CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Hypoxic-ischemic encephalopathy grading using novel EEG signal processing and machine learning techniques
Authors	Twomey, Leah
Publication date	2025
Original Citation	Twomey, L. 2025. Hypoxic-ischemic encephalopathy grading using novel EEG signal processing and machine learning techniques. MRes Thesis, University College Cork.
Type of publication	Masters thesis (Research)
Rights	© 2025 , Leah Twomey. - https://creativecommons.org/licenses/by-nc/4.0/
Download date	2026-11-05 14:38:59
Item downloaded from	https://hdl.handle.net/10468/17533

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



Hypoxic-Ischemic Encephalopathy Grading using Novel EEG Signal Processing and Machine Learning Techniques

Leah Twomey

A thesis submitted for the degree of

Master of Engineering in Electrical and Electronic Engineering

School of Engineering

University of College Cork

Head of School/Dept: Prof. Jorge Oliveira/Prof. Peter Parbrook

Supervisors: Dr. Emanuel Popovici, Prof. Andriy Temko

February 26, 2025

CONTENTS

I Declaration 9

II Acknowledgements 10

III Abstract 11

1 Introduction 12

1.1 Structure of the Thesis 14

2 Background Information and Preliminary Feature Analysis 16

2.1 Current State of the Art of HIE Grading 19

2.2 Current State-of-the-Art for 2D Neural Networks in Biomedical Fields 21

2.3 Current State of the Art for EEG Sonification 22

2.4 Dataset 23

3 Spectrogram-Based HIE Grading Using 2D CNN 35

3.1 Data Preprocessing 35

3.2 Machine Learning Model Performance Metrics 38

3.3 Model Training and Hyper-Parameter Tuning 40

3.3.1 **Model Architecture** 40

3.3.2 **Data Augmentation** 42

3.4 Results and Discussion 44

3.5 Conclusion 45

4 From EEG to Audio: A Novel Spectrogram-Based Approach for HIE Grading 47

4.1 Proposed Method 48

4.1.1 Pre-Processing using **Sonification** 48

4.1.2 Deep Learning using **2D CNN** 59

4.1.3 Post-Processing using **Optimised Rounder** 61

4.2 Statistical Analysis Methodology of Sonified EEG Signal 62

4.2.1 Visualisation Analysis Method 70

4.3 Training Methodology: Nested Cross-Validation Evaluation Framework . . . 71
4.4 Results . . . 73
4.5 Discussion . . . 79
4.5.1 Visual Analysis . . . 79
4.5.2 Hyper-Parameter Tuning Comparative Analysis . . . 86
4.5.3 Analysis of EEG Input Window Size . . . 92
4.5.4 Discussion on Reduction in Performance for Nested Cross-Validation Outer Fold 4 . . . 93
4.5.5 Comparison with 2D CNN Baseline . . . 96
4.6 Conclusion . . . 98

5 Model Quantisation and Implementation on Edge Device 100
5.1 Details of the Target Device . . . 100
5.2 Model Deployment . . . 101
5.3 Model Quantisation . . . 102
5.4 Analysis of Model with AMP . . . 105
5.5 Timing: Comparative Results . . . 106
5.6 Memory Usage Analysis . . . 109
5.7 Conclusion . . . 110

6 Conclusion and Future Work 112

References 116

LIST OF FIGURES

1 Current Clinical Diagnosis Workflow for HIE Grading 18
2 A 36 second window of EEG signal for all four HIE grades, the image was adapted to highlight the behaviours of the EEG signal dependant on the severity of HIE present. 24
3 Average Power in Each Sub-band of the PSD. 26
4 Feature Importance Using Scikit-Learn Random Forest Classifier of Top 20 Features from Feature set of 98 features of the EEG signal. 28
5 Box Plot for the Shannon Entropy Value for all Four HIE Grades 29
6 Cross-Validation Test Accuracy for Sweeping Range of Most Important Features Selected and Utilised for Random Forest Classifier. 30
7 Box Plot for the Top 9 Features for all Four HIE Grades. 31
8 Mel-Spectrogram of EEG Signal for the full 1 hour segment of EEG, with a sample for each HIE grade 38
9 Baseline 2D CNN Model Architecture using Mel Spectrogram representation of EEG signal. 41
10 Accuracy Results for Each ConvNext Model Architecture to Evaluate the Effect of the Model Size on the Accuracy Performance. 42
11 Mel Spectrogram representation of EEG signal with the data augmentation applied. The two black columns seen on the figure are the masked areas in the vertical direction. 43
12 Proposed Method Workflow 48
13 Comparison of EEG signal with spectrogram of audio signal after FM/AM transformation. 49
14 FM/AM Sonification Algorithm Overview 50
15 A 30 second window of an audio domain signal for all four HIE grades, highlighting the rhythmic behaviour of each grade after the sonification conversion process. 52
16 Comparison of the Original Audio Signal for a sample EEG with the Downsampled Audio signal in subplot 2. The original signal and downsampled signal are overlayed for direct comparison in subplot 3 of this figure. . . . 53

17 Spectrogram representation of all four HIE grades based on the optimal spectrogram parameters obtained experimentaly for the audio domain representation of the EEG signal. 57
18 Vertically Stacked Channel Spectrograms for the Sonified EEG Signal . . 59
19 Plot of the frequency mapping of all four HIE grades. 63
20 Scatter plot of the standard deviation of the frequencies in comparison to their true HIE grade. 64
21 MFCC Hyper-Parameter Tuning, with the number of MFCCs used to calculate the Random Forest Classifier feature set increasing in one step increments from 1 to 200. 67
22 PCA Analysis of the MFCC Features of the Sonified EEG Signal. 68
23 Regression Outputs of the MLP Model for MFCC Features. 69
24 Nested Cross Validation Workflow 72
25 Best Model Learning Curves 74
26 Normalised Confusion Matrix Representing Test Results of the Nested Cross-Validation Training Scheme 75
27 Nested Cross Validation Training and Validation Loss Learning Curves . . 76
28 Nested Cross Validation Training and Validation Accuracy Learning Curves 77
29 Nested Cross-Validation Outer Fold Test Accuracy for all Five Outer Folds 78
30 UMAP Analysis of Raw EEG Data, visualising the relationships and structures between the data in a 2D plane. 80
31 UMAP Analysis of Sonified EEG Data. 81
32 UMAP Analysis of the Last Feature Map output of the CNN model for the proposed method with FM/AM transformed EEG spectrogram 82
33 UMAP Analysis of the Last Feature Map output of the CNN model for the baseline 2D CNN model based on the Mel Spectrorgam representation of the EEG signal. 83
34 Validation Accuracy and Loss for various spectrogram parameters under analysis 88
35 Plot of a sample spectrogram for a chosen EEG signal with the data augmentation listed above applied. 91
36 Accuracy Per Grade per Fold for each Outer Fold 95

37 Model Implementation and Deployment Workflow 101

38 Violin Plot of the Regression Values for the Quantized Model 104

LIST OF TABLES

I Details of Dataset Distribution Across Grades 25
II Top Importance Sub-bands for HIE Grading 27
III Model Performance Metrics 44
IV Comparison of Original and Downsampled Signals 53
V Nested Cross-Validation Performance Results 74
VI Clustering Metrics for UMAP Analysis of Different Data Representations 84
VII Regression vs Classification Model Performance 89
VIII Test Accuracy for Each Augmentation and Probability 92
IX Nested Cross-Validation Performance for Different Input Window Size . . 93
X Comparison of Validation and Test Predictions for Outlier Samples 95
XI Comparison of Model Performance Metrics for Proposed Method and 2D Baseline CNN Model 97
XII Quantized Model Performance 104
XIII Training and Evaluation Metrics for Comparison of AMP Implementation 105
XIV Inference Time for Non-Quantized and Quantized Model on Qualcomm RB3 107
XV SNPE Execute Statistics (Averaged) 108
XVI Throughput Time for Non-Quantized and Quantized Model on Qualcomm RB3 108
XVII Memory Usage on Qualcomm RB3 Gen 2 Vision Development Kit 109

LIST OF ACRONYMS

HIE Hypoxic-Ischemic Encephalopathy

EEG Electroencephalography

MRI Magnetic Resonance Imaging

aEEG Amplitude-Integrated Electroencephalogram

NIRS Near-Infrared Spectroscopy

AI Artificial Intelligence

CNN Convolutional Neural Network

IBI Inter-Burst Interval

SWC Sleep-Wake Cycle

PSD Power Spectral Density

SVM Support Vector Machine

GMM Gaussian Mixture Model

AR Autoregressive

RMS Root-Mean Squared

Hz Hertz

MFCC Mel Frequency Cepstral Coefficients

LOO Leave-One-Out

RB3 Qualcomm RB3 (Snapdragon) Device

STFT Short-Term Fourier Transform

MSE Mean-Squared Error

TP True Positive

FP False Positive

FN False Negative

FIR Finite Impulse Response

FM/AM Frequency Modulation / Amplitude Modulation

FT Fourier Transform

AMP Automatic Mixed Precision

MP Mixed Precision

FP16 16-bit Floating Point

FP32 32-bit Floating Point

CV Cross-Validation

OOF Out-Of-Fold

SD Standard Deviation

DFT Discrete-Fourier Transformation

DCT Discrete Cosine Transformation

PCA Principal Component Analysis

UMAP Uniform Manifold Approximation and Projection

DBI Davies-Bouldin Index

CHI Calinski-Harabasz Index

ONNX Open Natural Network Exchange

I. DECLARATION

I, Leah Twomey, hereby declare that the work presented in this thesis, titled “HIE Grading using Novel EEG Signal Processing and Machine Learning Techniques” , is entirely my own except where explicityly stated otherwise. I have fully acknowledged and referenced all sources of information used in this document in accordance with academic and ethical guidelines.

I confirm that this work has not been submitted, either in part or whole, for any other academic degree or qualification at this or any other institition.

Leah Twomey

February 26, 2025

II. Acknowledgements

This thesis and my research throughout my masters is an attribute to the support and guidance I have received from many individuals and institutions. For this, I am extremely grateful.

I would like to start by expressing my gratitude to my supervisors, Dr. Emanuel Popovici and Prof. Andriy Temko. They have provided invaluable advice and guidance to me throughout the course of my research. Thanks to their expertise and insightful feedback I have been able to successfully complete my research, and I am sincerely thankful for their mentorship.

I would also like to extend my thanks to my colleagues and friends at the School of Engineering at University College Cork for their support, assistance and most importantly friendship throughout.

I would like to acknowledge Qualcomm for their generous funding and for providing the technological platform necessary to explore the intersection of deep learning and medical signal processing. Their support allowed me to push the boundaries of my research and gain practical insights into real-world applications.

Finally, I want to thank my family and friends, whose support and encouragement have been a constant source of motivation. Their understanding and belief in me were invaluable in helping me persevere through challenges and milestones alike.

To all those who have contributed to this journey, whether directly or indirectly, I would like to say thank you.

III. Abstract

HYPOXIC-ISCHEMIC ENCEPHALOPATHY (HIE) is a critical brain injury in newborns resulting from oxygen or blood supply deprivation during birth. Traditional diagnosis methods for HIE require specialized expertise from a trained medical professional, and immediate intervention of treatment is required within the first 6 hours post primary hypoxic-ischemic insult. Electroencephalogram (EEG) is the gold standard for analysing brain pathologies for neonates and large windows of this complex signal must be assessed for accurate HIE diagnosis. Using an openly available dataset, this thesis proposes an automated approach for HIE grading, leveraging signal processing techniques and Artificial Intelligence (AI) to provide accurate and timely assessments, with an implementation enabling edge-device deployment for real-world clinical utility.

Representing the 1 hour epoch of the EEG signal in the amplitude and frequency domain through Mel Spectorgram representation, the HIE grading task becomes an image recognition problem, where Convolutional Neural Networks have shown high accuracy and efficiency. An initial test accuracy of 84.85% is achieved. Further enhancement of the signals rhythmic behaviour is necessary to increase the grading potential of the signal, thus the FM/AM sonification algorithm was implemented, transforming the EEG signal into an amplitude and frequency modulated audio signal. This pre-processing is designed to enhance the background rhythmic pattern of the signal, an essential feature used by the clinician for visual inspection. The two dimensional (2D) CNN is designed as a regressor to map the input image to a value on the HIE grading scale to enhance the model's ability to leverage the monotonic relationship between the grades. An optimised rounding function is implemented to define the final clinical grade as a novel post-processing technique. An overall test accuracy of 89.97% based on a rigourous nested cross-validation evaluation framework is achieved, surpassing the current state-of-the-art by 6%. The robustness and generalisability of the proposed method is obvious from the weighted F1-score of 0.8985 and a Kappa score of 0.8219.

While enhancing the accuracy and speed of HIE diagnosis with this AI-driven approach, the goal is to make it readily available at the point of care. An inference time of 62.7 milliseconds is achieved for the quantized model on the Snapdragon processor highlighting its suitability for real-time, on-device HIE grading.

1. INTRODUCTION

HYPOXIC-ISCHEMIC ENCEPHALOPATHY (HIE) occurs in neonatal patients due to a critical shortage of blood or oxygen supply to the neonate's brain before, during or after birth [1]. HIE describes the complex physiological, cellular and molecular changes which occur to the neonate's brain following this anoxic injury [1]. This results in a lack of brain activity in the patient, leading to severe clinical outcomes [1]. Depending on the severity of HIE, the long-term outcomes of the patient will show varying levels of neurological impairment, with the most severe cases resulting in death [2]. 15 - 20 % of patients with severe HIE die within the first few days of life, and 25-30 % of neonates will suffer long-term neurological impairments [3] [2] [4] which include, but are not limited to, cerebral palsy, epilepsy, hearing loss, visual defects and cognitive impairments [5] [6].There is an incidence rate of 1-8 per 1000 live births in developed countries, however this figure increases to 26 per 1000 live births in developing countries [7] [8] [1]. In 2010, HIE was identified as the leading cause of long-term disabilities in children [9].

While significant research has been conducted into treatment methods for patients with HIE over recent decades, mild therapeutic hypothermia is the most widely adopted treatment [10]. Phase-specific treatment has been shown to increase effectiveness of treatment, with optimal results achieved when mild therapeutic hypothermia is implemented in Phase I of the HIE pathological and biochemical cascade [10]. Phase I is identified as the primary-energy failure, where energy provided to the cells decreases and a cascade of cellular damage occurs due to the shortage of oxygen or blood supply to the cells [11]. The impact of therapeutic hypothermia is evident from the decrease in mortality rates, where a study shows the mortality rate decreased from 27% to 12.8% after the initiation of therapeutic hypothermia [12]. However, when administered outside the window of less than six hours after birth, and thus outside the Phase I window, the effectiveness of the treatment decreases substantially [13]. The demand for early and accurate diagnosis methods of HIE severity is continually increasing worldwide, due to this demand for early intervention.

With this demand for early diagnosis, accurate HIE grading is proving to be a major challenge for medical professionals. This is in part due to the time constraint associated with early intervention in order for optimal patient outcome and also due

to the complexity of diagnosing HIE. A poor success rate at diagnosing HIE has been identified, highlighting this challenge, with approximately 20% of newborns with poor clinical outcomes not being diagnosed with HIE prior to the onset of severe impairments [11]. This is despite a number of high quality diagnosis methods used for HIE grading including data acquisition modalities such as Magnetic-resonance imaging (MRI), Electroencephalogram (EEG), amplitude Electroencephalogram (aEEG) and Near-infrared spectroscopy (NIRS) [1]. The following research uses EEG signal as the main data acquisition modality as it has shown superior sensitivity and specificity than other modalities [14], along with ease of monitoring bedside.

The struggle faced by medical professionals to diagnose HIE and the resulting poor outcome for patients is the primary motivation behind this study. The goal is to create a novel grading algorithm which can be used as an accurate and accessible tool to aid the physician in diagnosing HIE severity. This motivation lead the research detailed throughout this thesis, where the ultimate objective is to develop a novel approach to HIE grading. The current state of the art uses Machine Learning approaches for accurate HIE grading. Machine Learning methods are neccesssary due to the complex nature of the task, where simple statistical methods have shown inferior results. In order to achieve this, the following four research objectives were identified:

1) Analyse the raw EEG signal to identify relevant features and parameters associated with each HIE grade. Gain understanding of key metrics and attributes associated with each HIE grade.
2) Develop a novel, autonomous method for grading EEG signals that leverages machine learning. Explore suitable data representations and network architectures to enhance the accuracy and reliability of automated HIE grading.
3) Further development of the novel workflow to enhance the proposed pre-processing by using novel signal processing techniques. Highlighting the relevant features identified in the background analysis of the EEG signal to enhance the classification potential and thus increasing the overall accuracy performance.
4) Evaluate the potential for the proposed novel method to run on an edge device, assessing its potential as a decision support tool for the medical professional. Edge device implementation is necessary for retrofitting the existing technology in the clinical setting with modern smart AI features without requiring complete

new monitoring systems.

By completing these four research objectives this thesis aims to evaluate the performance of HIE grading using image-based deep learning networks and determine the optimal representation of the EEG signal for this task. In doing so, a novel method for HIE grading is proposed which surpasses the current state-of-the-art performance. The scope of this thesis will cover the analysis of EEG signals for the purpose of diagnosing HIE in neonatal patients and does not cover the analysis of other modalities such as MRI, aEEG or NIRS. A detailed explanation of the development of this novel method for HIE grading based on the patients EEG signals will be provided throughout this thesis. A publicly available dataset is utilised for this thesis, and the scope of this project includes analysis and usage of both the data and annotations provided in this dataset.

This thesis utilises the EEG signal dataset published for the purpose of developing grading algorithms for background EEG [15]. This dataset has 169 epochs, each one hour long continuous EEG segments and is made up of data from 53 patients [15]. The size of this public dataset is a limitation to this study, however rigorous training schemes have been implemented to account for this limitation, thus ensuring the robustness and generalisability of the developed approach.

A. Structure of the Thesis

This thesis will follow the structure below which details the background information of HIE grading, analysis of the data, and development of a novel baseline approach for HIE grading based on 2D neural networks. This is followed by a novel method for HIE grading based on the sonification representation of the EEG signal and finally the implementation and deployment of this novel approach on a edge device.

Chapter 2: Background Information and Preliminary Feature Analysis: This chapter focuses on the background of the research, detailing the current methods for HIE grading in clinical practice by medical professionals. It then discusses the current state of the art in HIE grading, the use of 2D neural networks in medical diagnosis and the use of sonification for EEG signal analysis. The HIE dataset used in this project is detailed and quantitative feature analysis is performed on the dataset to deepen the author's understanding of the correlation between these features and the HIE grade.

Chapter 3: Spectrogram-Based HIE Grading Using 2D CNN: This chapter looks at the introduction of a novel approach to HIE grading based on 2D neural networks. Spectrogram representation of the EEG signal is used as input to the image based networks, providing a time and frequency representation of the EEG signal. The training and evaluation of this novel approach are detailed in this chapter, providing the performance metrics for the proposed method, which will be later used as the baseline results for comparison.

Chapter 4: From EEG to Audio: A Novel Spectrogram-Based Approach for HIE Grading: Further development on the use of 2D CNNs for HIE grading is conducted with the focus on enhancing the signal prior to input to the CNN. Novel signal pre-processing techniques are used, converting the EEG signal to the audio domain using FM/AM sonification algorithm. In addition, novel signal post-processing techniques are deployed in the proposed workflow. Evaluation and discussion on this enhanced method for HIE grading is performed and new state-of-the-art results are provided, highlighting the benefit of the proposed method.

Chapter 5: Implementation on Qualcomm Snapdragon Board: Finally, the implementation of the trained model will be tested on the Qualcomm RB3 Gen 2 Vision Kit development board and quantization of the proposed method is conducted. Evaluation of the performance for the quantized model is performed and assessment of the deployment on the AI specific hardware of the target device is completed.

Chapter 6: Conclusion and Future Work: The final conclusion of the study will be provided in Chapter 6, summarising the key findings from the results and what the implications of these findings are. A suggestion for future work to be implemented will be discussed also to further this research.

2. Background Information and Preliminary Feature Analysis

ACCURATE grading of Hypoxic-Ischemic Encephalopathy (HIE) is crucial to ensure timely intervention and optimal patient outcomes, as highlighted in the introduction. However, diagnosing HIE remains a complex and time-sensitive process due to the intricate nature of the condition and the rapid progression of brain injury following the initial event. To tackle this, clinicians rely on a multi-modal approach that integrates various diagnostic tools and techniques to assess the severity of HIE in neonates. This chapter delves into the current methodologies for HIE grading in clinical environments, exploring the strengths and limitations of these techniques.

HIE diagnosis often involves a combination of physical examination, neuroimaging, electrophysiological studies, and biomarker analysis to provide a comprehensive assessment of the neonate's condition [16]. Each of these modalities plays a role in identifying the severity of brain injury, although each also presents its own limitations and challenges, especially given how early and accurate diagnosis is essential for effective treatment.

While physical examination, such as observations of abnormal pupil reactivity and motor responses, are highly predictive of poor outcomes, these are most effective 24 hours after the event [16]. The use of MRI and Diffusion Weighted Imaging (DWI) has been found to be effective for early diagnosis successfully detecting pathological changes and subacute lesions in the brain hours after birth [17]. However, the severity of the injury range identifiable on the MRI is limited, with difficulty of diagnosising HIE increasing for minor brain injuries [16]. This is clearly noted by the fact that 26% of neonates who had normal MRI findings and underwent therapeutic hypothermia treatment still suffered neurological impairments [16]. In addition, transportation to MRI machine is often not possible for some unstable neonates [18]. Neuroimaging techniques instead have shown to be better utilised for follow up care and treatment plan development by the clinician [18]. Biomarkers is another promising technique for early HIE detection, proving to be good indicators of HIE severity at an early phase [5]. A high sensitivity and specificity was achieved for electrophysical studies, using EEG and Amplitude-Integrated EEG (aEEG) with a sensitivity of 0.93 and specificity of 0.9 for aEEG [17]. A limitation to aEEG, however, is the potential decrease in amplitude for patients undergoing therapeutic hypothermia treatment, which compromises its suitabil-

ity for HIE diagnosis [19]. However, electrophysical studies possess an advantage over all other modalities due to the ease of monitoring, where the aEEG can be monitored bedside and, so, limiting any additional stress on the neonate due to monitoring its health [20].

While limitations and downfalls to each diagnostic modality exist, a combination of several of these diagnostic methodsis typically used to achieve an accurate diagnosis for the patient. This is a time consuming process, however, which is not optimal for the clinical setting since multiple studies have highlighted the time constraint on HIE diagnosis. This is further emphasized by the 100% improvement rate in patient outcome when treatment was received within 6 hours of the injury occurring [13]. No improvement was identified in patients with severe HIE who received treatment between 6 - 12 hours after birth [13]. An overview of the systematic workflow for how HIE is graded currently in a clinical setting is provided in Fig. 1.

