W. Ma, A. P. S. Kong, W. Y. So, S. Y.-Y. Fong and S.-P. Lam, "Mental morbidities and chronic fatigue in severe acute respiratory syndrome survivors: long-term follow-up," *Archives of Internal Medicine,* vol. 169, 2009.

[17] E. Olivera, A. Sáez, L. Carniglia, C. Caruso, M. Lasaga and D. Durand, "Alzheimer's disease risk after COVID-19: a view from the perspective of the infectious hypothesis of neurodegeneration," *Neural Regen Res,* vol. 18, no. 7, 2023.

[18] I. A. Pap, S. Oniga and A. Alexan, "Machine learning EEG data analysis for eHealth IoT system," 2020.

[19] R. Xiong, F. Kong, X. Yang, G. Liu and W. Wen, "Pattern recognition of cognitive load using EEG and ECG signals," *Sensors,* 2020.

[20] P. Bashivan, I. Rish, M. Yeasin and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," *International Conference on Learning Representations,* 2016.

[21] H. Liu, R. Shi, R. Liao, Y. Liu, J. Che, Z. Bai, N. Cheng and H. Ma, "Machine learning based on event-related EEG of sustained attention differentiates adults with chronic high-altitude exposure from healthy controls," *Brain Sciences,* 2022.

[22] F. Hassan, S. F. Hussain and S. M. Qaisar, "Epileptic seizure detection using a hybrid 1D CNN-machine learning approach from EEG data," *Journal of Healthcare Engineering,* 2022.

[23] R. A. Movahed and M. Rezaeian, "Automatic diagnosis of mild cognitive impairment based on spectral, functional connectivity, and nonlinear EEG-based features," *Computational and Mathematical Methods in Medicine,* 2022.

[24] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification," in *Proceedings of the IEEE Symposium on Biomedical Imaging*, 2018.

[25] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," in *Proceedings of the Neural Information Processing Systems Conference*, 2014.

[26] F. Fahimi, Z. Zhang, W. B. Goh, K. K. Ang and C. Guan, "Towards EEG Generation Using GANs for BCI Applications," in *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019.

[27] M. N. Fekri, A. M. Ghosh and K. Grolinger, "Generating Energy Data for Machine Learning with Recurrent Generative Adversarial Networks," *Energies,* vol. 13, no. 1, 2020.

[28] H. Ahuja, S. Badhwar, M. Litoiu, H. Edgell and L. Sergio, "Cognitive-motor performance and associated brain activity shows differences in individuals with Post Acute Sequelae of SARS-CoV-2 (PASC) or Myalgic Encephalomyelitis (ME)," in *Society for Neuroscience (SFN)*, San Diego, 2022.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv , 2014.

[31] M. Schuster and K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Transactions on Signal Processing,* vol. 45, no. 11, p. 2673–2681, 1997.