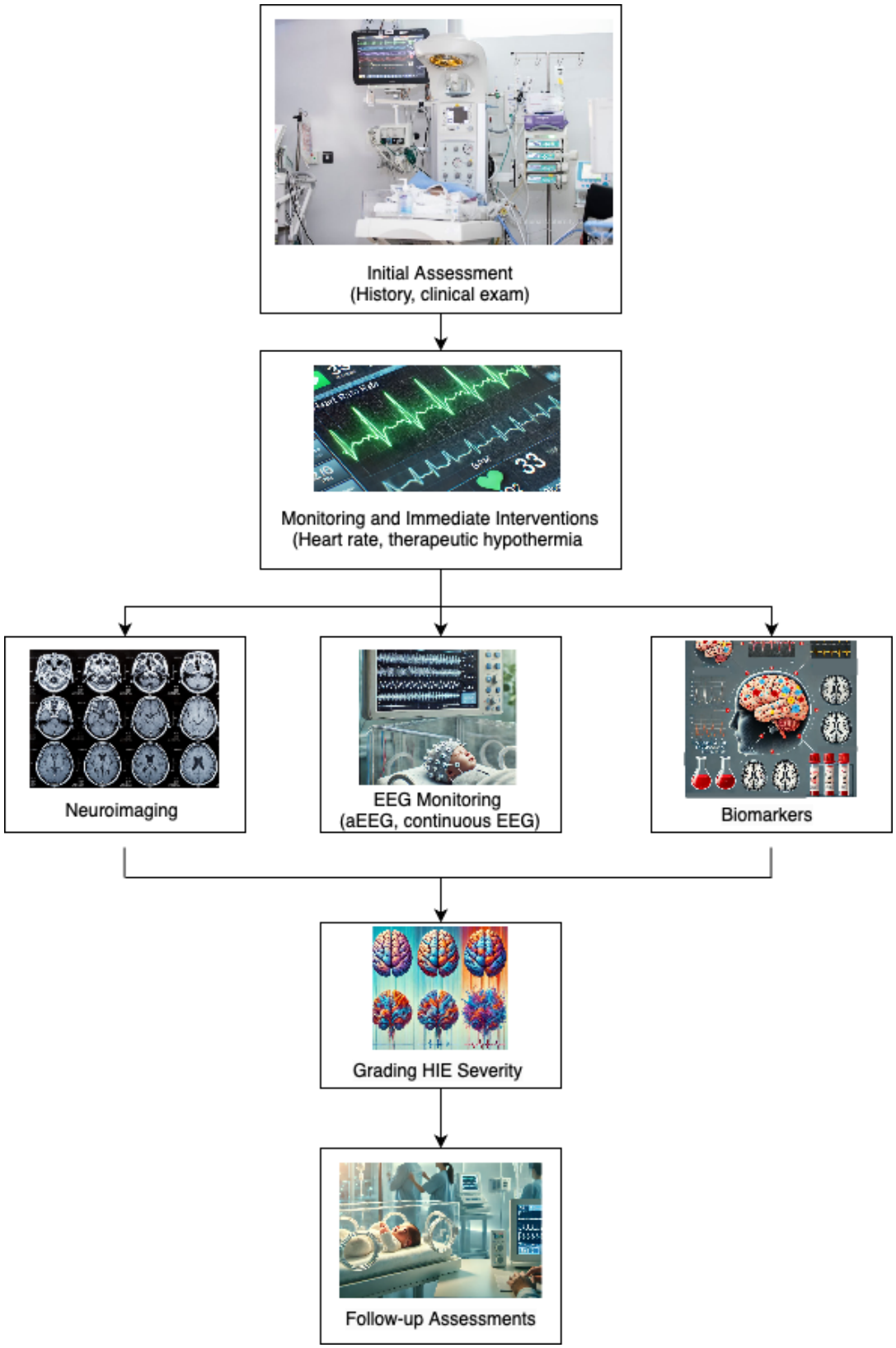


Fig. 1. Current Clinical Diagnosis Workflow for HIE Grading

This is a long and complex process, which is mentally taxing and time consuming on the physician. The review time of EEG signal alone has been reported to be anywhere from one seventh up to half the length of the original EEG recording, where abnormal

EEGs are on the longer range of this [21]. A minimum of 20 minutes of EEG was seen to be useful for HIE grading in [22], with longer windows showing improved results. In order to try to shorten this time consuming process, while improving patient outcome, significant research has been conducted on the development of an automated HIE grading system. The design to primary focus on using EEG signals for development of an autonomous grading system is based on two main reasons. The first is the widespread availability of EEG monitoring equipment in neonatal care units across the globe [20], and the second is due to the achievable sensitivity and specificity using the EEG signal [14]. Other diagnostic methods show lower individual performance. EEG signals have been found to show robust correlation to the HIE grade and to early clinical outcomes from early hours of life [23]. This further supports the reasoning for its use in automated HIE grading systems with current trends in research looking at the use of AI to develop these automated systems.

A. Current State of the Art of HIE Grading

The current state-of-the-art of HIE grading using autonomous methods now focus on the use of deep learning models, where an accuracy of 86.09% on a large unseen test dataset, containing 338 hours of patient data, has been reported by Yu et al [24]. Methods for HIE and background EEG grading based on patients EEG signals have developed significantly in recent years, where studies began initially evaluating the features of the EEG signal and their associated grade by hand [25]. Modern AI, machine learning and deep learning techniques are now used to determine the grade present in the EEG signal [24].

Quantitative EEG features were selected to determine their differentiability between the four HIE grades [26] [27], the amplitude and discontinuity pattern were shown in the post hoc analysis to have resulted in the best representation of the visual analysis of the EEG signal. Further development of automated grading systems explored combining the features of the EEG signals with machine learning techniques to determine the grade of the EEG signal [28]. Guo et al. [28] identified long-term features of the EEG signal, extracting 64 second windows of the EEG data as beneficial input features for Support Vector Machines (SVMs), achieving an overall accuracy of 79.5% and a Kappa score of 0.69.

By representing the EEG signal as a non-linear amplitude modulated signal for each 64 second window of the EEG signal and then further decomposing this signal into its amplitude and frequency components, Stevenson et al. achieved an accuracy of 83% [29]. By performing statistical analysis of these components, the authors defined a multi-class linear discriminant classifier to determine the severity of the HIE present [29]. Extracting long-term features from the sequences of short-term features, Ahmed et al [30] achieved an increase in accuracy to 87% by utilising the supervector approach of combining Gaussian Mixture Model (GMM) with discriminative SVM.

Improving on the work of Ahmed et al. [30], Raurale et al. [31] proposed a novel method for the feature selection through the use of a quadratic time-frequency distribution (TFD). Representing the features as inputs to a 2D CNN model achieved a leave-one-out (LOO) testing accuracy of 88.9%. However, an accuracy of 69.5% was achieved on a large private unseen dataset, consisting of 228 hours of patient EEG. This approach showed an improvement of the previous work of Ahmed et al [30], which achieved a test accuracy of 56.8% on the same unseen dataset.

Yu et al [24], proposed a method using fully convolutional layers for both the feature extraction and classification blocks, rather than using handcrafted features as proposed in [30] or using TFD for feature extraction, as in [31]. Yu et al achieved an accuracy of 86.09% on a large unseen dataset, a relative increase in accuracy of 23.65% compared to the previous baseline on this dataset by Raurale et al [31]. This dataset contains 338 hours of patient data. This was achieved by proposing a method to maintain temporal position information while reducing the number of parameters in the model.

Validation of the growing role of AI to enhance the decision making process in healthcare is evident from the findings in [32]. This study shows the potential of open-source grading algorithms on a large unseen test dataset at both predicting whether HIE is present and also grading any cases where the model believes a HIE diagnosis exists [32]. This study has significant implications for highlighting the applicability of autonomous HIE grading methods in a clinical setting.

This analysis of the literature on autonomous HIE grading methods shows the potential increase in performance by using deep learning networks. However this research has not yet explored in-depth analysis of 2D spectrogram based CNNs to their full potential. Indeed, time-frequency image representations of the signal have been used as

CNN input only once in the literature thus far (by Raurale et al in [31]), where 5 minute input windows of the EEG signal were used. Very promising results for spectrogram based 2D CNNs have been seen in other biomedical fields, thus the motivation for this study to harness their power for HIE grading.

B. Current State-of-the-Art for 2D Neural Networks in Biomedical Fields

Recent advances in the use of artificial intelligence in the biomedical domain have been seen in processing spectrogram images derived from physiological signals, primarily EEG and ECG data. By transforming one-dimensional time series signals into two-dimensional time-frequency spectrograms a rich representation of a non-stationary signal can be achieved. By capturing both frequency and time variations, an increase in the information available for analysing complex biomedical data aids with diagnosing diseases such as cardiac arrhythmias and neurological disorders [33] [34].

2D CNNs based on spectrogram parameters of the EEG signal have been used in a wide variety of applications from seizure detection [35] [36] to emotion classification [37] to identifying neurological disorders such as autism, epilepsy, Parkinson's disease and schizophrenia [38]. These studies all show that by representing the EEG signal as a spectrogram image, sufficient information for classification can be learnt by the CNN model. This suggests that there may be other clinical diagnoses that could be automated by this configuration, and here its becoming a primary enabler for the analysis and research conducted in this thesis.

In cardiology, 2D CNNs have been very effective in detecting arrhythmias from ECG signals when represented as spectrogram images [39] [40] [33]. A high accuracy in arrhythmia detection was identified, with results equal to and surpassing the state-of-the-art based on 1-D neural networks for the same task. This research shows the potential of spectrogram based representations to highlight rhythmic behaviour of the physiological signal, thus aiding the neural network in detecting and classifying these patterns. This is particularly relevant to the HIE grading task in this thesis, where a defining parameter of the HIE grading is the rhythmic pattern between active and inactive segments of the EEG signal. Evaluation of methods to further enhance this rhythmic pattern were reviewed in the literature with the goal of increasing the grading potential, prior to plotting the spectrogram image.

C. Current State of the Art for EEG Sonification

By converting the EEG signal to the audio domain, the rhythmic behaviour of the EEG signal can be extracted, thereby aiding diagnosis, particularly in the field of seizure detection. This has been explored for both neonatal and adult EEG signals, where sonification is used as a tool to enhance diagnosis [41] [42]. It has been seen as a useful tool to decrease the review time and enhancing the information available at the decision making step [41] [43]. The motivation for representing EEG signals in the audio domain stems from the human ability to perceive different pitches and slow evolutions in frequency. Different methods for EEG signal sonification have been tested. The overall consensus shows that, with a short training time for the medical professional, a high accuracy was achieved for seizure prediction based on the audio representation of the EEG signal [41] [43] [44].

A number of sonification algorithms have been presented in the literature to date for conversion of EEG signal to the audio domain, those of note are outlined here. Gomez et al [41] details two methods for EEG sonification. The first method, the Phase Vocoder, performs very well at capturing the temporal evolution of the EEG signal with the second method, FM/AM sonification, being tuned to highlight rhythmic patterns in the EEG signal. In [44], auditory differentiation between the normal and abnormal EEG activity was devised. A pentatonic scale is used to transform the EEG signal into musical notes and the EEG values are then linearly mapped to fit the scale. Event-based sonification is used by Baier et al where the rhythmic events of epileptic seizure is emphasized and the background activity is suppressed [43].

While research into sonification for seizure detection has shown promise, the possibilities of this have not yet expanded to the exploration of background EEG signal grading. This thesis explores the expansion of sonification methods for EEG grading, with a particular focus on amplitude and frequency modulation to enhance the specific features and rhythmic patterns in the background of the EEG signal, while reducing the impact of insignificant fluctuations in the signal. When using such techniques in conjunction with deep learning networks a new state of the art in autonomous HIE grading was achieved. However, in order to implement this, a high quality dataset is required. This dataset must be reflective of real-world environments in order to develop a robust grading algorithm, suitable for real-world application. For this, the publicly

available dataset published by O'Toole et al is used [15].

D. Dataset

There were 53 patients evaluated in this background EEG dataset, each with a diagnosis of HIE. These EEG signals were collected from neonatal care units in Ireland, Netherlands, Sweden and the UK. Of these 53 patients, 40 weeks was the median gestational age, with 32 male patients and 21 female patients [15]. More than half underwent therapeutic hypothermia treatment [15]. For all 53 patients, the EEG analysis commenced in less than 48 hours of the neonate's birth using either the NicoletOne ICU monitor or the Neurofax EEG-1200 EEG monitors [15]. A maximum of 5 epochs per patient were included in the dataset, with a median of 3 epochs per neonate [15].

This dataset accumulates to 169 hour-long epochs, measured using a compressed representation of the 10:20 international system [45], with nine EEG channels gathered from the F3, F4, C3, C4, O1, O2, T3 and T4, Cz locations [15]. For the Neurofax recordings an average of C3 and C4 was taken as the reference channel and FCz was used for the NicoletOne recordings, which was a midline placement between Fz and Cz [15]. A sampling rate of 200 Hz and 256 Hz for the respective monitoring systems was used [15]. For analysis in this study, an 8-channel bipolar representation of the data is utilised.

The dataset consists of an uneven distribution across the four HIE grades, with grade 1 being most dominant with 104 epochs (61.5%). The distribution of the data across all four grades is detailed in Table I. The data was graded into four categories: normal/mild, abnormal, moderately abnormal and majorly abnormal based, on the grading system detailed by Murray et al [46]. An increase in severity as the grade increases is noted and the key specifications of each grade are defined as:

- **Grade 1 (Normal / Mild):**
  - Grade 0 and Grade 1 are combined in this dataset to a single grade.
  - Continuous EEG background pattern
  - Mild disturbance to normal physiological features
- **Grade 2 (Abnormal):**
  - Interburst Interval (IBI) of less than 10 seconds
  - Disturbed Sleep-Wake Cycles (SWC)

  - Asymmetry and asynchrony

- **Grade 3 (Moderately Abnormal):**
  - Attenuation of background EEG patterns
  - No Sleep-Wake Cycles (SWC) present
  - IBI between 10 to 60 seconds
- **Grade 4 (Majorly Abnormal):**
  - Severe discontinuity
  - IBI greater than 60 seconds
  - Attenuation of background EEG to below 10 microvolts

The key features listed for each HIE grade can be visually identified based on the 36 second window of EEG signal plotted in Fig. 2. The most notable feature based on visual inspection is the increase in discontinuity and lack of activity in the signal as the severity of the condition increases, along with the depreciation in signal amplitude.

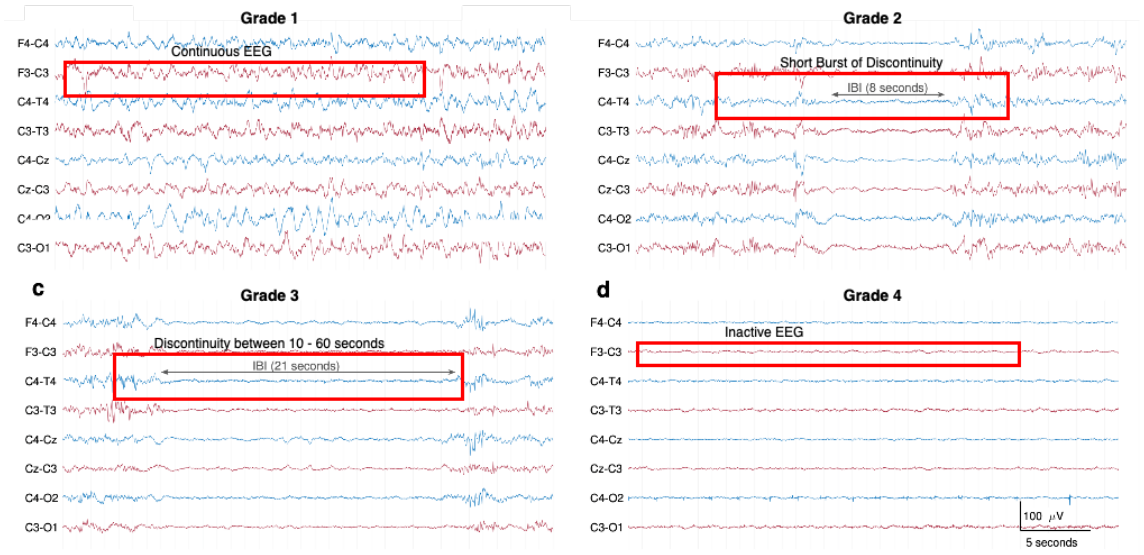


Fig. 2. A 36 second window of EEG signal for all four HIE grades, the image was adapted from [15] to highlight the behaviours of the EEG signal dependant on the severity of HIE present.

All data epochs in this dataset were independently graded by two clinical physiologists. Any disagreements in grading were resolved by a joint review by both clinicians and a consensus grade was determined [15].

TABLE I

DETAILS OF DATASET DISTRIBUTION ACROSS GRADES

	Grade 1	**Grade 2**	**Grade 3**	**Grade 4**
Severity	Normal / Mild	Abnormal	Moderately Abnormal	Majorly Abnormal
Hours of Data	104 h	31 h	22 h	12 h
Distribution Percentage	61.54%	18.34%	13.02%	7.1%

A deeper understanding of the EEG signals in this dataset and the relationship between their features and the HIE grades was desired by the author and so feature extraction and analysis of the signals was conducted. This was necessary for the development of the signal pre-processing phase in the machine learning pipeline, in order to emphasise the differentiability between the grades. In order to increase the author's understanding of the EEG signal parameters and how they fluctuate for various HIE grades, comparative analysis of the features across the four grades is performed and any differentiating features are noted. This was implemented as background research prior to the development of novel methods for the task of HIE grading. Initial inspection looked at identifying the power in specific frequency bands of the Power Spectral Density (PSD) of the overall EEG signal. The PSD, $S_X(f)$, of the signal is defined in Equation 1 [47].

$$S_X(f) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau \tag{1}$$

where $S_X(f)$ is the power spectral density as a function of frequency f, $R_X(\tau)$ is the autocorrelation function of the signal $X(t)$, τ is the time lag, j is the imaginary unit, and f is the frequency in Hertz.

The average power in each sub-band is calculated on a per-grade basis to determine any patterns associated with the HIE grades, the results of which are shown in Fig. 3.

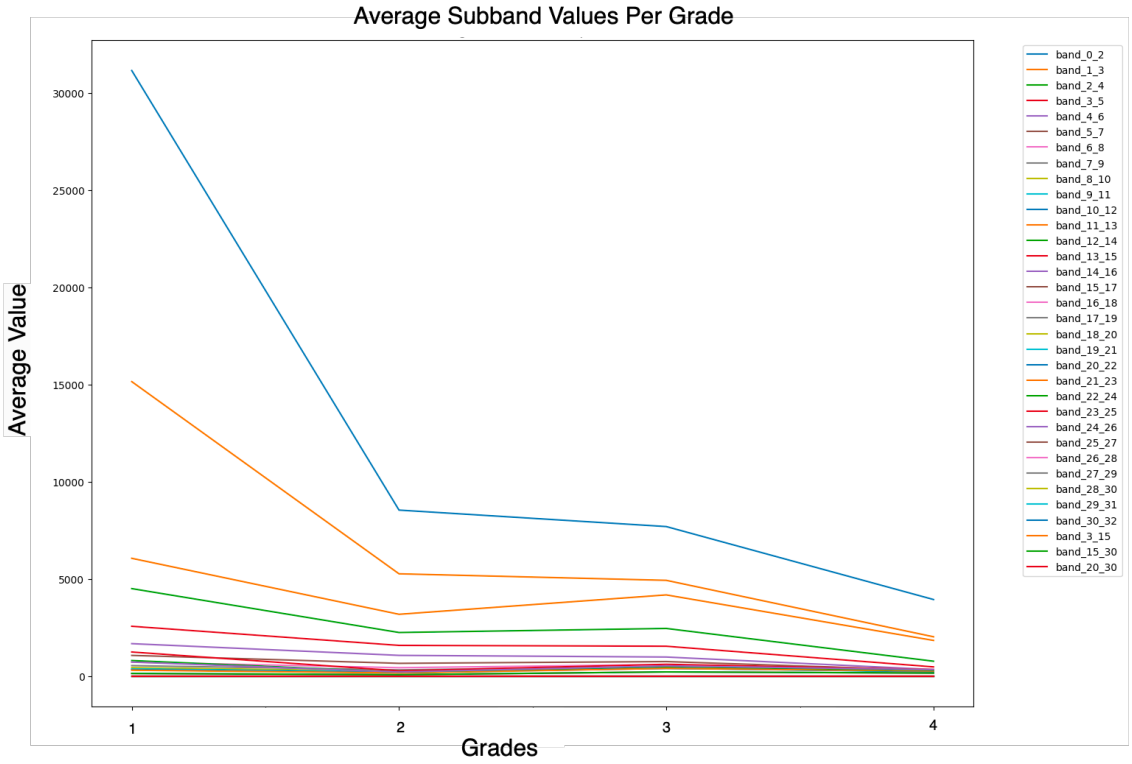


Fig. 3. Average Power in Each Sub-band of the PSD.

It is evident from Fig. 3, that sub-bands between 0 - 2 Hz and 1 - 3 Hz show a significant drop in power between grades 1 and 2. The power difference between grades 2 and 3 and grades 3 and 4 is significantly lower than the drop which occurred between grades 1 and 2. This is indicative of the power in the Delta frequency band, which is usually defined as 1-4 Hz being a key feature in determining the HIE grade [48]. In neonates, elevated delta activity can indicate disruptions in normal brain function, reflecting alterations in corticol function. It is also clear from this figure that the powers in the higher frequency sub-bands are less significant in determining the HIE grade, aligning with the knowledge that maximal power in neonatal brain activity declines as the frequency increases [49]. Performing feature importance analysis on the PSD sub-band data alone, with Scikit-Learn's Random Forest Classifier [50], showed the sub-bands of most relevance to the HIE grade being those listed in Table II.

TABLE II

TOP IMPORTANCE SUB-BANDS FOR HIE GRADING

Importance	**Band Frequencies (Hz)**
1	0 - 2 Hz
2	1 - 3 Hz
3	2 - 4 Hz
4	3 - 5 Hz
5	4 - 6 Hz
6	5 - 7 Hz
7	6 - 8 Hz
8	7 - 9 Hz
9	8 - 10 Hz
10	3 - 15 Hz

While the above sub-bands show the highest feature importance, in order to evaluate their ability to determine the HIE grade a Random Forest Classifier was trained, with an initial cross-validation accuracy of 0.70 ($\pm$ 0.04). It is important to note that a patient independent model was not designed in this initial testing as it was conducted purely to determine the effectiveness of the features in determining the HIE grade. This analysis showed that the frequency component of the signal is a strong indicator of the HIE grade and a key feature to highlight later in the research when representing the signal for input to the AI model.

Increasing the feature set from the PSD of specific frequency bands, the overall best features for HIE grading are determined, with a grid search analysis to determine the optimal number of features for best HIE grading. 98 features in total were assessed with the feature importance of the top 20 features based on Random Forest selection being illustrated in Fig. 4.

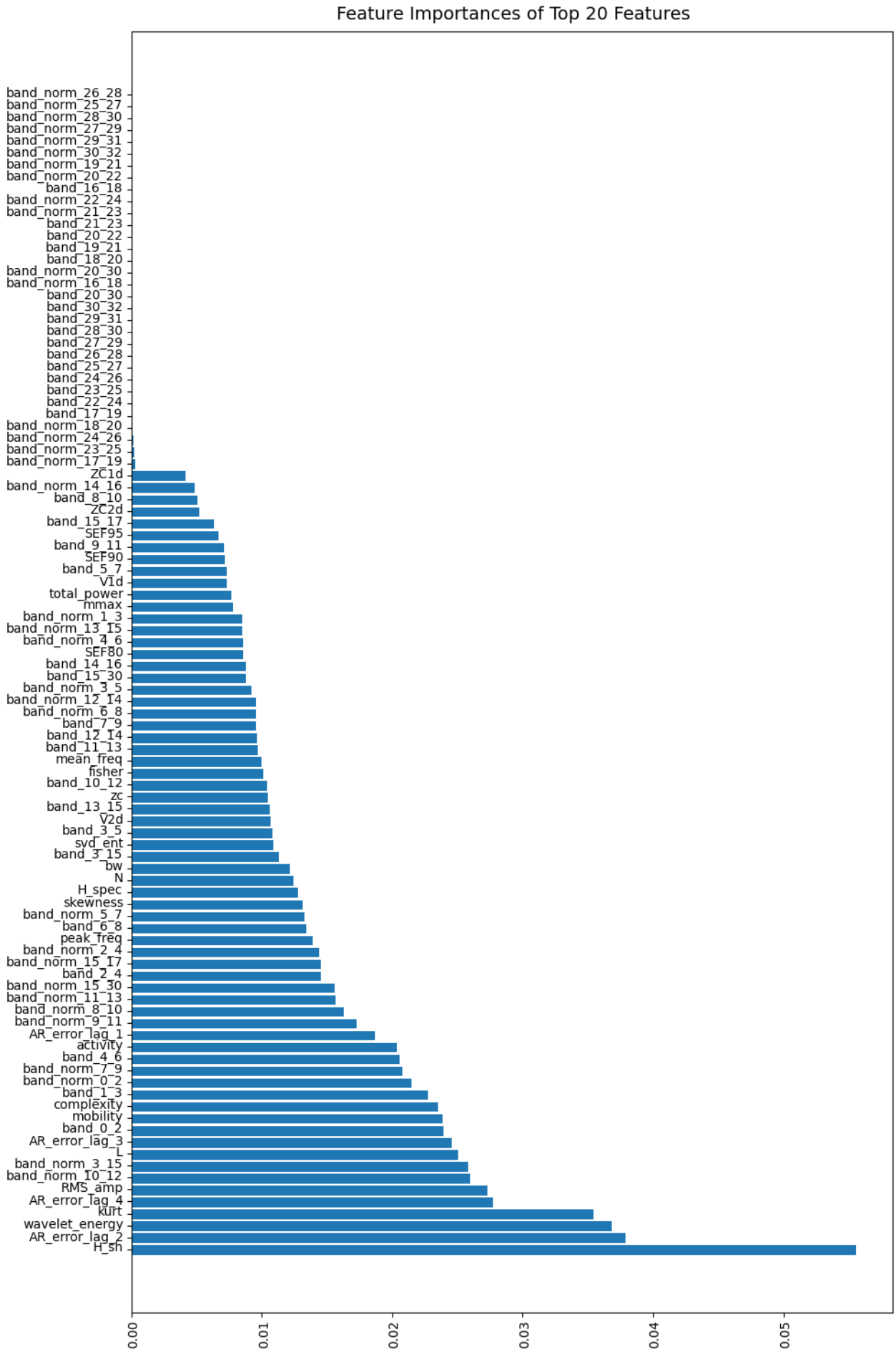


Fig. 4. Feature Importance Using Scikit-Learn Random Forest Classifier of Top 20 Features from Feature set of 98 features of the EEG signal.

The most important feature is clearly Shannon Entropy (H_sh), a measure of the signal's uncertainty or unpredictability based on its probability distribution [51]. In the context of the EEG signal, this metric captures the amplitude and frequency variation characteristics of different HIE grades. High entropy values indicate a sporadic, high-amplitude, and high-frequency EEG pattern, often observed in less severe HIE cases. In contrast, lower entropy values reflect the flatter, more predictable patterns found in severe HIE grades, such as grade 4. Therefore, Shannon Entropy not only aligns with the hypothesized role of rhythmicity in HIE differentiation but also quantifies the distribution complexity that may underlie the varied clinical presentation across HIE grades. The box plot in Fig. 5 shows the distribution of the Shannon Entropy values across the four HIE grades.

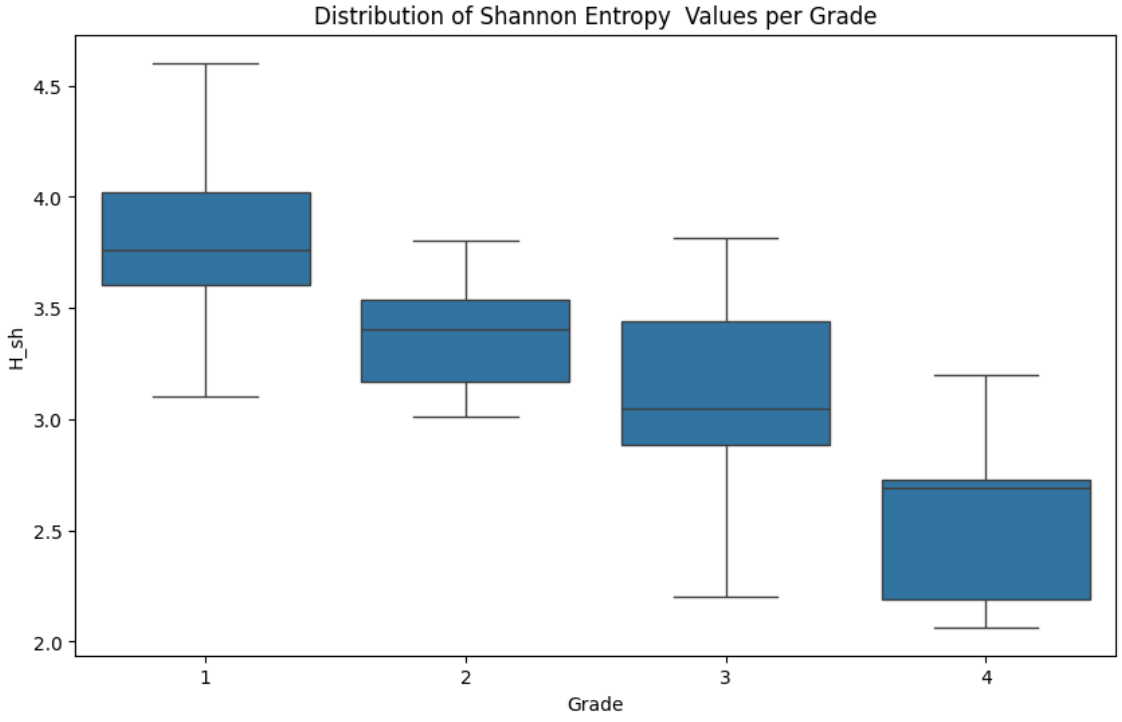


Fig. 5. Box Plot for the Shannon Entropy Value for all Four HIE Grades

From the box plot in Fig. 5, the differentiability of the Shannon Entropy can be noted from the separated average value for each HIE grade. However, it is evident, due to the high inter-quartile range of the data, that this feature alone is not strong enough to use as a differentiator for HIE grading. This is further supported by the cross-validation accuracy of approximately 67% achieved when this single feature is used in the Random Forest classifier, as seen in Fig. 6. The overall trend of the entropy values across the

HIE grades is clear, where grade 1 shows the highest entropy values, with a decrease in entropy values as the HIE severity increases. This is due to the highly active brain signal of grade 1 patients, seen in the time series representation as sporadic and highly chaotic signals. The average entropy values for grade 4 is negative, indicating a highly predictive signal which aligns with the knowledge of grade 4 HIE, which is primarily a flat and inactive EEG signal. While the Shannon Entropy value is a good indicator of the different HIE grades, there is some overlap between the interquartile ranges for each class seen in Fig. 5. This is expected due to the highly non-linear behaviour of the EEG signals and the fact that no exact metrics define each HIE grade but, instead, each class is a range on a grading scale. The nature of this means that when analysing the EEG signals it can often be very difficult to differentiate the signals from neighbouring classes, especially when considering only a single feature of the signal. Evaluating the grading performance of a Random Forest Classifier for a range of 1 to 98 features, the test accuracy is plotted in Fig. 6.

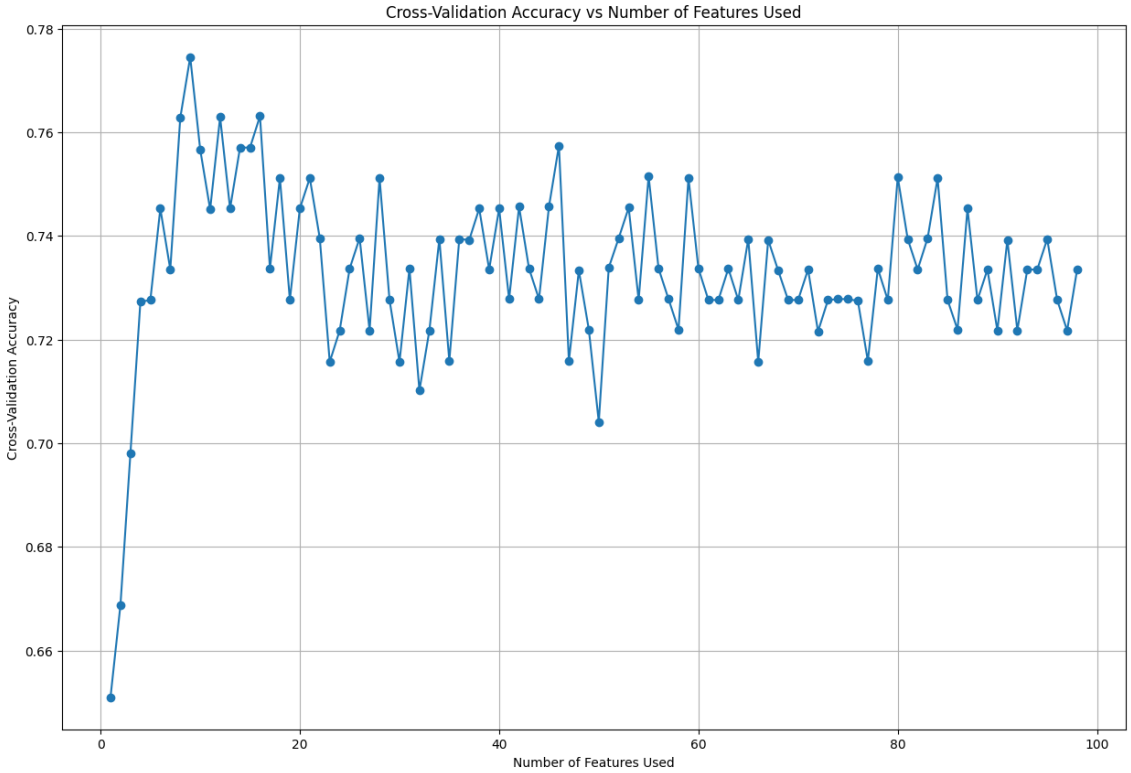


Fig. 6. Cross-Validation Test Accuracy for Sweeping Range of Most Important Features Selected and Utilised for Random Forest Classifier.

Optimal results were achieved when the top 9 features were used in the Random Forest Classifier, as seen in Fig. 6, with a test accuracy of 77%. As previously discussed, this

feature importance test was performed based on Random Forest. However, SelectKBest from Scikit-Learn [52] and Bonferroni Corrections [53] were tested where 9 different top features were returned. This resulted, however, in equal average test accuracy for nested cross-validation but with an increase in variance across the folds and so the Random Forest Classifier returned optimal performance. A list of the 9 most important features is provided below, along with an evaluation of the relevance of the feature to the evolution of severity of HIE present in the EEG signals. In addition, the box plot of these 9 features across the four HIE grades can be seen in Fig. 7, representing the distribution of the features for each HIE grade and so supporting the interpretation of the features as the severity of the brain injury evolves.

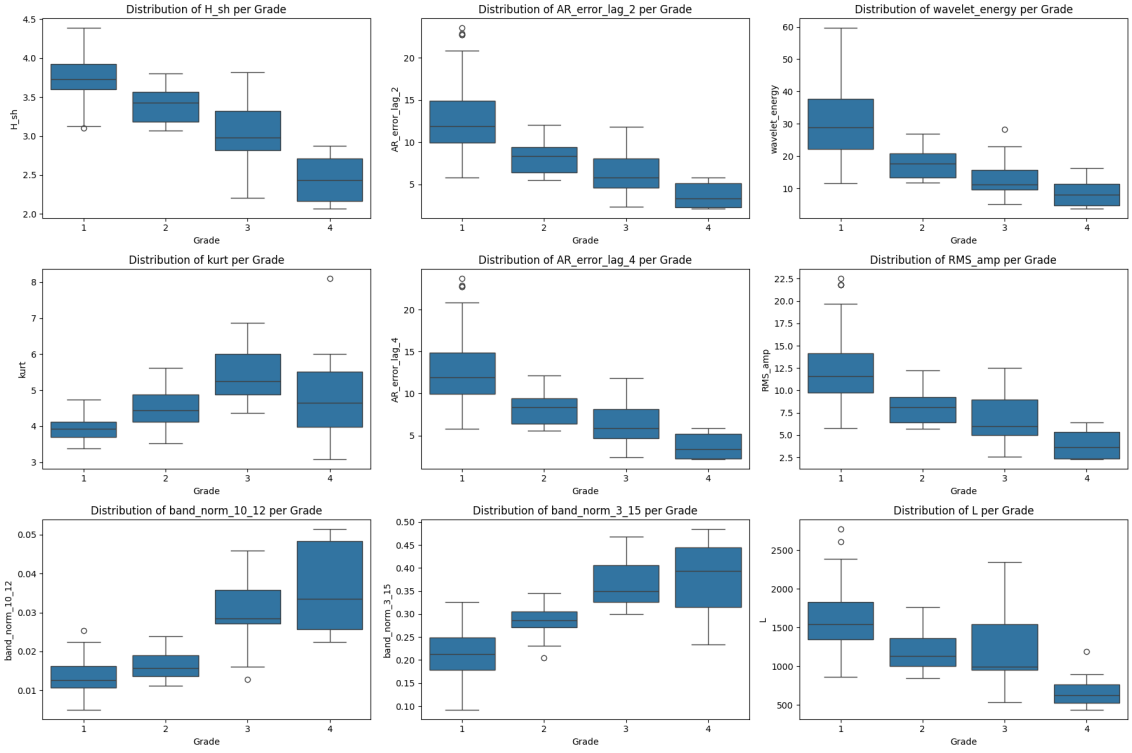


Fig. 7. Box Plot for the Top 9 Features for all Four HIE Grades.

- **H_sh**: Shannon Entropy

  Shannon Entropy is a quantification of the randomness or disorder of the EEG signal, with higher values indicating a higher complexity associated with the signal [51]. As previously discussed, the Shannon entropy level decreases as severity of HIE increases due to the lack of random activity present in the EEG when the brain is severely injured.

- **AR_error_lag_2**: Autoregressive (AR) prediction error at lag 2

The AR calculates the error between the actual EEG signal and its predicted value using an Auto Regressive model with a lag set to 2. The AR uses past values to predict future values, with a higher error indicating a larger difficulty in predicting the signal based on past events and, so, indicating an unpredictable signal is present [54]. The error decreases as the severity of the HIE increases, aligning with the results of the Shannon Entropy value, which shows that the more severe the HIE grade, the more predictable the signal becomes due to its limited activity.

- **wavelet_energy**: Wavelet Energy

  The Wavelet Energy is a measurement of the energy content of the EEG signal across a defined frequency scale with wavelet decomposition applied. The sharp decrease in wavelet energy as the HIE increases suggests a reduction in frequency activity as brain function deteriorates, supporting the findings in Fig. 3, where as the HIE grade increases, the power in each frequency band decreases.

- **kurt**: Kurtosis

  Kurtosis is a statistical measure of the skew in the distribution of the EEG signal, capturing the presence of extreme values, or bursts of activity, in the signal [55]. Given the continuous highly active nature of grade 1 signals, a low Kurtosis value is received. The kurtosis value increases for grades 2 and 3, as expected due to the pattern of the EEG being sporadic with varying periods of activity and discontinuity. This aligns with the expected nature of the IBI of the signal for grade 2 and 3, with durations of inactivity of the signal increasing, despite sporadic periods of brain activity still being present [46]. The kurtosis decreases again for grade 4, where the EEG signal is highly inactive and so the distribution of the signal is once again more uniform.

- **AR_error_lag_4**: Autoregressive (AR) prediction error at lag 4

  The lag is set to 4 here, allowing for the capture of slightly longer dependencies in the signal, in comparison to the AR error at lag 2 previously discussed. The error follows the same trend as the AR error at lag 2, with the error decreasing as the HIE grade increases, indicating a more predictable signal as activity in the signal decreases and the HIE grade increases.

- **RMS_amp**: Root Mean Square (RMS) Amplitude

  A measure of the overall signal strength is provided by the RMS value, allowing

for analysis of the intensity of the EEG signal [55]. As expected, the RMS value decreases as the HIE increases with lower grades representing weakened or suppressed brain activity, in accordance with, the definition of the four HIE grades [46]. This shows the impact of the signal amplitude as a defining characteristic for HIE grading.

- **band_norm_10_12 and band_norm_3_15**: Normalized Power in Band 10-12 Hz and Band 3-15 Hz

  The normalized power in the frequency range of 10 - 12 Hz and the broader range of 3 - 15 Hz is calculated, indicating how much of the total signal power is concentrated in these specific frequency ranges [56]. The frequency range of 10 - 12 Hz encompasses the majority of the Alpha band of the EEG signal [57]. The second band covers a larger window of frequencies, capturing the Delta, Theta, Alpha and Beta bands of the EEG signal [57]. Analysis of the normalized power in these bands help to highlight changes in brain function, with analysis of the EEG frequency bands having shown promising results for disease diagnosis in the past [58] [59] [60].

- **L**: Curve Length

  The curve length captures the signal's variability and fluctuation over time by measuring the cumulative length of the EEG signal curve [61]. A shorter curve indicates decreased brain activity and so is associated with a more severe HIE grade, as less fluctuation is present in this signal.

This analysis helped to increase the understanding of the EEG signal features most relevant to determining the four HIE grades present in the EEG signal. It also shows promising initial results for grading while highlighting the importance of rhythmic behaviour, signal amplitude and frequency in distinguishing HIE grades. Through feature analysis, Shannon Entropy emerges as a central measure of the signal's unpredictability, capturing fluctuations in amplitude and frequency that vary across HIE grades. Additionally, autoregressive errors and frequency-specific power in low-frequency sub-bands highlight shifts in predictability and signal strength as the severity of HIE progresses. This aligns with the distinct rhythmic patterns of each grade. This research was implemented as a primary baseline and provided a foundational understanding of the metrics of interest for each HIE grade. It is evident from this set of 9 most important

features that the amplitude of the signal, signal frequency and the rhythmic pattern of the signal are the key defining characteristics of the four grades. This informed the development of a novel deep learning approach that utilizes spectrogram representation of the signal to highlight the amplitude and frequency components of the signal in an image representation. This data-driven approach can be used to train 2D CNNs and enhance the grading performance. In addition, the information gained throughout this background research guided the development of the FM/AM transformation of the EEG signal to enhance the signal representation prior to the AI analysis, the details of which are discussed later. The next chapter will focus on both the methodology and results of the novel approach, where a 2-dimensional (2D) CNN is trained based on Mel Spectrogram representations of the EEG signal.

3. Spectrogram-Based HIE Grading Using 2D CNN

THIS chapter details the development of a baseline approach to HIE grading using an image based 2D CNN trained on spectrogram representations of the EEG data. While current research in the field of HIE grading has looked to use deep learning for both feature extraction and classification, little work outside that of Raurale et al [31] has been published on image based neural networks. To harness the true power of 2D CNNs for this HIE grading task, the EEG data must first be pre-processed in a suitable image format. A recent Kaggle competition released by Havard Medical School on Harmful Brain Activity Classification [62] showed the vast potential for spectrogram image representations in EEG classification, as well as in conjunction with classical time series representations. However, this was not focused on HIE, nor was it implemented for neonatal patients. Instead, this competition focused on detecting seizures along with five other brain activity patterns and used on very short EEG windows, looking at windows of 1 minute duration. This chapter will focus on the workflow from spectrogram creation to HIE grade output of the developed algorithm using 1-hour EEG input windows. The results and discussion on the performance of this algorithm conclude this chapter.

A. Data Preprocessing

The first step in the signal pre-processing was to filter the raw EEG signal. A notch filter is applied to remove any noise present in the signal due to the power supply, and a high-pass filter is used to attenuate low frequency drifts below the high-pass cut-off frequency using a first order Butterworth filter. A low-pass filter is then implemented to suppress the high-frequency noise in the signal using a finite impulse response (FIR) filter. After this, the filtered signal could be used to compute the time-frequency representation of the signal using spectrograms. Mel Spectrograms have been selected for use in this analysis.

The Mel Scale is a non-linear scale that approximates the way in which humans perceive pitch [63], where the Mel Scale puts emphasis on the frequency component of the signal smaller than 1 kHz. This is the range in which human hearing is most sensitive. It also highlights the rhythms present in the EEG signal, thereby improving the HIE grade differentiability, which is discussed in more detail later. The Mel Spectrogram can be derived from the Short-Term Fourier Transform (STFT), by slicing the signal

into short window frames, as in Equation 2, where the frequencies are represented on the Mel Scale.

$$X(t,f) = \sum_{n=0}^{N-1} x(n)w(n-t)e^{-j2\pi fn} \tag{2}$$

where $X(t,f)$ is the STFT of the signal $x(n)$, $w(n-t)$ is the window function applied to the signal, f is the frequency and t is the time shift or frame index. The power spectrogram is then computed based on the magnitude of the STFT, as per Equation 3.

$$|X(t,f)|^2 = \text{Re}(X(t,f))^2 + \text{Im}(X(t,f))^2 \tag{3}$$

A series of triangular Mel filter banks are applied to the frequency spectrum obtained from the power spectrogram in order to map the frequency components to the Mel scale. The triangular filters are centred around frequencies spaced evenly in the Mel scale, with the filter response defined as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \leq k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{4}$$

where $H_m(k)$ represents the Mel filter response for the m-th filter and the frequency bin k. The triangular shape of the filter stems from their design to linearly increase from $f(m-1)$ to $f(m)$, with $f(m)$ the centre frequency, and linearly decrease from $f(m)$ to $f(m+1)$. The centre frequencies of the Mel filters are computed based on the non-linear Mel Scale, as defined in Equation 5.

$$f(m) = \left(\frac{N}{f_s}\right) F_{\text{mel}}^{-1}\left(F_{\text{mel}}(f_l) + m\frac{F_{\text{mel}}(f_h) - F_{\text{mel}}(f_l)}{M+1}\right) \tag{5}$$

where N is the length of the FFT, f_s is the sampling frequency, f_l and f_h are the lower and upper frequency limits of the Mel Scale respectively, F_{mel} is the Mel Scale function and M is the number of Mel filters. The Mel Scale function converts a linear

scale frequency, f, in Hz, into the corresponding frequency in Mel units, as per Equation 6. In this case, it is converting the frequencies from the power spectrogram of the EEG signal. To represent the human auditory system's decreased sensitivity to pitch difference at high frequencies, the Mel Scale function compresses the frequency scale at higher frequencies.

$$F_{\text{mel}}(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{6}$$

Finaly, the logarithm of the filter bank outputs is calculated to compute the Mel Spectrogram as per Equation 7.

$$S_m(t, m) = \log\left(\sum_{k=0}^{N-1} |X(t, k)|^2 H_m(k)\right) \tag{7}$$

While using the Mel spectrogram for raw EEG signals is not common practice, it has been shown to provide significant advantage in highlighting the features relevant to each HIE grade. It aids in capturing the important rhythms present in the Delta, Theta, Alpha and Beta frequency bands of the EEG. These four frequency bands of the EEG signal are known to define multiple neural states and so are relevant discriminative features across the four HIE grades. In addition, the Mel Spectrogram allows for representing the signal amplitude along with the frequency in a single image representation, and so encompasses the three elements of the signal identified as relevant for HIE grading from the background research. To recap, these are the signal amplitude, frequency and rhythmic pattern.

A representation of the Mel Spectrogram for each grade is shown in Fig. 8 below. The parameters of the Mel Spectrogram used to obtain the 256×512 input image for the eight vertically stacked spectrograms are a window/shift ratio of $1024/512$ and 32 Mel filters. This uses the full 1-hour segment of the EEG and represents all eight bipolar channels as a single image.

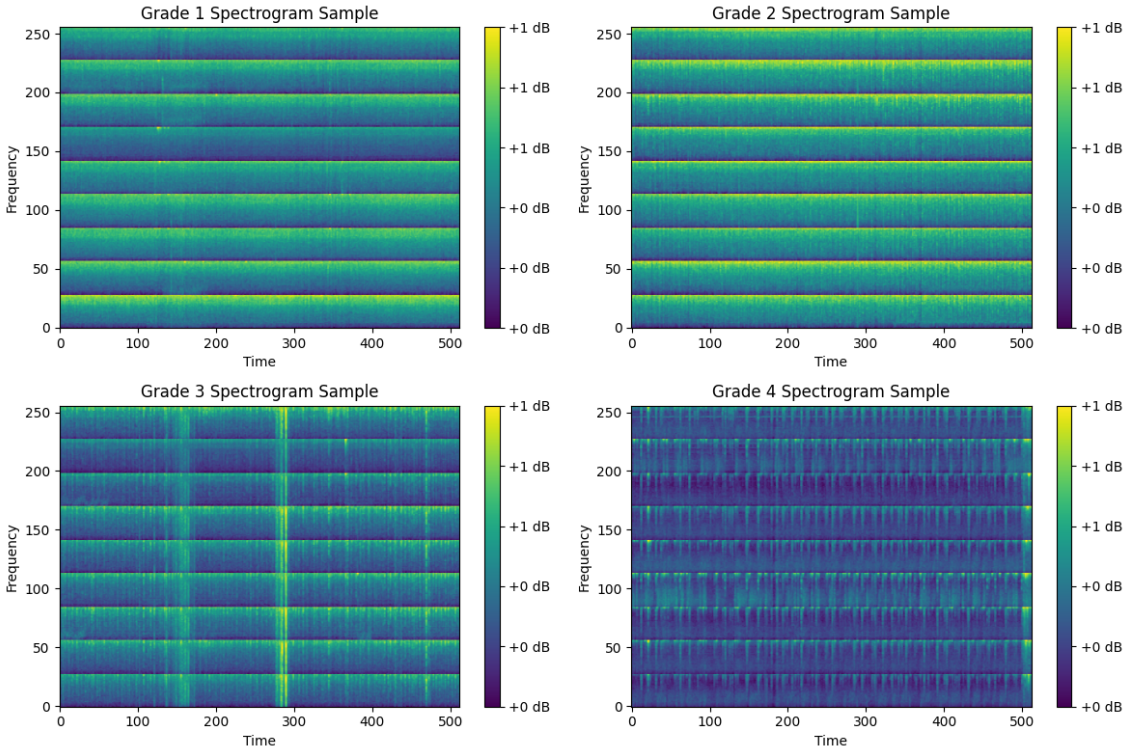


Fig. 8. Mel-Spectrogram of EEG Signal for the full 1 hour segment of EEG, with a sample for each HIE grade

B. Machine Learning Model Performance Metrics

The performance of the model is evaluated primarily based on its loss and accuracy metrics when comparing different hypothesis tests to determine the optimal model parameters. Given that a regression model is implemented, the continuous regression value is rounded based on the optimal thresholds identified during training. This discrete class value is then used to calculate the accuracy score of the model. A mean-Squared Error (MSE) loss function was applied during the model training, as defined in Equation 8 [64].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

where n is the number of samples, y_i is the true value and $\hat{y}_i$ is the predicted value. The MSE loss function is used to measure the average squared difference between the true and predicted values, with a lower MSE indicating a better fit of the model to the data.

In addition to the model loss and accuracy calculations, the precision, recall and F1-score of the model were calculated. The precision is defined as the ratio of correctly

predicted positive instances to the total number of instances predicted by the model, given mathematically in Equation 9 [65].

$$\textit{Precision} = \frac{TP}{TP + FP} \tag{9}$$

where TP is the true positive rate and FP the false positive rate.

Recall is used to assess the model's ability to correctly identify all relevant positive instances in the dataset. It is defined as the ratio of correctly predicted positive instances to the total number of actual positive instances in the dataset, as shown in Equation 10 [66].

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{10}$$

where TP is the true positive rate and FN the false negative rate.

The F1-score is the harmonic mean of the precision and recall, and is used to provide a balance between the two metrics, as shown in Equation 11 [67]. A higher F-1 score defines a greater balance between both the precision and recall of the model.

$$\textit{F1} = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \tag{11}$$

While the final predictions are defined as discrete classes, a regression model has been developed. The motivation to use a regression model is based on the graded scale relationship between the HIE grades, where a monotonic relationship exists between the activity of the EEG signal and the HIE grade. In contrast, a classification model with 4 discrete classes would not highlight this scaled relationship as effectively.

In order to evaluate the regression model behaviours, the R-squared value (R^2), also called the coefficient of determination was calculated, as per Equation 12.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value and $\bar{y}$ is the mean of the true values. The R^2 value is used to determine the proportion of the variance in the dependent variable that is predictable from the independent variable. A value of 1 indicates a perfect fit of the model to the data [68].

The final performance metric used was the Cohen's Kappa score, a statistical metric which measures the level of agreement between the prediction from the model and the ground truth. The Cohen's Kappa score is defined as per Equation 13 and is a robust metric in comparison to a simple accuracy score calculation, due to the fact it takes account of the possibility of an agreement occurring by chance [69].

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{13}$$

where p_o is the observed agreement and p_e is the expected agreement. A Kappa score of 1 is defined as a perfect agreement between raters, 0 indicates agreement is no better than by chance. A Kappa score lower than zero indicates an agreement which is worse than that expected by chance.

C. Model Training and Hyper-Parameter Tuning

An initial fixed data split was utilised for the training and evaluation of the performance of the CNN model for HIE grading based on the Mel Spectrogram images of the EEG signal. The data was initially split 80/20, with 20% used as the unseen test set and the remaining 80% was further divided in an 80/20 split to define the training and validation datasets. A patient independent model was ensured by splitting the data with groups defined by the patient ID, guaranteeing all data epochs for a patient are kept in the same dataset whether it be training, validation or test. A stratified split was then implemented to maintain the distribution of the overall dataset in the training, validation and test sets, and so the model is reflective of the true data. From here, the model was trained and evaluated based on the training and validation data and the best model, selected based on the validation loss, was tested on the unseen test data.

*1) **Model Architecture**:* The ConvNext CNN architecture backbone [70] is a well known model with state-of-the-art performance for both image classification and computer vision tasks. It is trained in this study for HIE grading based on the Mel Spectrogram input image of the EEG signal. The pretrained models were used from Timms library [71], with the model then being fine-tuned by retraining on the Mel Spectrogram images of all four HIE grades. Model hyper-parameter fine tuning was conducted to optimise the model for the HIE grading task, with the final model architecture being

defined as the ConvNext Nano backbone with MSE loss and Adam optimizer with weight decay of $1 \times e^{-4}$. The model was trained for 80 epochs with a batch size of 32 and a learning rate of 1×10^{-4}. In order to identify this optimal model architecture, a series of tests were conducted to determine the optimal hyper-parameter settings for the model.

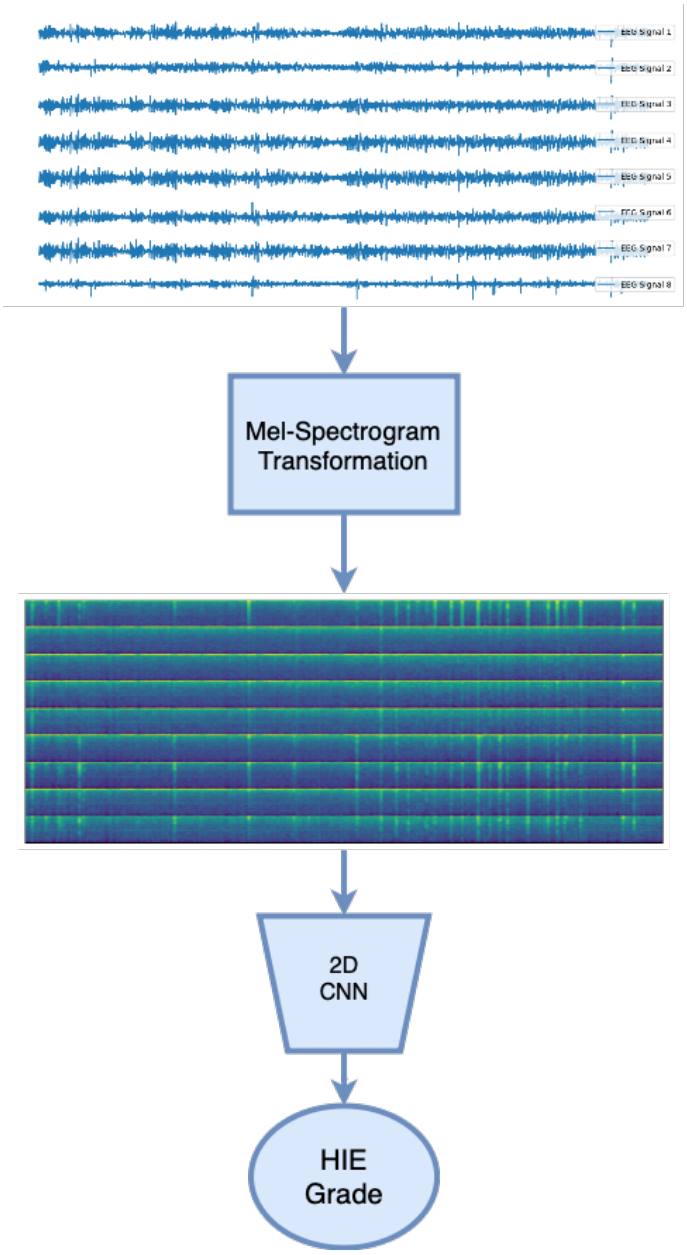


Fig. 9. Baseline 2D CNN Model Architecture using Mel Spectrogram representation of EEG signal.

Initial testing of the appropriate model size for the task was assessed by running training of the CNN for a variety of different ConvNext backbones, with the training, validation and test accuracy results shown for each experiment in Fig. 10. All other parameters of the model configuration and pre-processing phase remained unchanged

throughout this testing. The ConvNext backbone architectures evaluated are:

- ConvNext Atto;
- ConvNext Tiny;
- ConvNext Nano;
- ConvNext Small;
- ConvNext Base.

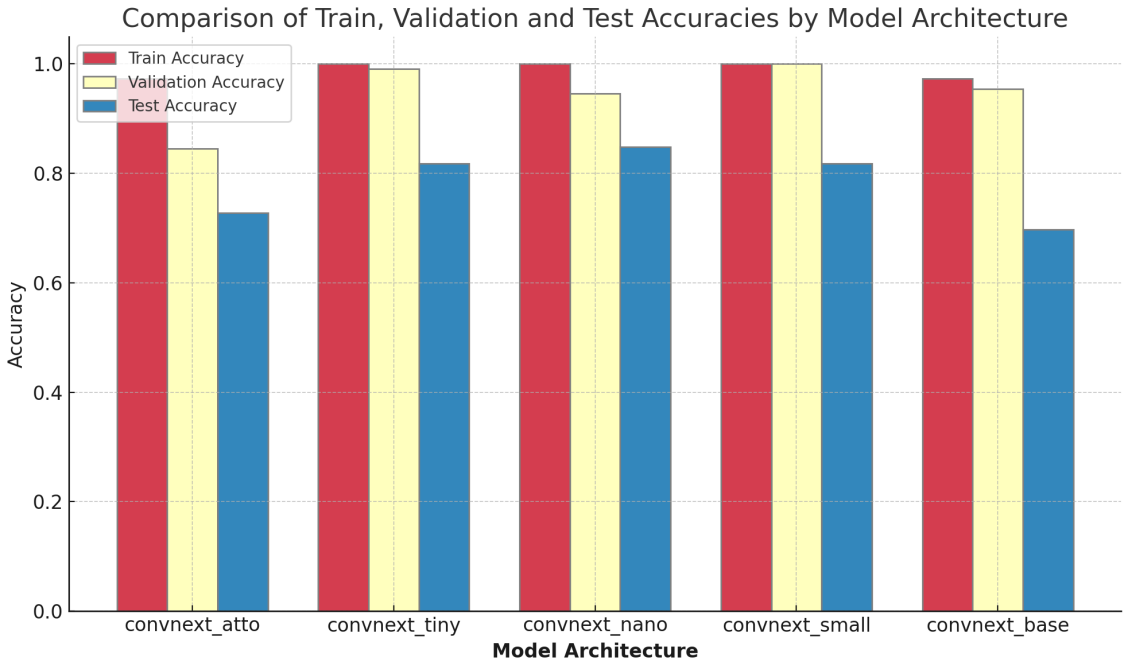


Fig. 10. Accuracy Results for Each ConvNext Model Architecture to Evaluate the Effect of the Model Size on the Accuracy Performance.

Comparing the results in Fig. 10 it is evident that the model which generalises best to the unseen test data is Convnext Nano, where a test accuracy of 84.85% was achieved. However, there is still a significant drop in accuracy between the validation and test data, indicating the model is potentially overfitting to the training data. The other backbone architectures show a larger difference in accuracy between the validation and test data, which indicates that these models generalise poorly to the unseen test data.

*2) **Data Augmentation**:* Data augmentation based on the Albumentations library [72] was implemented in an attempt to decrease the gap between the training and test accuracy and, thus, develop a more robust HIE grading model. Data augmentation was tested to assess whether the model was overfitting to the training data. Overfitting is a possibility due to the complexity of the selected CNN and the limited size of the

available dataset. The goal of data augmentation was to then slow down the model's learning by increasing the complexity of the training data. This helps the model to learn more relevant features for HIE grading and, thus, perform better on the unseen test data. The data augmentation scheme was applied to the Mel Spectrogram which resulted in the best model performance being a combination of the following transformations:

- Random Horizontal Flip;
- Random Vertical Flip;
- Channel Masking.

A sample spectrogram image with the data augmentation scheme applied is shown in Fig. 11.

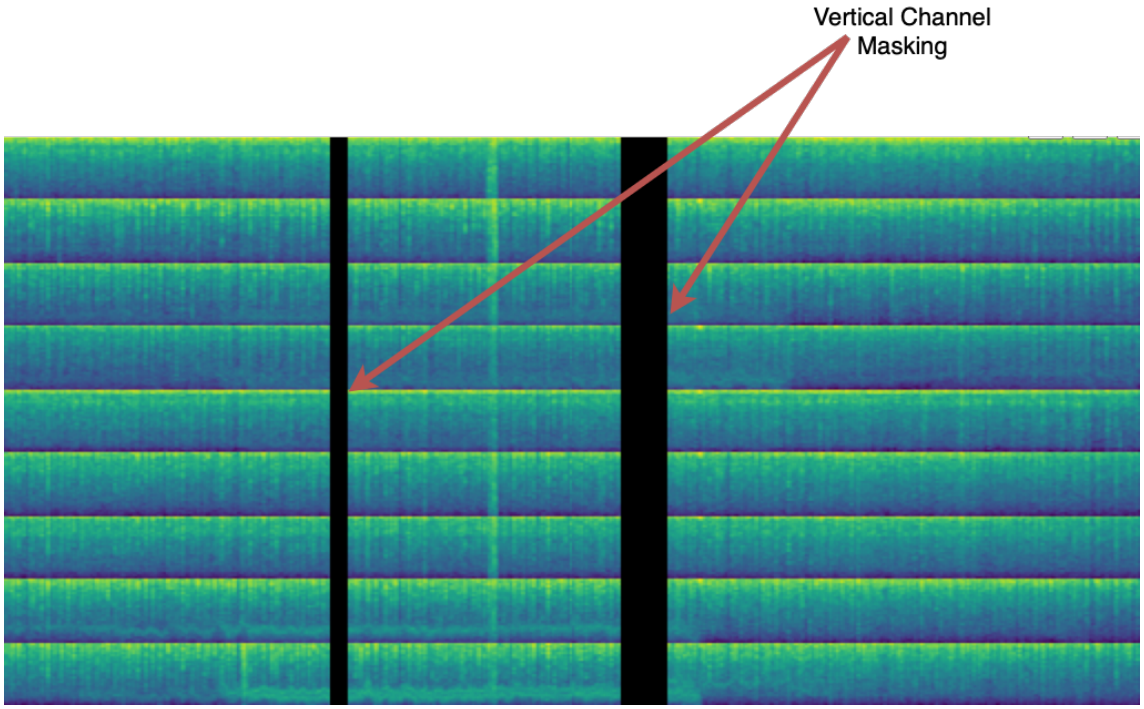


Fig. 11. Mel Spectrogram representation of EEG signal with the data augmentation applied. The two black columns seen on the figure are the masked areas in the vertical direction.

Each augmentation was applied to the training data with a probability of 10% and so not all samples would be augmented in each epoch, and the samples with augmentation are randomly selected. It was important to find an appropriate balance in the amount of data augmentation applied so that the model did not learn the noise of the training images due to the data augmentation applied. The gap between the training and validation loss decreased due to the data augmentation scheme applied and is within a satisfactory range for the model to be evaluated on the unseen test data. The difference between

training and validation loss was less than 0.1. The unseen test data remained fixed at an accuracy of 84.85% and, so, while the test accuracy did not improve due to the data augmentation, the generalisability was still better, based on evaluation of the loss function. Further investigation into the model performance is provided in the next section.

D. Results and Discussion

The proposed method achieves an accuracy of 84.85% on an unseen test set, with the performance metrics of the model being detailed in Table 3.4. A good balanced performance is seen across all evaluated metrics, showing the model is clearly learning meaningful features with which to classify the four HIE grades and can generalise well to new unseen data.

TABLE III

MODEL PERFORMANCE METRICS

Metric	**Value**
Precision	0.7858
Recall	0.7576
F1-Score	0.7547
Kappa Score	0.5622
R-squared	0.7574

The sensitivity of the model states that the model correctly identifies around 75.76% of true positive cases, meaning that 24.24% of positive cases are missed and so there is the need for potential fine-tuning of this model, or an increase in dataset size for the model to learn deeper features for each grade. With the F1-score of 0.7547, it is clear that the model demonstrates a good balance between the precision and recall however there is the potential for further optimisation to reach a higher overall performance. The Kappa score is quite low, at a value of 0.5622. While this shows the model performs substantially better than random guessing, there is still significant room for improvement. This score implies that class imbalances might be affecting the model's ability to correctly differentiate between categories. This moderate Kappa score suggests the need for further development of the method and workflow for HIE grading using 2D spectrogram based CNN. However, it is a very strong baseline developed to assess

the performance of 2D CNNs for HIE grading. Further development of this method lead to the novel approach detailed in Chapter 4, where significant enhancement to the pre-processing, deep learning model and post-processing phases of the method is performed.

E. Conclusion

The 2D CNN model showed promising performance for the implementation of autonomous HIE grading based on spectrogram representation of the EEG signal, achieving a test accuracy of 84.85%. This is 1.44% below the current state-of-the-art for HIE grading published by Yu et al [24] on a large unseen dataset which contained 338 hours of patient data [24], double the size of the publicly available dataset used in this study. It is 4.56% below the leave-one-out test accuracy proposed by Raurale et al, which used a similar approach of representing the EEG signal in the time-frequency domain although Raurale et al used much smaller input windows [31]. Ruarale et al's work used an input window of 5 minutes EEG signal [31], the work detailed in this study, however, focused on using the full 60 minute window of the signal as a single input to the CNN. The same model proposed by Ruarale et al achieved a test accuracy of 69.5% [31] on the same private dataset used by Yu et al [24] and, so, the method proposed here shows a relative increase in performance of 22.07%.

This work is a foundational baseline method for HIE grading and highlights the potential of 2D CNNs for autonomous grading tasks, as it shows comparable results to state-of-the-art methods proposed in the literature to date. To surpass the performance of these methods, however, further development of the signal representation and model architecture are required, as is done in Chapter 4 of this thesis. The findings of the background research to this thesis showed the importance of amplitude and frequency in distinguishing the HIE grade, and this was supported by the high test accuracy for the Mel spectrogram representation classified by the 2D CNN. The rhythmic behaviour of the signal is highlighted but not given sufficient emphasis for differentiating the four grades. The Mel spectrogram representation and the initial background research showed the relevance of detecting the pattern of activity and inactivity for seperability between the four grades. This lead to the desire for enhancement of the pre-processing phase to enhance all three metrics of the signal and thus the goal is to improve the overall

classification performance. The next chapter builds on this, proposing a novel method for HIE grading and detailing an in depth analysis of the models performance.

4. From EEG to Audio: A Novel Spectrogram-Based Approach for HIE Grading

This chapter builds on the foundational research from Chapter 3, where 2D CNNs demonstrated promise for HIE grading. Informed by the importance of rhythmic behavior and amplitude patterns identified in our background research and intial feature analysis, we introduce a novel spectrogram-based approach for HIE grading. The proposed method employs innovative signal processing techniques to translate the EEG data into the audio domain, enhancing frequency and amplitude features critical for distinguishing HIE grades. Sonification techniques have shown potential for aiding clinical diagnosis in the field of seizure detection, where the rhythmic behaviour of the EEG signal is enhanced through the conversion to the audio domain. By focusing on amplitude and frequency modulation of the EEG signal, long term patterns of the signals activity and continuity are highlighted. The goal being to emphasize the differentiation between the four HIE grades, with these patterns of continuity and discontinuity being one of the primary visual metrics for grading by a clinician. State-of-the-art CNNs are then used for grading, with novel techniques used to optimise the model for the specific task at hand.

This chapter covers all aspects of the proposed method, providing a technical breakdown of the three main phases in the machine learning pipeline:

- **Pre-processing phase**:
  - Conversion of the EEG signal into the audio domain using a sonification algorithm.
  - Generation of spectrogram images representing the signal in the time and frequency domains.
  - Generation of a single stacked image representing spectrograms of all eight channels of the EEG signal.
- **Modelling phase**:
  - Hyperparameter tuning and optimization of the 2D neural network.
  - Configuration of the final model architecture, containing regressor head, mixed precision operations, Stochastic Weight Averaging and the ConvNextV1 Nano backbone architecture.

- **Post-processing phase**:
  - Conversion of the CNN regression output into a discrete clinical grade using a post-processing algorithm.
  - Discussion of the algorithm for translating continuous predictions into clinical categories.

Finally, the chapter discusses the results and performance of the proposed method, including the rigorous training scheme implemented to achieve a robust grading system.

A. *Proposed Method*

An overview of the workflow for the proposed method is shown in Fig. 12, where the pre-processing, deep learning network and the post-processing steps are clearly outlined.

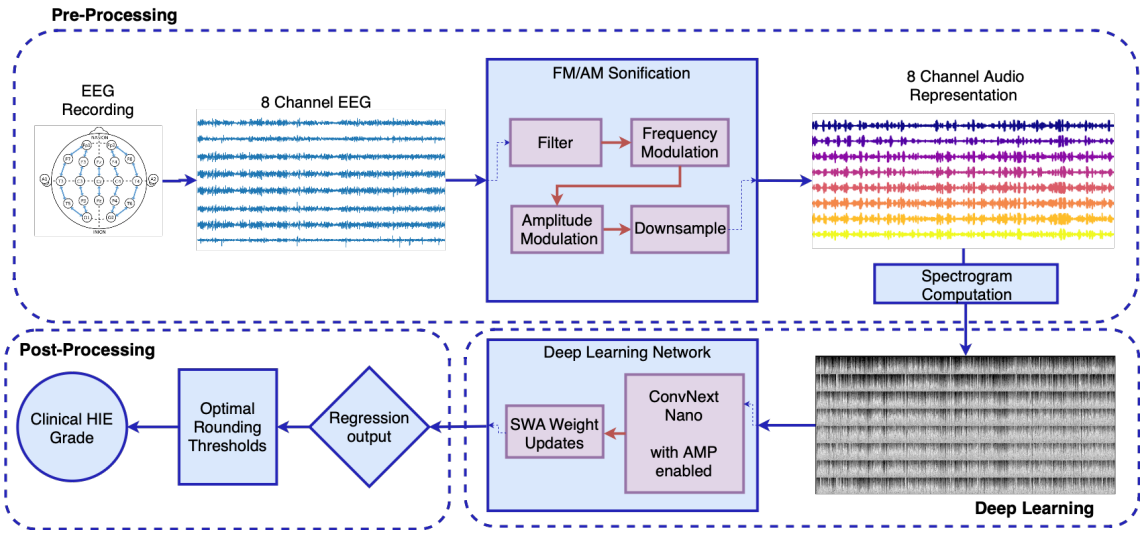


Fig. 12. Proposed Method Workflow

*1) Pre-Processing using **Sonification**:* Sonification is a signal processing technique which has been shown to provide useful benefits in seizure detection by processing the information in a novel way, complementary to the standard visual inspection of EEG signals [73], as previously detailed in Chapter 2. However, no publications to date have explored the use of sonification in the analysis of background EEG signal grading. This chapter focuses on the conversion of the EEG signal to the audio domain using sonification, then utilising this novel data representation to train and evaluate deep learning networks for HIE grading. In this study, the EEG signal was converted to the audio domain based on an adaptation of the FM/AM sonification algorithm,

first proposed by Gomez et al [41], with the goal to compress the EEG signal in the time domain while also highlighting the rhythmic behaviour of the EEG signal. The objective is to enhance the features of the EEG signal most relevant to each HIE grade. The FM/AM sonification algorithm was first proposed for the work on seizure detection in neonatal patients, where the seizure was distinguishable by the human ear through frequency and amplitude changes in time. Surveys were conducted with a number of both medical and non-medical professionals to determine the advantage of using sonified signals in detecting seizures in comparison to visual inspection of the EEG signal [41]. The FM/AM sonification holds a particular advantage for identifying patterns in the background of the EEG signal over long windows in time, in comparison to visual inspection of the EEG signal. This is due to the FM and AM synthesis which varies a carrier wave frequency and amplitude proportionally to a modulating wave. As can be seen in Fig. 13, the waveform amplitudes of the EEG signal become the spectra of the audio signal and, so, any structure in the signal will be perceived as a rhythm in the resulting audio. This is important as the rhythmic pattern of activity and inactivity, along with the amplitude of this signal, are known features of interest for the clinician when diagnosing the severity of HIE.

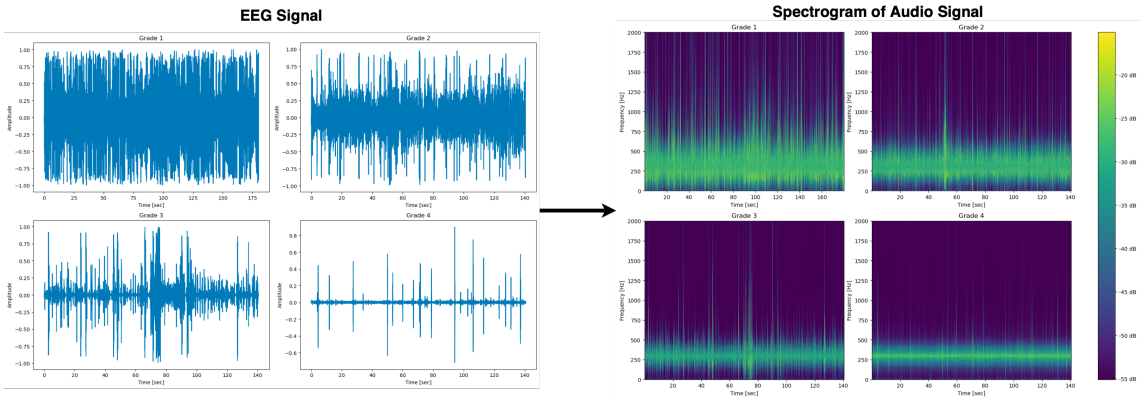


Fig. 13. Comparison of EEG signal with spectrogram of audio signal after FM/AM transformation.

An overview of the FM/AM Sonification algorithm is provided in Fig. 14, detailing the key phases of the algorithm discussed below.

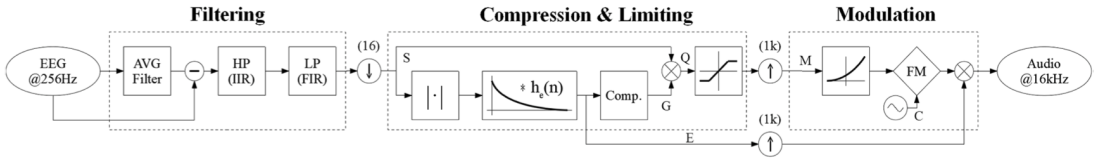


Fig. 14. FM/AM Sonification Algorithm Overview [41].

The FM/AM sonification technique begins with preprocessing and filtering the EEG signal. A notch filter is applied to remove any noise present in the signal due to the power supply of the monitoring device. Standard EEG filtering is applied to filter the signal between 0.5 to 7.5 Hz, downsampling the signal then to 16 Hz in order to decrease the computational load for the remaining signal processing [41]. The signal is then compressed to reduce the dynamic range of the EEG signal to prevent any amplitude spikes in the EEG signal from distorting the Frequency Modulation (FM) stage. The envelope $E[n]$ of the EEG signal is computed by convolving the absolute value of the signal with an exponential impulse response. This creates a smooth envelope which captures the energy of the signal. A compression ratio, r, is applied to the envelope to reduce the dynamic range of the signal when the amplitude exceeds a threshold T. The envelope is calculated as per Equation 14, and is later used to modulate the amplitude of the signal.

$$E'[n] = \begin{cases} E[n] & \text{if } E[n] \leq T \\ T + \frac{E[n]-T}{r} & \text{if } E[n] > T \end{cases} \tag{14}$$

The FM step is a process which effectively changes the tone of an otherwise fixed carrier signal, and the EEG signal is used as the modulating wave. The frequencies of this signal are then passed through an exponential transform. The relationship is defined in Equation 15 transforms the frequencies to semitones. This exponential transform is applied to follow the musical definition of an octave.

$$f[n] = 500 \cdot 10^{V[n]} \tag{15}$$

where $f[n]$ is the frequency of the modulated signal at time step n and $V[n]$ is the EEG signal at time step n.

The FM signal is then amplitude modulated (AM) to translate the long-term amplitude of the EEG signal into the audio amplitude. The FM signal is modulated based on the envelope of the EEG signal, $E[n]$, previously calculated. This step is crucial in emphasising the long-term rhythmic behaviour of the EEG signal, a feature noted as important in grading the severity of HIE in the EEG signal.

The EEG signal is compressed in the time domain by the FM/AM sonification algorithm, an advantage for rapid reviewing of the EEG data. This is also useful here since a reduced window size covers the information of the full EEG signal. The time compression is applied to the algorithm by adjusting the down-sampling factor of the signal, thus increasing the playback speed of the output audio signal. For example, a speed of 10 will linearly reduce the audio signal in the time domain by a factor of 10, and so a 60 minute window of EEG is represented as 6 minutes in the audio domain.

A sample of the audio domain signals for all four HIE grades, with a time compression of 20, is shown in Fig. 15. Identification of each grade is possible from Fig. 15, based on visual inspection of the activity levels of each signal. In addition to the variation in activity levels, the discontinuity patterns of the more severe HIE grades are also noted, based on visual inspection.

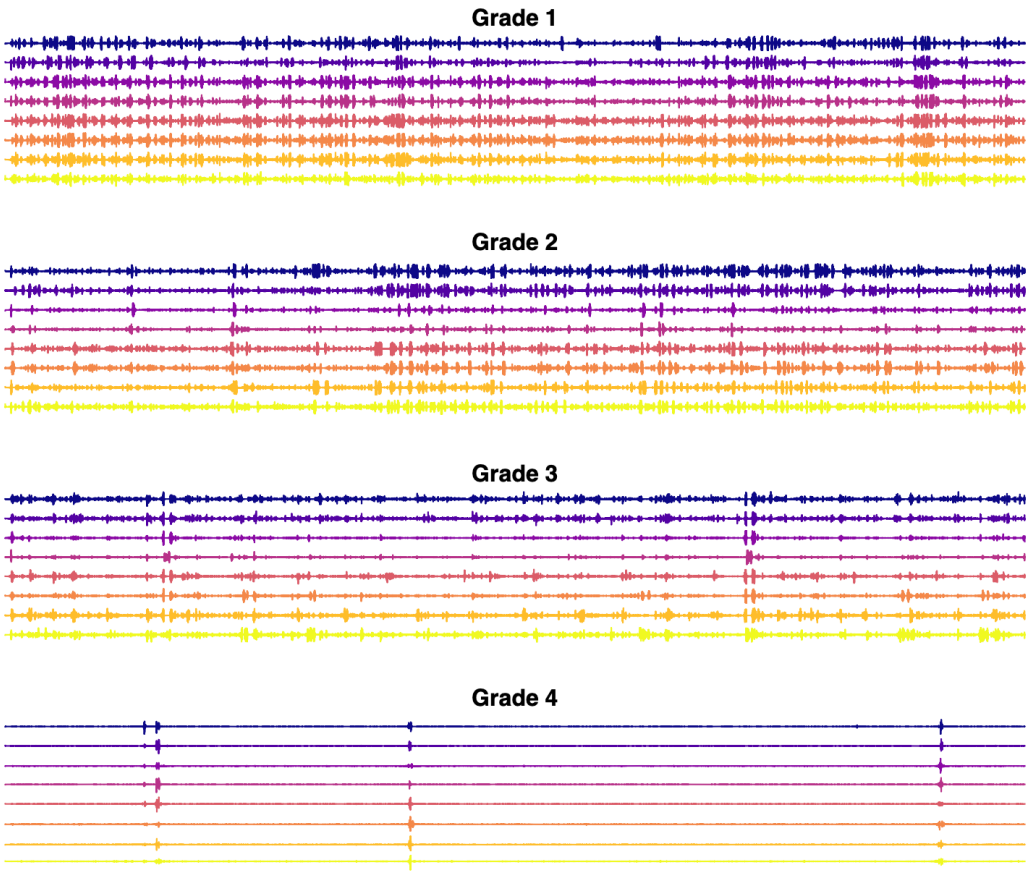


Fig. 15. A 30 second window of an audio domain signal for all four HIE grades, highlighting the rhythmic behaviour of each grade after the sonification conversion process.

The audio signal was downsampled from 8 kHz to 512 Hz to reduce data size while preserving the signal's integrity. The resampling process was performed using the Scipy Python library function [74], which relies on Fourier Transform-based interpolation. This method models the signal as a sum of sinusoids, enabling accurate resampling without introducing significant artifacts. A trade-off of downsampling is the reduction in temporal resolution. However, since the final goal is to generate a spectrogram of the signal, reducing the temporal resolution allows us to allocate more resolution to the frequency domain in the spectrogram, preserving critical information about the signal's frequency content.

A plot of the original signal for one patient sample is shown in Fig. 16, along with the downsampled signal. This figure highlights the limited change in the audio representation of the EEG signal due to downsampling. The most prominent difference is a decrease in the amplitude of the downsampled signal, identified from the overlayed signals in subplot 3 of Fig. 16.

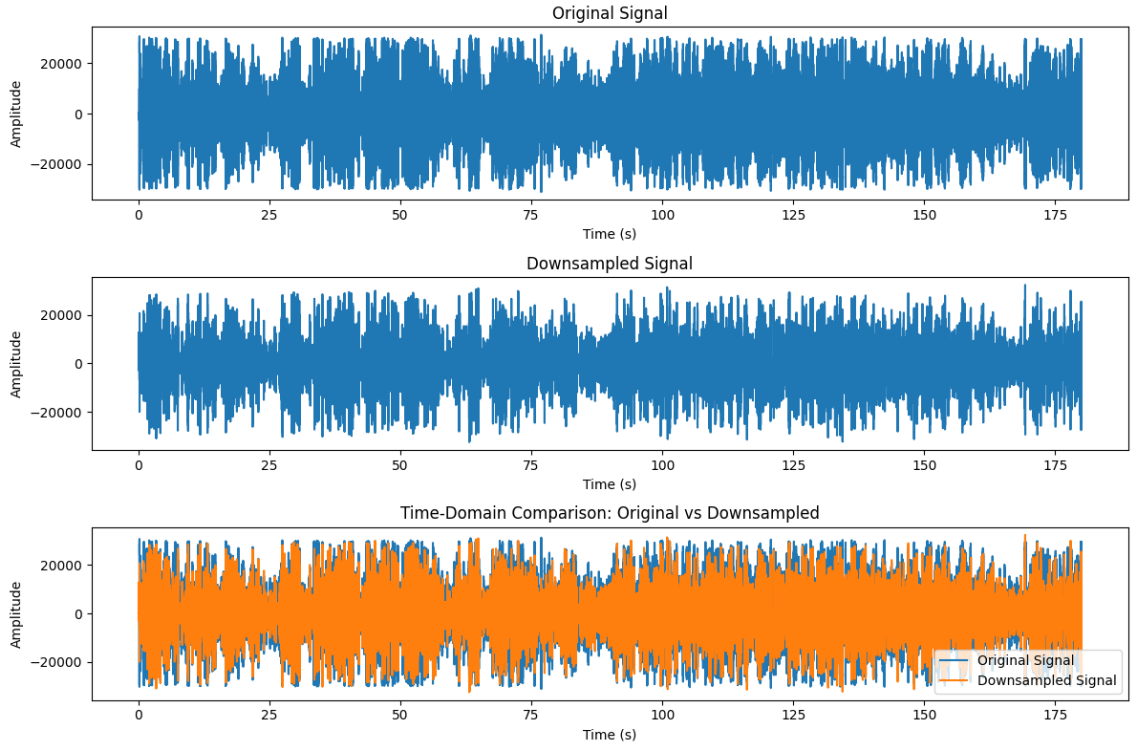


Fig. 16. Comparison of the Original Audio Signal for a sample EEG with the Downsampled Audio signal in subplot 2. The original signal and downsampled signal are overlayed for direct comparison in subplot 3 of this figure.

While the time domain representation of the audio signal showed no significant changes, due to downsampling, outside of an accepted tolerance range, it was vital to ensure the downsampling satisfies Shannon-Nyquist Sampling Theorem [75]. The results of the maximum frequency of the original signal and the Nyquist rate of the downsampled signal are shown in Table IV.

TABLE IV

COMPARISON OF ORIGINAL AND DOWNSAMPLED SIGNALS

Metric	**Value**
Maximum Frequency in Original Signal	169.28 Hz
Nyquist Rate of Downsampled Signal	256.00 Hz
Satisfies Shannon-Nyquist Theorem	Yes

Given that the maximum frequency of the original signal is 169.28 Hz, the improvement in frequency resolution due to downsampling is significant for the spectrogram computation. The frequency resolution of the spectrogram for the original signal is much coarser, potentially missing important nuances in the frequency representation that are essential for accurately grading the signal.

The spectrogram representation of the EEG signal is performed to generate the 2D image for input to the CNN. Spectrograms are used widespread in audio signal analysis, converting the time domain signal into a visual representation in the time-frequency domain. This conversion is performed based on the computation of a series of short-time Fourier Transforms (STFTs) [76]. STFT is useful in cases where the goal is to extract the narrow-bandwidth information from a non-stationary signal [77], and converts the long signal into smaller overlapping segments, prior to computing the Fourier Transform (FT) for each segment. This is true for the audio domain representation of the EEG signal. The STFT of a signal $x[n]$ is defined as per Equation 16.

$$\text{STFT}\{x[n]\}(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{16}$$

where $x[n]$ represents the input signal, $w[n-m]$ is the window function applied to the segment of the signal, ω denotes the angular frequency, and m refers to the time shift or frame index.

A single frame of the spectrogram is generated by each STFT calculation for the segmented portion of the original audio signal. The squared magnitude of the STFT then computes the spectrogram of the signal, as described in Equation 17. This results in a two-dimensional image, where the x-axis represents time, the y-axis represents frequency and the colour intensity represents the magnitude of the signal at that time and frequency [76].

$$S(m,\omega) = |\text{STFT}\{x[n]\}(m,\omega)|^2 \tag{17}$$

The number of Fourier transforms per window, the overlap and the window length are three highly influential parameters in determining the resolution of the spectrogram in both the time and frequency domains. Details of the effects of these three parameters on the differentiability of the signals for the four HIE grades is evaluated in the discussion of this chapter, where a grid search of the spectrogram parameters is performed and the accuracy of the resultant CNN model is evaluated. Discussion as to the impact of each feature on the resultant spectrogram image is provided here.

- **Number of FFT Points:** The frequency resolution is controlled by the number

of FFT points, or the size of the FFT window, with the frequency resolution Δf defined as:

$$\Delta f = \frac{f_s}{N} \tag{18}$$

where Δf is the frequency resolution of the spectrogram, N is the number of FFT points and f_s is the sampling frequency of the audio signal. With an increase in the number of FFT points, a finer distinction between frequencies is achieved, thus resulting in an improved frequency resolution. However, a drawback to this is that the time resolution will suffer as the FFT window covers a larger number of samples. The proposed model uses 128 points per frame and results in a frequency resolution of 4 Hz per frame.

- **Hop Length:** The hop length determines the overlap of the window, defining the size of the step between any two adjacent windows. The time resolution is dependent on the hop length, with a smaller hop length resulting in a larger overlap between windows and, so, a finer time resolution is achieved. This relationship is shown in Equation 19.

$$\Delta t = \frac{\text{hop_length}}{f_s} \tag{19}$$

where Δt is the time resolution and f_s is the sampling frequency of the audio signal. It is important to note that the better the time resolution the greater the increase in computational costs of the spectrogram as the number of windows increases. A time resolution of 125 milliseconds is achieved in the proposed method, with a hop length of 64 defined.

- **Window Length:** The amount of the signal considered at any one time is determined by the window length. A window function, the Hamming window in the case of this study, is applied to each window of the signal to reduce the impact of edge effects and a window length of 128 is defined for the proposed method. Both the time and frequency resolution are again dependent on the window length, with a larger window length increasing the frequency resolution while also decreasing the time resolution. The window length can be computed based on Equation 20

with α and β chosen to reduce the largest side lobe [78].

$$w[n] = \alpha - \beta \cos(\frac{2\pi n}{N}) \tag{20}$$

where $w[n]$ is the window function, N is the number of samples in the window, and α and β are the window parameters.

An example of the audio signal being represented as a spectrogram image, with the optimal spectrogram parameters being a window length/shift of 128/64, is shown in Fig. 17. Based on visual inspection of the four plots in Fig. 17, the key characteristics that allow for differentiation across the grades are visible, where the rhythmic pattern, along with the amplitude level of the signal can be noted. Comparing these images to the Mel spectrogram representation of the EEG signal, in Fig. 8, the discriminatory features between the grades are clearly easier to identify in the FM/AM transformed spectrogram of Fig. 17. Particularly, the differentiation between grade 2 and grade 3 is easier to identify.

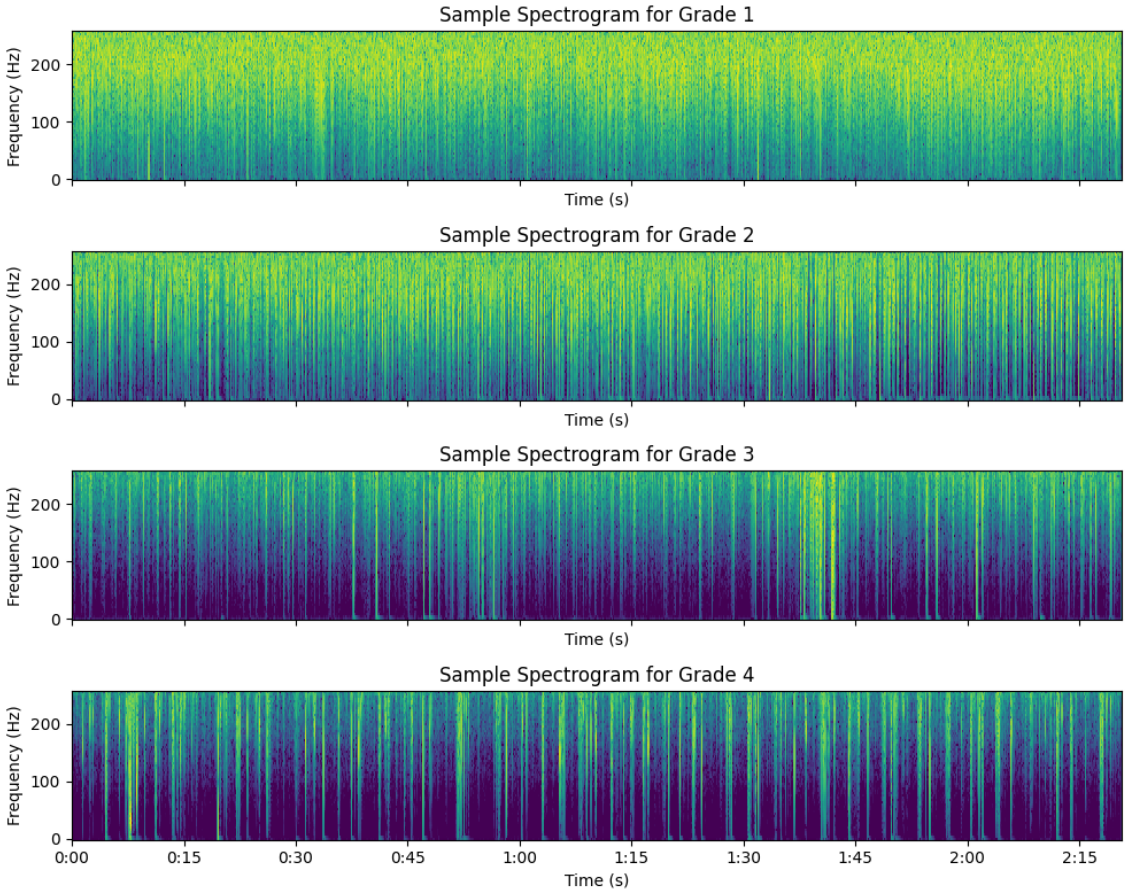


Fig. 17. Spectrogram representation of all four HIE grades based on the optimal spectrogram parameters obtained experimentaly for the audio domain representation of the EEG signal.

As shown in Fig. 17, the audio signals for each grade of EEG are represented as spectrograms, capturing the distinct features that allow for effective differentiation of the signals. Each spectrogram provides a visual pattern based on frequency and time, highlighting the unique characteristics of each grade.

- **Grade 1** exhibits a uniformly distributed power across the frequency range, with higher intensity in the upper frequencies.
- **Grade 2** shows a similar distribution, though with a slight reduction in the intensity of upper frequencies, indicating early disruptions in the EEG signal which may correspond to mild abnormalities, such as patterns of discontinuity, as expected.
- **Grade 3** introduces more variability over time, with clear fluctuations in the lower frequencies. This suggests more significant disruptions, aligning with the definition of grade 3 HIE, as in [46].
- **Grade 4** presents the most prominent changes, with concentrated energy in the lower frequencies and attenuation in the upper frequencies. This pattern reflects

the heavily disrupted nature of the EEG signals, where grade 4 is expected to be a primarily flat signal.

The differentiating features, particularly in the energy distribution across frequencies, provide a clear visual indicator of the severity of HIE. This highlights the utility of this FM/AM transformation in aiding the signal classification, as the improved differentiability of the signal on input to the CNN aids it in the classification process.

With the optimal spectrogram parameters identified for a single sonified EEG channel, all eight channels must then be concatenated at the spectrogram image level. The motivation for concatenating the spectrograms of all eight channels stems from the annotation of the dataset, where a single annotation is provided for all eight channels. For accurate interpretation of the CNN outputs, the input to the CNN must reflect this annotation metric. When a single channel spectrogram is used, late decision aggregation is applied to identify the average output across all eight channels. However, this is computationally demanding along with being time consuming as, for a single patient, inferening must run eight consecutive times. Therefore, a single image representing all eight signals is favourable. Analysis of various image concatenation techniques were explored, including:

1) **X-Dimension Concatenation:** The eight EEG spectrograms are vertically stacked to create a single image of shape $8H \times W$, where H is the height of a single spectrogram image and W is the width of a single spectrogram.
2) **Y-Dimension Concatenation:** The eight EEG spectrograms are horizontally stacked to create a single image of shape $H \times 8W$. The rectangular dimensions of this image are unusual for a CNN model, where most pretrained models are trained on square images; however, it was tested in this analysis, where the model performed well with the rectangular shape of the single-channel input image.
3) **Average Spectrogram Image:** The spectrograms of all eight EEG channels are computed, and then their average is taken to create a single input image of shape $H \times W$.
4) **Channel-Level Aggregation:** Each spectrogram image is a greyscale, single channel image, so the CNN backbone for all previous samples is set with input channel dimensions as 1. For channel-level aggregation, the spectrograms are stacked at the tensor channel dimension, resulting in an input tensor of shape

$[batch_size, 8, H, W]$, where each channel represents the greyscale spectrogram of a single EEG channel.

The final model used eight vertically stacked spectrograms, concatenating the spectrogram images in the x-dimension. A representation of this is shown in Fig. 18.

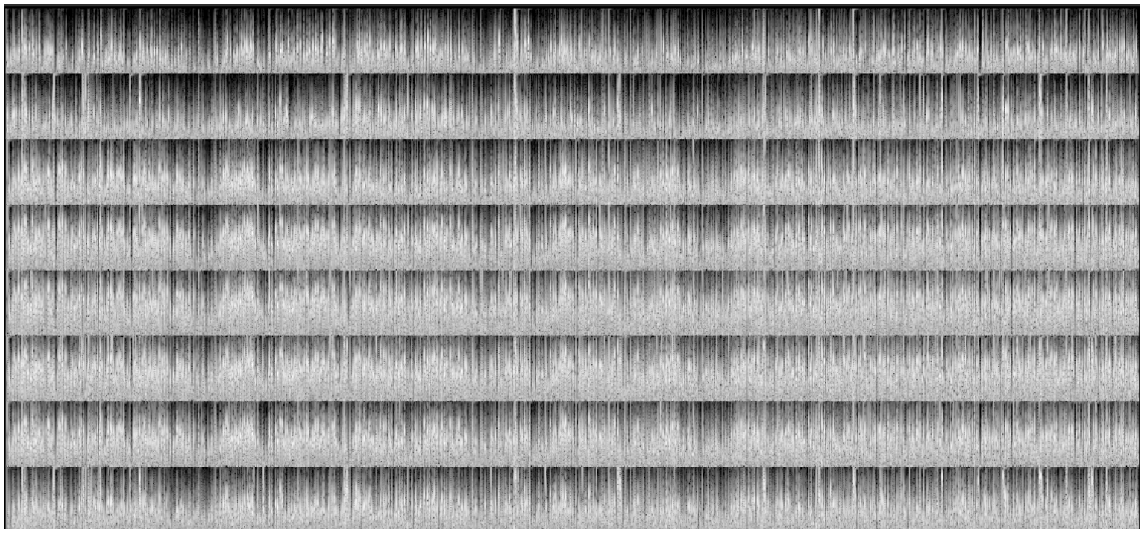


Fig. 18. Vertically Stacked Channel Spectrograms for the Sonified EEG Signal

In summary, the pre-processing phase of the proposed method leverages FM/AM Sonification to convert EEG signals into the audio domain, thereby emphasizing the rhythmic patterns critical for distinguishing between different HIE grades. The sonified signal is then represented as a spectrogram, converting the temporal and frequency characteristics of the EEG into a visual format. Through parameter optimization of the short-time Fourier Transform (STFT), a spectrogram configuration is identified that best highlights the features relevant to HIE grading. Finally, the input image to the CNN is eight vertically stacked spectrograms in the x-dimension which creates a robust input representation for deep learning.

*2) Deep Learning using **2D CNN**:* The ConvNext Nano backbone architecture pre-trained from the Timms library [71], was used as the base architecture, as was used in the baseline method proposed in Chapter 3. The model is trained on the spectrogram images depicting the audio domain signals. Significant research went into determining the optimal hyper-parameters and model configurations to achieve the best performance for the proposed method. The final configuration is detailed here, with findings supporting these hyper-parameter values provided in the subsequent discussion.

Since there is a continuous relationship between the HIE grades, this regression model was designed to capture this relationship. A linear layer is applied as the regressor after the global average pooling layer of the ConvNextV1 Nano architecture. Superior performance of the regression model was achieved over the classification header, where comparitive analysis showed lower training and validation losses and, thus, improved generalisability. The improved performance of the regression header aligns with expectations due to the nature of the grading task, where the relationship between the grades is monotonic. The features associated with each HIE grade exist on a spectrum, defined within a fixed range, thus it is important for the model to learn these relationships. A regression model thus captures this continuous variability, thereby allowing for more precise predictions that better align with the subtle gradations in severity observed in clinical practice.

A cosine Annealing learning rate scheduler [79] with warm restart was implemented after showing superior performance. By applying warm restarts to the learning rate during training, the model returned optimal performance by allowing the model to escape local minima and plateaux in the loss function. By applying warm restarts, the model weights were excited at periodic intervals without random initialization, this meant that the information the model had learnt to that point was not lost [79]. The model can then learn additional information as the loss function converges to the local minimum again, thus resulting in increased model learning and overall performance improvement. The cosine annealing learning rate scheduler used is defined as per Equation 21 [79].

$$\eta_t = \eta^i_{\text{min}} + \frac{1}{2}\left(\eta^i_{\text{max}} - \eta^i_{\text{min}}\right)\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_i}\pi\right)\right) \tag{21}$$

where η_t is the learning rate at epoch t, η^i_{min} is the minimum learning rate, η^i_{max} is the maximum learning rate, T_{cur} is the current epoch and T_i is the total number of epochs in the cycle.

In order to reduce the fluctuations of the model weights and improve generalisability of the model, Stochastic Weight Averaging (SWA) is applied [80]. SWA is based on the fundamental process of averaging the weights at different points in the training, while applying a cyclic learning curve to achieve the optimal performance without increasing the computational complexity of the model [80]. SWA is implemented with a Stochastic

Gradient Descent (SGD) optimiser in [80], however, it is used in conjunction with the Adam Optimizer in this study. SWA is applied once training has already reached peak performance, with the goal to stabilise the fluctuations in the model.

By employing Cosine Annealing with warm restarts, and SWA, the overall model generalisability and computational efficiency is enhanced. By leveraging these techniques, the regression model was able to capture the subtle changes between the HIE severities. Post-processing is essential as the next step, to align the continuous output predictions with clinically discrete HIE grades. The developed optimised rounding function ensures that the continuous predictions are effectively translated into clinically relevant categories, enhancing the overall model performance.

*3) Post-Processing using **Optimised Rounder**:* An optimised rounder class was implemented as a post-processing step in the machine learning pipeline for the proposed method to convert the continuous regression value of the model to a discrete clinical grade. This idea was first implemented in the PetFinder Kaggle competition [81]. Optimal rounding thresholds which best separate the continuous predictions into their respective classes are computed based on Nelder-Mead Optimization [82]. Nelder-Mead is a popular method for optimisation and is a minimisation algorithm which seeks to make the difference between the predicted value and the target value as small as possible [82]. The fundamental principles of the Nelder-Mead algorithm is that it is a derivative-free approach that utilises a heuristic search to evaluate the function at different points and move towards the optimal solution.

The optimised rounder is designed to identify a set of thresholds $t_1, t_2, ..., t_{C-1}$ such that the continuous predictions are separated into C discrete classes, where C is 4 in this task, due to the number of HIE grades. Fine-tuning the thresholds enhances the alignment between the predicted and true class, thus improving the model's overall performance.

Significant analysis was performed to design and developed this proposed method. Additional results and experiments conducted to support these designs are provided in this discussion. This starts with the analysis of the sonified EEG signal to understand the benefits of representing the EEG signal in the audio domain for HIE grading based on the FM/AM transformation.

B. Statistical Analysis Methodology of Sonified EEG Signal

In order to gain a deeper understanding of how the EEG signal behaves after the FM/AM conversion to the audio domain, statistical analysis of the signal, including standard deviation analysis, was performed. The standard deviation (SD) is a measure of scatter in a set of data points, and is used to quantify the amount of variation or dispersion of a set of values [83]. The SD, σ_x, of a signal $X(t)$ is calculated as per Equation 22 [84].

$$\sigma_X = \sqrt{\frac{1}{T}\int_0^T \left(X(t) - \mu_X\right)^2 dt} \tag{22}$$

where μ_X is the mean of the signal $X(t)$ and T is the total time of the signal.

The standard deviation was calculated from the point in the FM/AM algorithm where tone mapping occurs and prior to the generation of the cosine carrier signal and amplitude modulation. A representation of each HIE grade at this point in the sonification algorithm is provided in Fig. 19, where the differentiability of all four grades is possible through visual inspection, based on the differences in the signal amplitudes. This supports the motivation for selecting the standard deviation as the statistical analysis metric where, based on visual inspection, the fluctuation of the signal at this phase in the sonification process appears to be a strong differentiating factor between the four HIE grades.

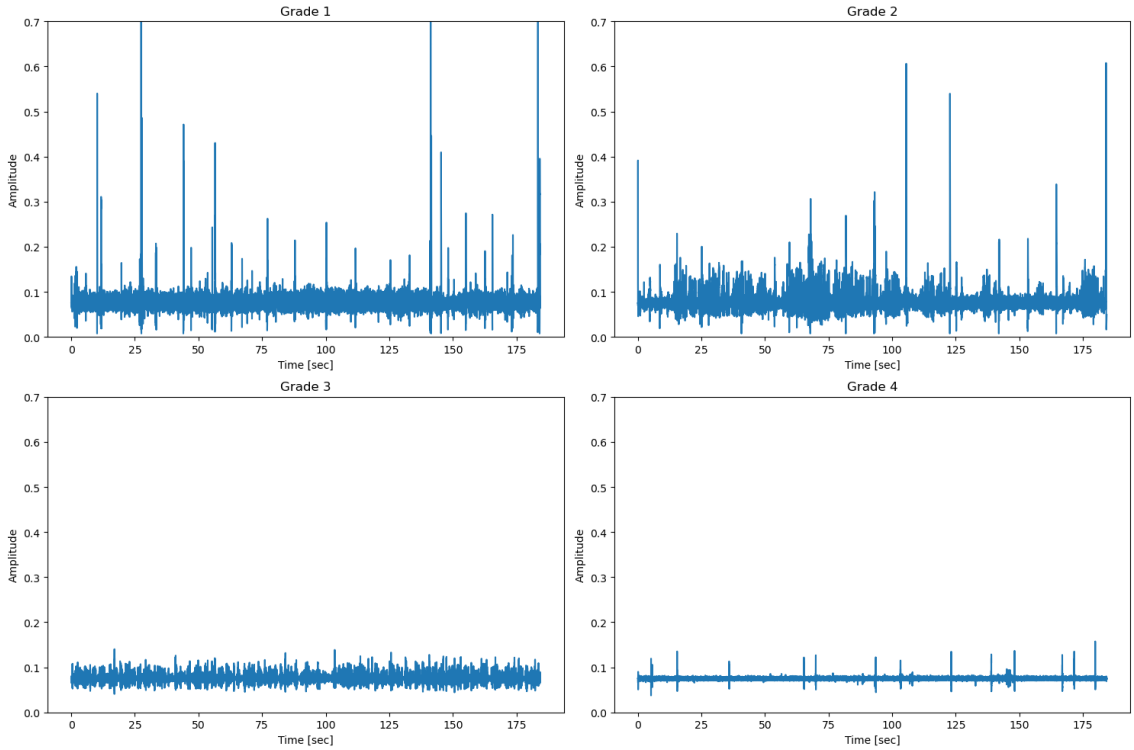


Fig. 19. Plot of the frequency mapping of all four HIE grades.

After calculating the standard deviation, a new optimised rounding function is trained and implemented to convert the standard deviation values to integer values representing each HIE grade. The technical details of the functionality behind this rounding function are presented in the post-processing section of the proposed method in this Chapter. This explains how the function identifies the optimal rounding thresholds for a set of training data based on their label and a given data metric, in this case the standard deviation value of the transformed signal. These rounding thresholds can then be applied to a test dataset to grade the audio signal based on the standard deviation value of the signal. A test accuracy of 66.4% was achieved, which is greater than a random guess, although not significant enough to be used as a standalone diagnostic tool. Evaluating the standard deviation of the frequencies in comparison to their true HIE grade in the scatter plot in Fig. 20 shows an overlap between the SD values for each grade.

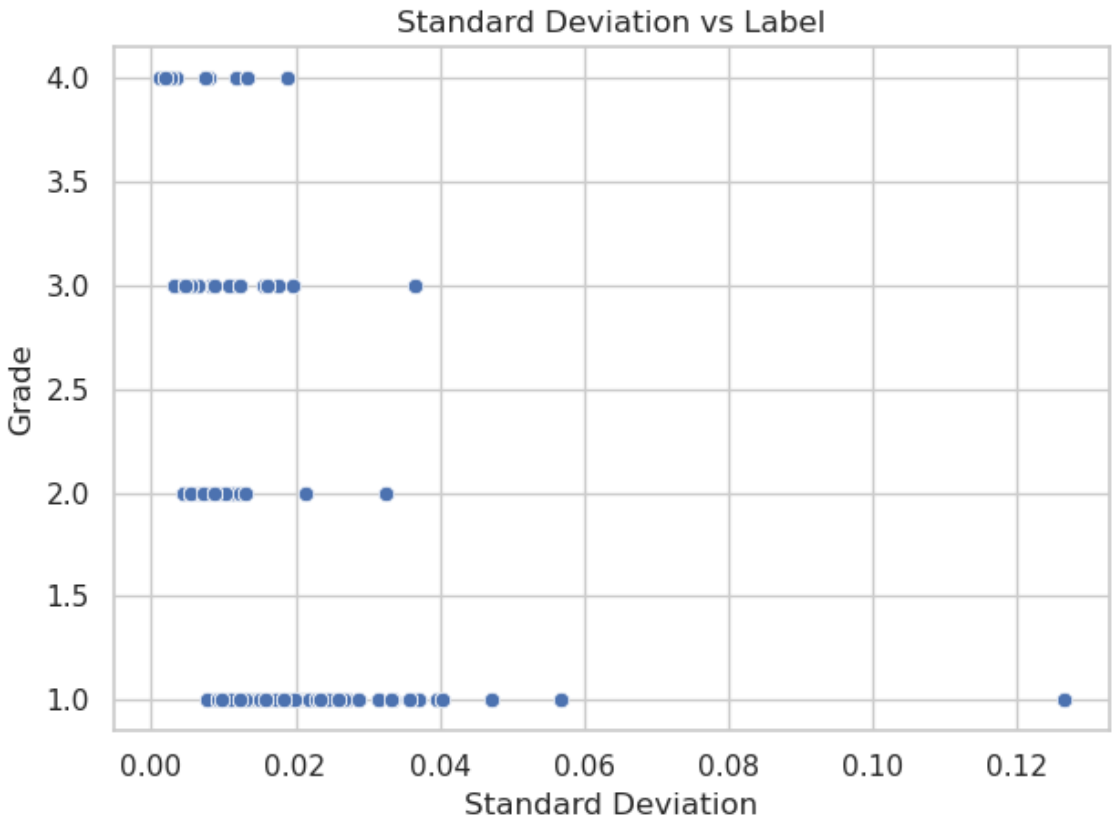


Fig. 20. Scatter plot of the standard deviation of the frequencies in comparison to their true HIE grade.

The frequency mappings plotted for the four HIE grades in Fig. 19 show a clear differentiability. However, these four samples were selected for demonstration purposes. When using the full dataset of 169 epochs, it is clear that there is a lot of fluctuation across the standard deviation values of each grade, as seen in Fig. 20. There is significant overlap in the standard deviation values of the four grades, which is likely due to the nature the HIE grades, which are on a graded scale. Classification into four discrete classes needs to reflect these relationships between the various features and so, individually, the standard deviation alone is not a strong enough indicator if the HIE grade present. In addition, as is true with real world data, some samples may lie on the decision boundary between grades and so additional knowledge is also necessary to determine these grades. Despite not resulting in a high test accuracy, information and increased understanding of the signal's parameters after FM/AM transformation can be achieved from this statistical analysis.

The correlation coefficient of the standard deviation and grade shows a moderate negative correlation of -0.527. This is a significant result as it shows that the standard deviation of the frequencies is a differentiating factor between the four HIE grades,

when represented in the audio domain. This is due to the FM/AM transformation of the signal emphasising both the pattern of activity and the amplitude of this pattern while also reducing insignificant fluctuations in the signal. As a result, an increased differentiability of the signals across the four grades is achieved when represented in the audio domain. The negative correlation is expected due to the nature of HIE grades where an increase in severity, and so a higher grade value, means lower brain activity and so a lower standard deviation of the frequencies is expected. This analysis highlights the challenge associated with the primary task and focus of this thesis to differentiate the HIE grade present in the EEG signal. The demand for a more complex and robust method to diagnose HIE based on the EEG signal is thus evident, thereby supporting the use of CNN deep learning architecture in the proposed method.

Given the demand for a deeper analysis of the sonified signal for accurate grading, analysis of the audio domain features was conducted, where the main focus looked at Mel-Frequency Cepstral Coefficients (MFCC) of the sonified signal. MFCCs have been used across a wide range of applications to date including gender recognition, voice recognition and speaker recognition [85] [86] [87]. The MFCCs of the signal were computed in this study using the Librosa library [88], with a technical explanation of MFCC computation being detailed in the following five steps:

- **Pre-emphasis**

  Amplification of the high frequencies of the signal to balance the frequency spectrum and compensate for any compression that might occur during the signal acquisition [89].

- **Signal Framing and Windowing**

  The signal's short-time stationary properties are captured by dividing the signal into short frames, usually 20 ms, with a 50% overlap between frames. The high overlap percentage allows the temporal characteristics of the signal to be captured and tracked [89]. In order to reduce spectral leakage, a window function is applied to the signal, with the Hamming window used in this study [90].

- **Power Spectrum**

  The frequency spectrum of each frame is computed based on the Discrete Fourier Transform (DFT), and the power spectrum is then calculated based on the result of this DFT [91].

- **Mel-Frequency Wrapping**

  The Mel-Frequency bank, discussed in the previous section, is applied to the power spectrum to map the frequencies to the Mel scale, in order to mimic the human auditory system's response to different frequencies [89].

- **Discrete Cosine Transform**

  The Discrete Cosine Transform (DCT) is applied to the Mel filter bank and is computed based on Equation 23.

$$X(k) = \sum_{n=0}^{N-1} x_n \cdot \cos\left(\frac{2\pi jnk}{N}\right), \quad k = 1 \tag{23}$$

  where $X(k)$ is the DCT coefficient at index k, x_n is the Mel filter bank output at index n and N is the number of Mel filter banks [92].

A simple random forest classifier was trained on the MFCC features of the sonified EEG signal to evaluate their potential for grading the severity of the background EEG signal, with best performance showing a test accuracy of 80.8%. Examination as to the benefits of representing the EEG signal in the audio domain are further assessed through this analysis. The number of Mel filter banks used has a significant effect on the MFCC analysis of the signal and, so, initial hyper-parameter tuning looked to identify the optimal number of Mel-filter banks to use prior to training the Random Forest classifier, with the results from 1 to 200 Mel-filter banks being shown in Fig. 21.

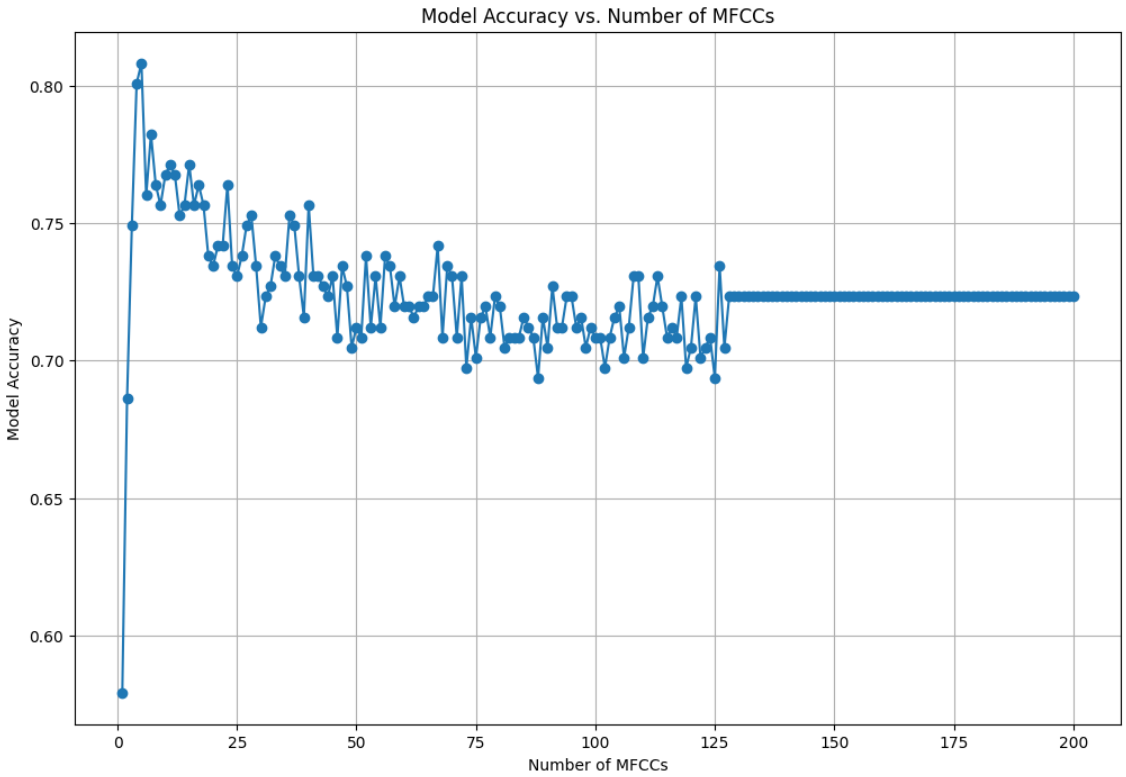


Fig. 21. MFCC Hyper-Parameter Tuning, with the number of MFCCs used to calculate the Random Forest Classifier feature set increasing in one step increments from 1 to 200.

It is clear that five Mel-filter banks returned optimal results, with a test accuracy of 80.8% from the Random Forest classifier. The performance decreases beyond this value, with the test accuracy saturating at approximately 72% for high values of Mel-filter banks used. The random forest was used as the initial machine learning algorithm implemented. However, further improvement of the model performance was desired and, so, Multi-Layer Perceptron (MLP) neural network architecture was tested [93]. A grid search was applied to identify the optimal tuneable parameters of the MLP, where the number of hidden layers, number of neurons per layer, dropout rate and learning rate value were implemented. A test accuracy of 76.752% was achieved, and so a decrease from the random forest classifier. Assessment of the performance for a regression model was then conducted, where an increase of 1.48% was achieved, bringing it to an accuracy of 78.23%. Hyper-parameter tuning of the optimiser, loss function and learning rate for the regression model were then performed and increased the test accuracy to 81.549%. This analysis of the MFCC features of the sonified signal, based on the simple artificial neural networks discussed here, shows the distinguishing powers of the MFCC features across the four HIE grades. In order to assess the further potential for increasing the

performance of this MLP for HIE grading, Principal Component Analysis (PCA) of the MFCC features for each signal was conducted.

PCA is a method for data analysis where the data is transformed into a new coordinate system to identify the most important features of the data [94]. The goal of PCA is to reduce the dimensionality of the data, while retaining as much of the variance as possible. The relationships between the data can be described by several inter-correlated quantitative dependent variables [94].The first principal component is the direction in which the data varies the most, with the second principal component being the direction in which the data varies the second most and so on. The PCA analysis of the MFCC features of the sonified EEG signal is shown in Fig. 22.

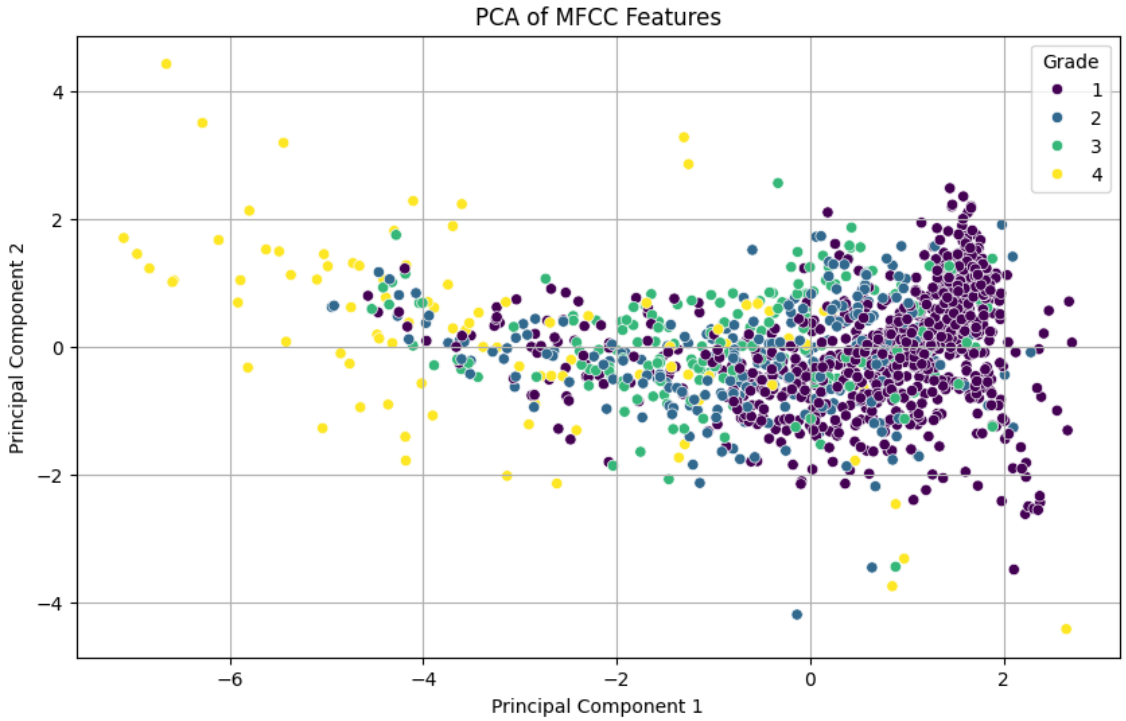


Fig. 22. PCA Analysis of the MFCC Features of the Sonified EEG Signal.

Grade 1 is clearly the most homogeneous of all the four HIE grades, containing the most distinct and separable features in comparison to the remaining three grades. This is evident based on the large grouping of grade 1 segments in the right side of Fig. 22. It is clear that the MFCC features of grade 2, 3 and 4 are more intertwined and so increasing the complexity and challenge of identifying the grade based on these features alone. When comparing the true grade versus the predicted grade based on the MLP model for MFCC features, a clear overlap of the regression values for grades 2, 3 and

4 are evident in Fig. 23.

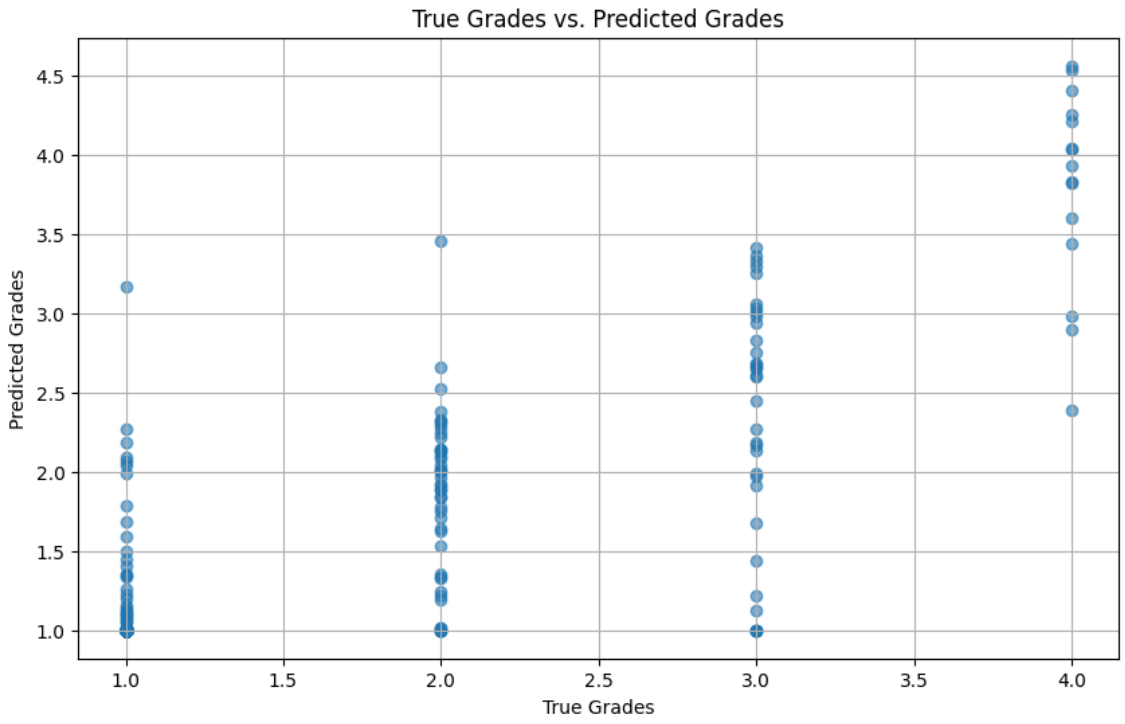


Fig. 23. Regression Outputs of the MLP Model for MFCC Features.

This supports the interpretation of the PCA visualization where it shows that there are very subtle differences between the MFCC features and they may require additional analysis, or higher dimensional features, to differentiate between the grades. Capturing a greater complexity of the data by increasing the number of components selected for the PCA analysis resulted in similar trends identified as when two components are used and so further methods of analysis are required as the PCA analysis is not sufficiently differentiating the data between the four grades.

In order to further explore the structure of the audio signal and compare this structure and its relationships between the different grades in the dataset, Uniform Manifold Approximation and Projection (UMAP) analysis was conducted. PCA helps to identify maximising variance and linear correlation. It is clear that the audio domain signal is highly non-linear, hence the poor performance for PCA analysis to separate the features of the signal based on the HIE grade. UMAP analysis has the potential for providing deeper insights into the dataset. The goal with applying UMAP analysis was to capture a more nuanced representation of the underlying structure of the audio data and also provide a methodology for direct comparison with the EEG signal behaviour.

1) Visualisation Analysis Method: UMAP is a mathematical dimensionality reduction technique used to preserve the global and local structures of data in a lower-dimension space [95], with significant benefits for visually representing data in a low dimensional plot. UMAP analysis is constructed based on two main mathematical stages, the construction of a weighted graph with each data point connected to its nearest neighbour and then the optimization to a low-dimensional embedding, which preserves the structure of the high-dimensional graph [95]. The algorithm holds two primary goals which are to preserve local distances in order to maintain structure within local neighbourhoods and to approximate the global structure of the data, maintaining the global connectivity between points.

UMAP assumes that the data lies on a manifold, this approximation allows UMAP to compress the dimensionality of the data while maintaining the structural relationships [95]. Uniform distribution in this manifold is maintained by adjusting the distance between nearest neighbour points in the construction of a k-nearest neighbour graph. This graph is then weighted, with the edge weights representing the strength of connections between nearest neighbour points, and is done so based on the probability that they belong to the same local neighbourhood [95]. The concept of fuzzy simplicial set is then used to represent the topological structure of the data, where each data point is considered as a member of a local fuzzy set that represents the probability of connectivity with other points. A Gaussian kernel function based on the distance function between points x_i and x_j is used to calculate the local connectivity probability p_{ij}.

$$p_{ij} = \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j) - \rho_i}{\sigma_i}\right) \tag{24}$$

where ρ_i is the distance to the nearest neighbour, and σ_i is a scaling factor that is determined such that a fixed number of neighbours are considered within a certain probability threshold. A global fuzzy simplicial set, p_{ij}^{sym}, is created based on the symmetrized local connectivity, as per Equation 25.

$$p_{ij}^{\text{sym}} = p_{ij} + p_{ji} - p_{ij}p_{ji} \tag{25}$$

Low-dimensional optimization is then implemented with the goal to find a low-

dimensional embedding Y that preserves the fuzzy topological structure from the high-dimensional space. The low-dimensional probability, q_{ij}, between points y_i and y_j, is defined using a Laplacian kernel function (26).

$$q_{ij} = \frac{1}{1 + a\|y_i - y_j\|^b} \tag{26}$$

Where a and b are parameters controlling the tightness of the embedding. UMAP minimizes a cross-entropy loss to align the high-dimensional probabilities p_{ij} with the low-dimensional probabilities q_{ij} (27).

$$C = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) + (1 - p_{ij}) \log\left(\frac{1 - p_{ij}}{1 - q_{ij}}\right) \tag{27}$$

This cost function encourages the preservation of both local and global structures by penalizing discrepancies between p_{ij} and q_{ij}. In order to minimize the cross-entropy loss, UMAP uses stochastic gradient descent (SGD), an iterative process which adjusts the low-dimensional embeddings y_i to better match the high-dimensional structure, represented by the fuzzy simplicial set.

C. Training Methodology: Nested Cross-Validation Evaluation Framework

To account for the limitation of the dataset size, and to ensure overfitting was not present in the model, a robust training implementation was defined. Nested Cross-Validation (CV) has been proven to provide unbiased performance metrics, regardless of the dataset size [96] and is deemed the most robust evaluation framework for small datasets, as in this study. Nested CV was implemented with K fold inner and K fold outer CV. Stratified Group K Fold CV was employed, with the groups determined based on the patient IDs. This ensured a patient independent model was developed, where no data from the same patient appeared in both the training and testing sets within any fold.

By utilising stratified group k fold CV, the distribution of the dataset across the folds was maintained. This is crucial, given the skewed distribution of classes amongst the dataset, which is reflective of patients in neonatal care units [97], where the number of higher severity cases are decreased. The workflow for the nested CV is detailed in Fig. 24.

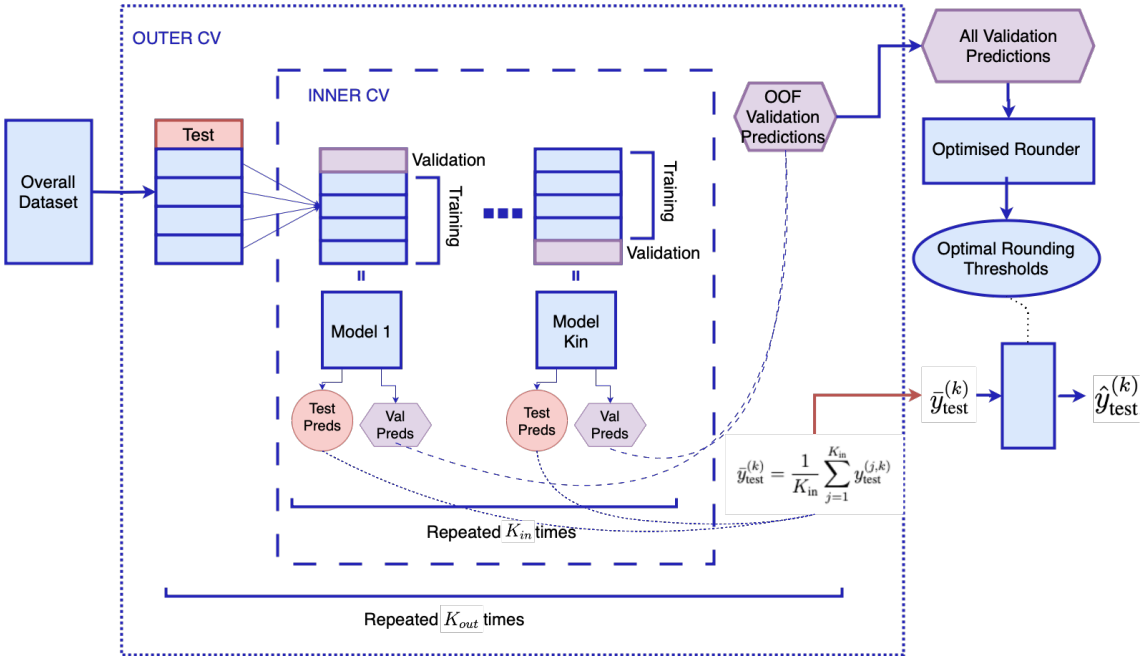


Fig. 24. Nested Cross Validation Workflow

The entire dataset is divided into K_{out} folds, where $K_{out} = 5$ for this study, stratified by grade and grouped by patient ID. K_{out} is set to 5 here due to the distribution of the dataset, where five patients contain samples of grade 4 [15], and so ensuring each outer fold has one patient with grade 4 samples present. A subset of the data $D_{test}(k)$ is held out as an unseen test set in each outer fold and the remaining data $D_{train(K)}$ is used for training and validation, where $D_{train}(k) = D - D_{test}(k)$, where D is the overall dataset. Within the inner CV, $D_{train}(K)$ is further divided into K_{in} folds, where $K_{in} = 4$ here. This is chosen based on the remaining four patients with grade 4 samples after one of these patients was excluded (for unseen testing). The model is trained on $D_{train-in}(k)$ and validated on $D_{val-in}(k)$, with the model performance evaluated on $D_{test}(k)$. This process is repeated for each fold of the outer CV, with the performance of each inner loop model, M_{in}, being evaluated on the test set of each outer fold. The predictions for each $D_{val-in}(k)$ are stored in the out-of-fold (OOF) array *oof_pred*. The performance for every outer fold (k) is measured based on the average aggregation of all test predictions $\bar{y}_{\text{test}}^{(k)}$, with $\bar{y}_{\text{test}}^{(k)}$ calculated as per Equation 28, for each file ID in each inner fold, *j*.

$$\bar{y}_{\text{test}}^{(k)} = \frac{1}{K_{\text{in}}} \sum_{j=1}^{K_{\text{in}}} y_{\text{test}}^{(j,k)} \tag{28}$$

The Optimized Rounder algorithm is trained on the validation predictions stored in *oof_pred* and the optimal rounding thresholds are determined to convert the continuous model predictions to a discrete clinical grade. The aggregated predictions, $\bar{y}_{\text{test}}^{(k)}$, are passed to the trained optimised rounder and the clinical grades are denoted as $\hat{y}_{\text{test}}^{(k)}$. The overall accuracy for each outer fold k is computed based on $\hat{y}_{\text{test}}^{(k)}$ to the corresponding ground truth labels, as per Equation $y_{\text{true}}^{(k)}$.

The final overall model performance is determined by aggregating the accuracies from all K_{out} folds, as per Equation 29. In Equation 29, $\mathbb{I}$ denotes the conditional statement where when the condition is true a value of 1 is returned, else 0 is returned.

$$\text{Accuracy} = \frac{1}{N} \sum_{k=1}^{K_{\text{out}}} \sum_{i \in \mathcal{D}_{\text{test}}^{(k)}} \mathbb{I}\left(\hat{y}_{\text{test},i}^{(k)} = y_{\text{true},i}^{(k)}\right) \tag{29}$$

This nested CV appproach ensures the model's performance can be evaluated in a robust and unbiased manner, in order for the model performance to confidently reflect how well it will perform on new, unseen data.

D. Results

Initial results look at evaluating the model performance for a fixed data split of training, validation and test sets. This was a patient independent split with the distribution of the overall dataset maintained in each split. All hyper-parameter testing and evaluation was performed on this fixed data split. The best model, defined by the configuration in the proposed method, resulted in an accuracy of 93.94% on an unseen test set, where the test set is made up of 20% of the overall dataset. The learning curves of this plot are shown in Fig. 25, where the trend of the training and validation loss are closely mapped, indicating good fit of the model to the validation data. The model was trained for 120 epochs, with the best model selected based on the validation loss as the chosen metric and so was extracted at epoch 109. The test accuracy for the best model is also plotted in Fig. 25, where it is seen to closely map the validation accuracy curve. This highlights the generalisability of the model, which is successfully predicting on unseen data.

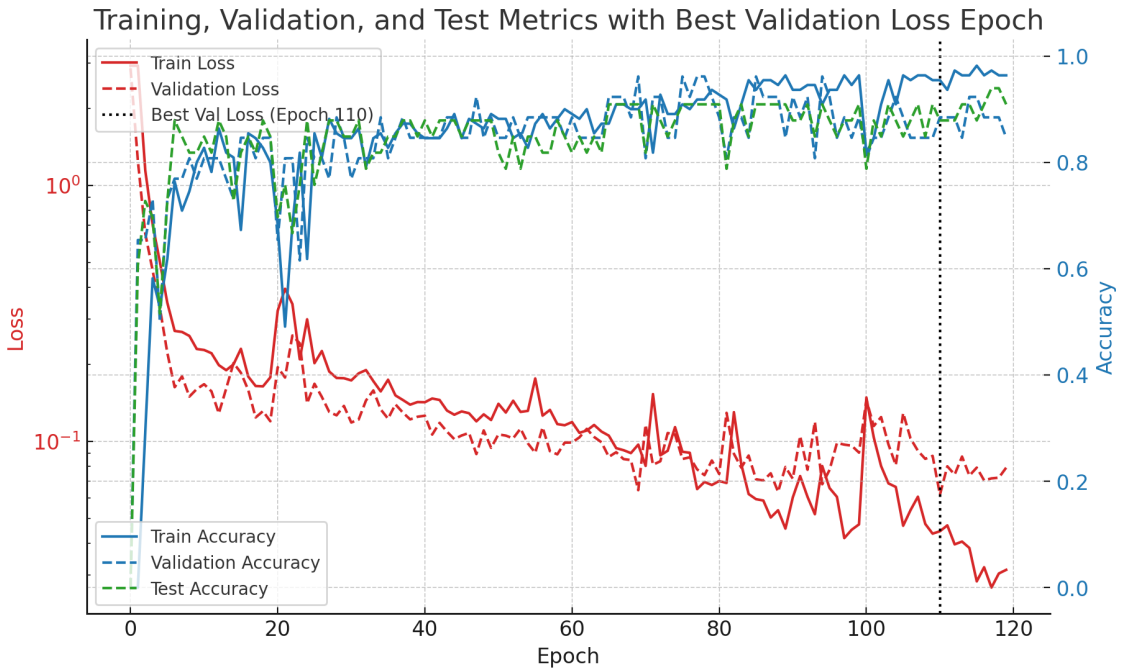


Fig. 25. Best Model Learning Curves

The nested cross-validation was performed and an overall test accuracy of 89.97% was achieved. The results of the model performance are provided in Table V, where initial inspection shows high overall model performance, indicated by all performance metrics surpassing the 0.8 boundary for good performance.

TABLE V

NESTED CROSS-VALIDATION PERFORMANCE RESULTS

Metric	**Value**
Precision	0.9079
Recall	0.8994
F1-Score	0.8985
R-Squared	0.8507
Cohen's Kappa Score	0.8219

The model demonstrates strong performance in identifying true positives, indicated by the high precision score of 0.9079, showing that there are relatively few false positives. The recall score is slightly lower at 0.8994, reflecting the model's ability to detect most relevant instances, though it misses a small number of positive cases. The balance between precision and recall is well captured by the F1-score of 0.8985, suggesting a good trade-off between these metrics. Furthermore, the R-squared value of 0.8507

indicates that the model explains a substantial proportion of the variance in the data, highlighting the accurate fit of the model, along with its reliability. The Cohen's Kappa score of 0.8219 further confirms the model's effectiveness, as it demonstrates a strong agreement beyond chance.

The confusion matrix of the model is then evaluated, to understand the relationship between true positives, true negatives, false positives and false negatives across the four HIE grades. Fig. 26 shows the normalised confusion matrix for the overall model of the nested cross-validation.

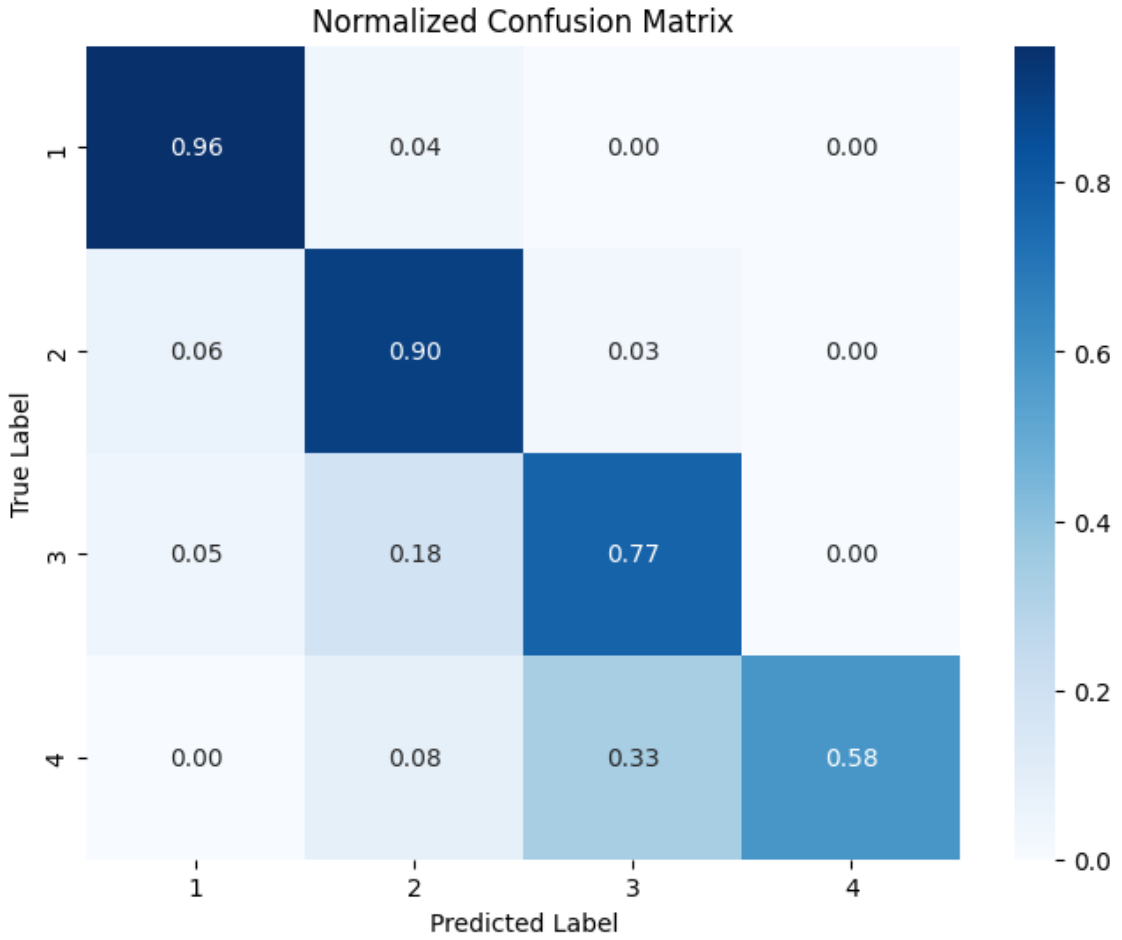


Fig. 26. Normalised Confusion Matrix Representing Test Results of the Nested Cross-Validation Training Scheme

Analysing the results of the confusion matrix it is clear the model performs well at identifying grade 1 samples, with a slight challenge present classifying between grade 1 and 2. Four grade 1 samples are wrongly classified as grade 2. A noticeable confusion increases amongst grades 2, 3 and 4, however. The main reason for this is likely due to the limited available samples of each of these grades in comparison to the number of grade 1 samples in the dataset. It is also important to note that there is some asymmetry

present in the normalised confusion matrix, where the sum of the elements above the diagonol is smaller than the analogous elements below the diagonal. This indicates that if an error is present in the model prediction, the error would typically tend to underestimate the HIE grade, rather than overestimate it. These results are all stated on the overall performance of the nested CV scheme. However, individual performance of each fold in the CV is also assessed.

The training and validation loss learning curves are plotted in Fig. 27, where the fluctuation between folds is highlighted by the shaded area. Despite fluctuation across folds, which is mainly due to the different data splits, both the training and validation loss functions follow similar trends. In addition, the standard deviation range for the validation loss is smaller than that of the training loss, showing consistency of the validation performance across folds. This highlights the model's ability to generalise well to the validation data, therefore the same is expected for the unseen test data. The training and validation accuracy responds as expected, based on the loss learning curves, where both curves follow the same trend and, so, showing the model's ability to learn relevant features in the spectrogram image related to each HIE grade. The accuracy learning curves are plotted in Fig. 28.

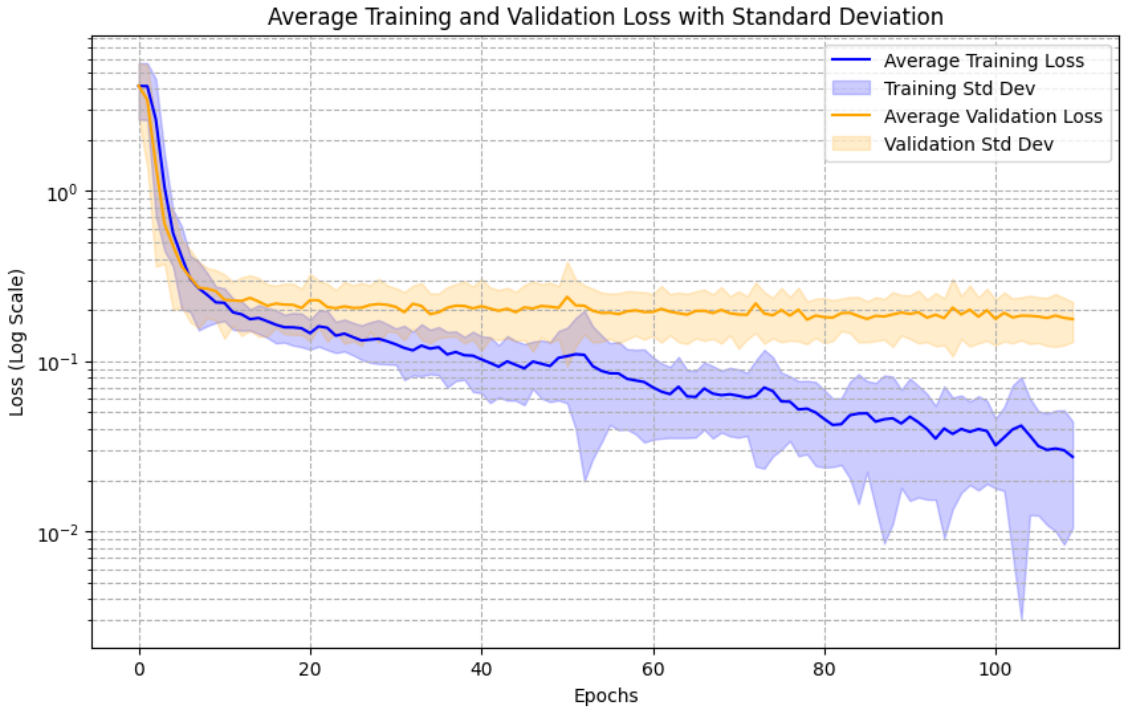


Fig. 27. Nested Cross Validation Training and Validation Loss Learning Curves

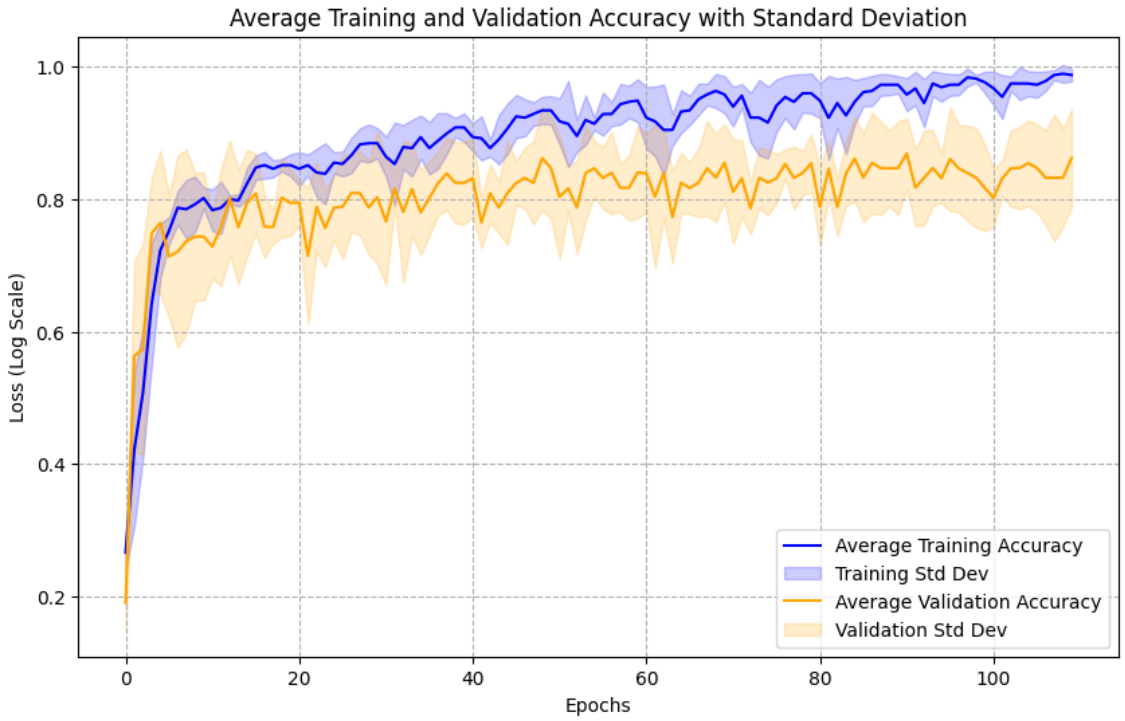


Fig. 28. Nested Cross Validation Training and Validation Accuracy Learning Curves

The test accuracy for each outer fold of the nested cross validation is shown in Fig. 29, where the test accuracy for each outer fold is calculated, based on rounding the average prediction for each file segment across the four inner folds. The accuracy is computed based on the comparison of the rounded average predictions to the ground truth for each label.

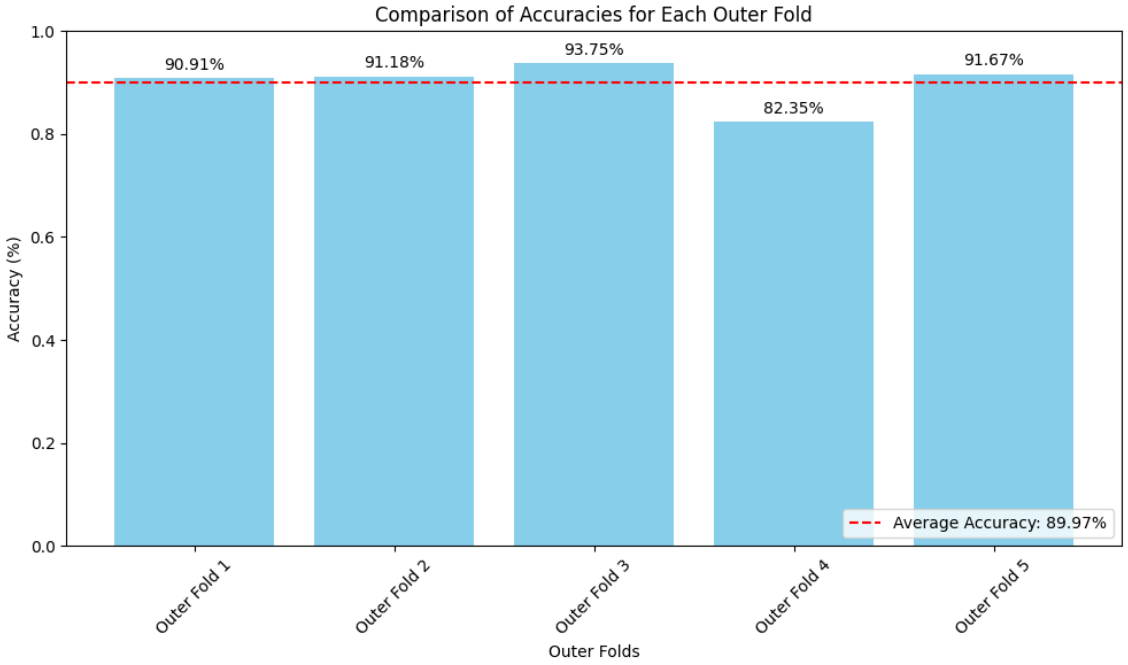


Fig. 29. Nested Cross-Validation Outer Fold Test Accuracy for all Five Outer Folds

As is clearly evident from Fig. 29, there is a significant drop in test accuracy for outer fold 4. This indicates that there is something causing significant challenge to the model in predicting the test data segments in this fold, in comparison to the test segments of all other outer folds. The test accuracy for outer fold 4 has a relative decrease of 8.47% from the overall average accuracy. The test accuracies of all other folds are closely mapped to one another. In order to identify the cause of this decrease in performance for outer fold 4, deeper analysis into this fold is performed and the findings are provided in the discussion of this chapter.

While the overall test accuracy for the nested cross-validation was computed to be 89.97%, the 95% Confidence Interval was evaluated to understand the reliability of the overall accuracy result [98]. The confidence interval is computed as per Equation 30, where $\bar{A}$ is the mean accuracy, σ is the standard deviation, Z is the z-score for the 95% confidence interval (1.96) and n is the number of folds.

$$\text{Confidence Interval} = \bar{A} \pm Z \times \frac{\sigma}{\sqrt{n}} \tag{30}$$

A 95% confidence interval of 89.15% to 90.73% was achieved, thus stating that we can be 95% confident the true accuracy of the model lies within this range.

E. Discussion

The above results show the proposed method achieves a test accuracy of 89.97%, with a 95% confidence interval of 89.15% to 90.73%. This is a significant result, surpassing the current state-of-the-art for autonomous HIE grading based on EEG signal analysis, recently stated as 86.09% on a large unseen test set by Yu et al [24]. In addition, the proposed method surpasses the leave-one-patient-out performance of 88.9% achieved by Raurale et al in [31], which also used a CNN to determine the HIE grade based on a time-frequency representation of the signal also. However Ruarale et al used much smaller windows of EEG at 5 minute segments. The work proposed throughout this chapter uses the FM/AM transformation representation of 60 minutes of EEG. While the advantage of the proposed method in improving the state-of-the-art accuracy for autonomous HIE grading is obvious, direct comparison between the results stated by Yu et al in [24] cannot be performed. This is due to the different datasets used in both studies, where Yu et al used a large unseen dataset, double in size to that of the publicly available dataset used here, and which is not publicly available [24]. However, the nested cross-validation method implemented in this study accounts for the limited size of the dataset. This increases the robustness and generalisability of the model, thus increasing the comparability of the two methods where the method proposed here shows superior performance.

1) Visual Analysis: The UMAP analysis of the EEG signal is performed as the baseline metric to understand whether any patterns or trends can be identified from the EEG signal to separate them into the four HIE grades. Fig. 30 shows the UMAP analysis for all EEG signals in the dataset, where no clear trends are apparent after visual inspection.

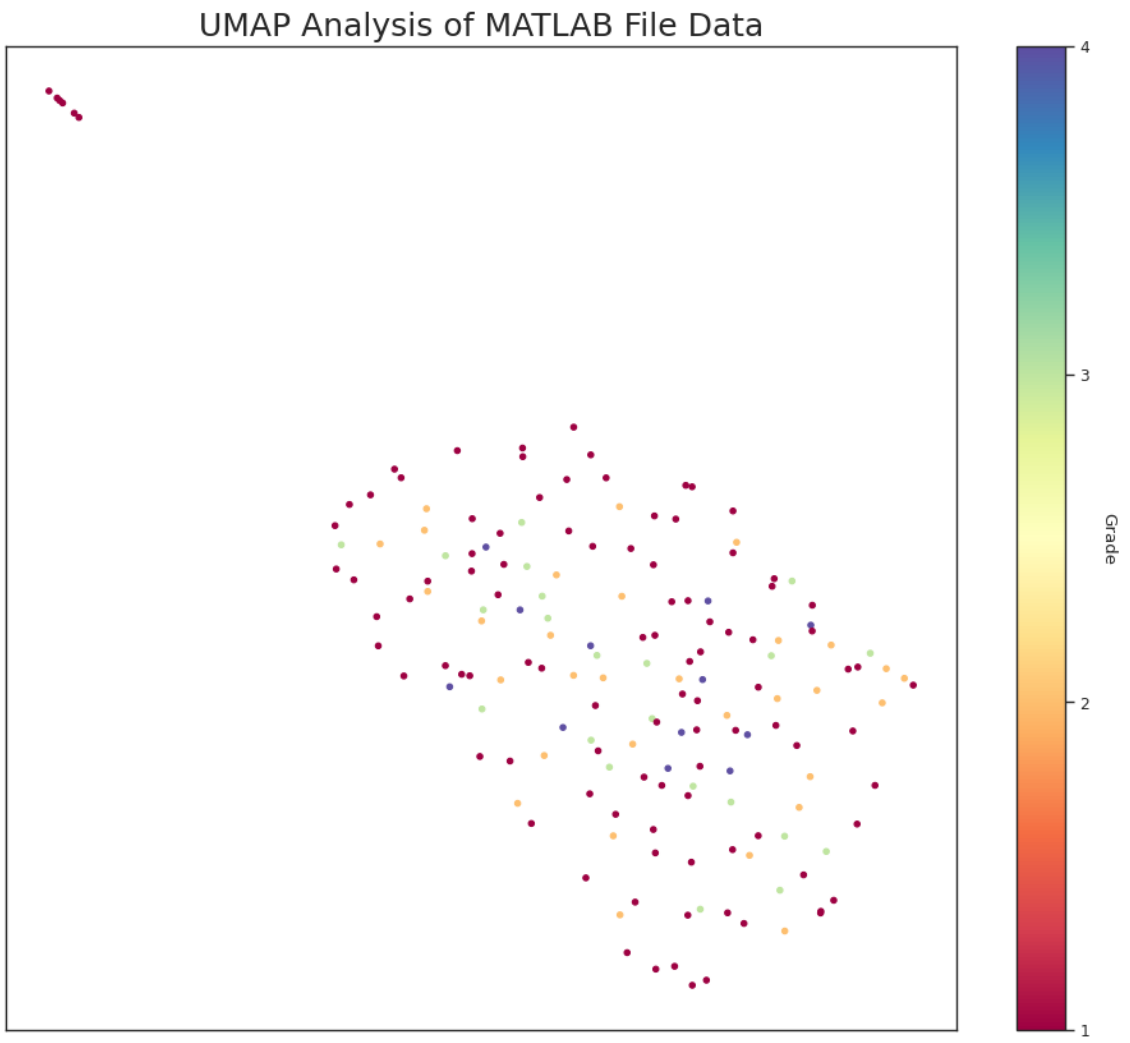


Fig. 30. UMAP Analysis of Raw EEG Data, visualising the relationships and structures between the data in a 2D plane.

By representing the multi-channel and multi-dimensional signal of the EEG in the 2D space, the challenge with separating the HIE grades based on the EEG signal was identified. The lack of clustering or trends present in Fig. 30 shows that the EEG signal, without any pre-processing, does not effectively differentiate between the grades. The UMAP analysis of the sonified EEG signal is performed and visualised in Fig. 31, where it is assesed on an individual channel basis. A more structured grouping of the signals, based on the HIE grades, is noted in Fig. 31, indicating the benefit of representing the EEG signal in the audio domain for the classification task. While overlap between the grades is still present, observable trends can be identified, which is an improvement in differentiability in comparison to the raw EEG signal seen in Fig. 30.

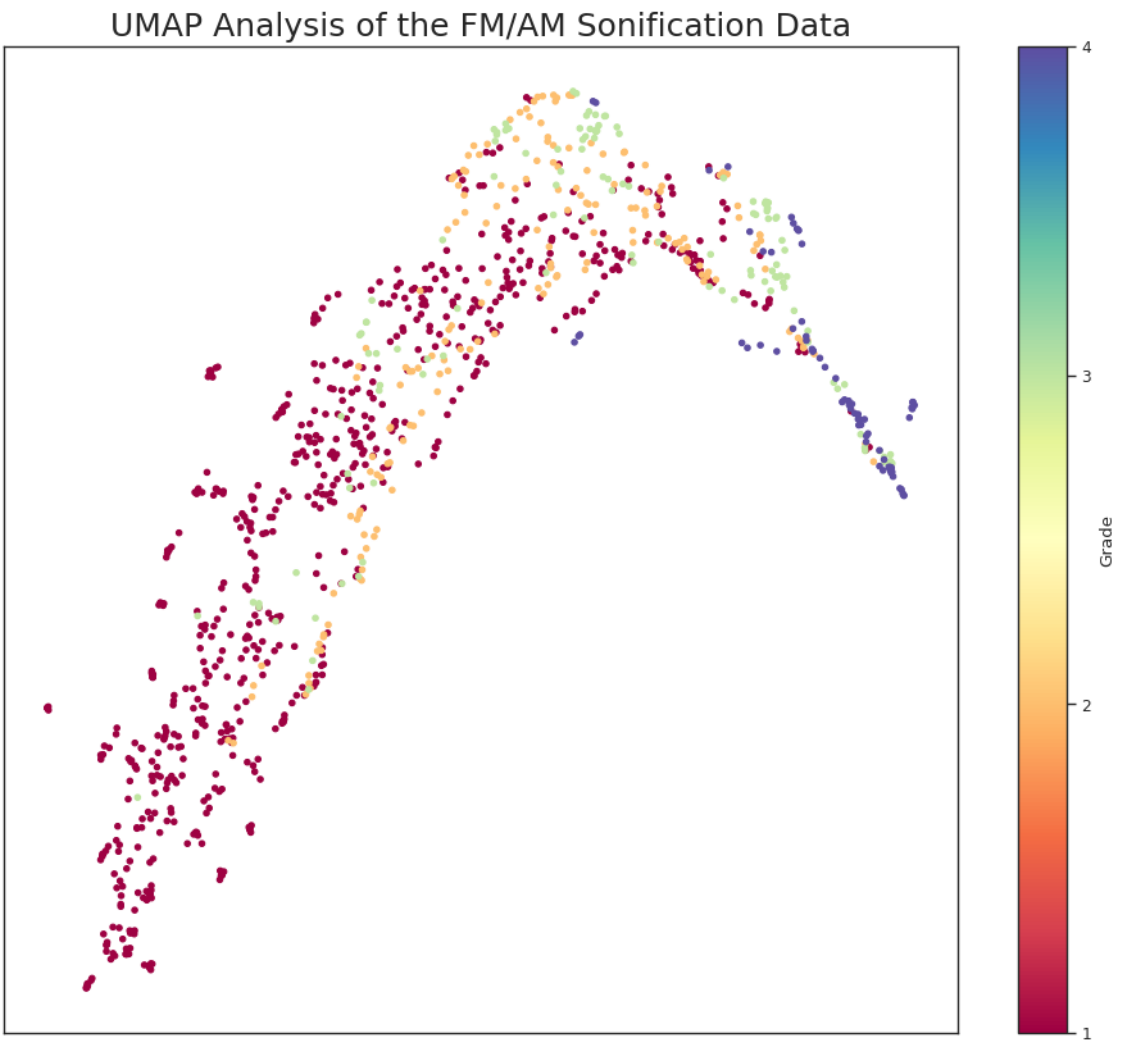


Fig. 31. UMAP Analysis of Sonified EEG Data.

Grade 1, as indicated by the red points on the plot, are grouped closely together, starting in the bottom left corner of the graph and gradually dispersing across the centre of the plot. Grade 2, as indicated by the yellow points, are then spaced out across the mid range of the grade 1 grouping and then mix with those data points of grade 3, identified as green. Grade 4 is again more dispersed, showing that it is most different from grades 1 and 2, as expected, with a few samples of grade 3 mixing in this grouping. Unlike the UMAP plot of the raw EEG, where there is little to no groupings in the dataset, the sonified signal showed a clear increase in the separation of the data points based on their HIE grade. This analysis showed the benefits of the FM/AM sonification algorithm in highlighting the rhythmic behaviour of the EEG signal and, thus, enhancing the emphasis on the continuity and discontinuity of the EEG signal.

While the sonification of EEG signal to the audio domain shows potential for im-

proved separation between the grades, the need for more sophisticated methods to accurately identify the HIE grades is necessary. This further supports the need for the deep learning backbone in the proposed method. In order to evaluate just how well the CNN is performing at increasing the separation between grades, UMAP analysis is performed at the feature map level of the last convolutional layer of the proposed CNN architecture. Visualisation of this UMAP analysis is given in Fig. 32.

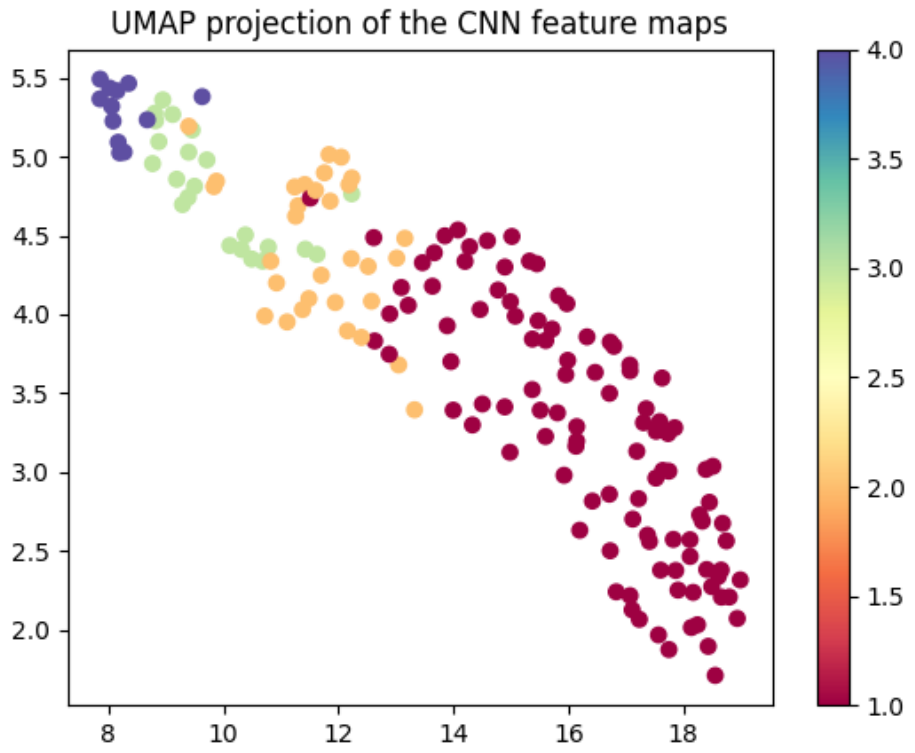


Fig. 32. UMAP Analysis of the Last Feature Map output of the CNN model for the proposed method with FM/AM transformed EEG spectrogram

Improvement in the separation between HIE grades at the feature map level of the CNN is evident from visual inspection. There is still a small amount of overlap, with the cluster boundaries all very close together. However, overall there is a very significant improvement from the baseline UMAP of the unprocessed EEG signal. While this analysis shows the benefit of the proposed method which includes the FM/AM transformation and the 2D regression network, in order to assess the improvement in separation of the CNN alone, UMAP analysis of the baseline HIE grading method proposed in Chapter 3 is evaluated. This evaluates the separation behaviour of the convolutional layers when the Mel Spectrogram representation of the EEG signal is

used instead of the spectrogram representation of the audio domain signal. Fig. 33 plots the visualisation of this UMAP analysis, where tight clustering is noted. However, strong overlap between grades exists.

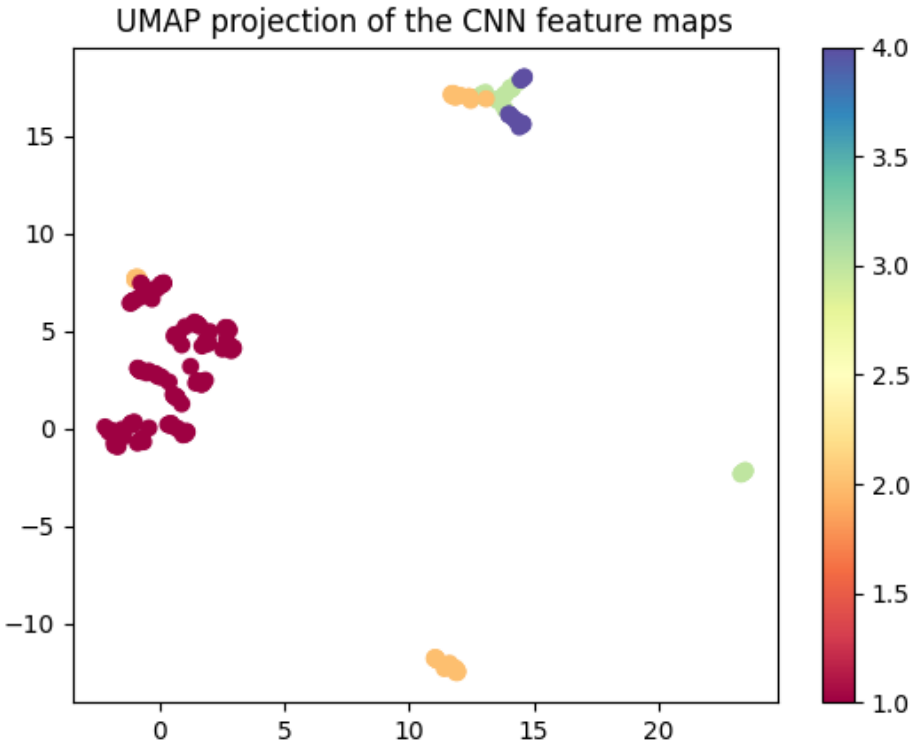


Fig. 33. UMAP Analysis of the Last Feature Map output of the CNN model for the baseline 2D CNN model based on the Mel Spectrorgam representation of the EEG signal.

Three quantitative metrics to assess the UMAP performance for different data representations are used to evaluate the clustering performance of the UMAP analysis in addition to visual inspection. These metrics are listed as follows:

1) **Silhouette Score:** The silhouette score is a quantitative metric which measures the separation between clusters in the UMAP embedding [99], which in the case of this analysis is the separation between HIE grades. The silhouette score is a measure between -1 and 1, which describes how similar a point is to its own cluster, compared to other clusters. The silhouette score is calculated based on Equation 31 using the Scikit-learn python library [50], where $|a|$ is the mean intra-cluster distance and $|b|$ is the mean nearest-cluster distance of each sample.

$$s = \frac{b - a}{\max(a, b)} \tag{31}$$

2) **Davies-Bouldin Index:** Unlike the silhouette score which determines how similar each data point is to its own cluster, the Davies-Bouldin Index (DBI) is a ratio metric which focuses on the ratio between within-cluster scatter to between-cluster scatter [100].
3) **Calinski-Harabaz Index:** The Calinski-Harabasz Index (CHI) is the final quantitative metric used to asses the performance of the UMAP separation analysis in this study. This provides an insight into the overall picture of cluster separation and compactness from the UMAP analysis, as described in Equation 32, and helps to assess the general cluster structure of the data [101].

$$CH(K) = \frac{B(K)}{W(K)} \times \frac{N - K}{K - 1} \tag{32}$$

where $B(K)$ is the between-cluster scatter, $W(K)$ is the within-cluster scatter, N is the number of data points and K is the number of clusters.

The silhouette scores, DBI and CHI values for all four data representations are provided in Table VI, where the feature map of the convolutional layer for the spectrogram representation of the EEG signal performs best overall.

TABLE VI
CLUSTERING METRICS FOR UMAP ANALYSIS OF DIFFERENT DATA REPRESENTATIONS

Data Representation	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Raw EEG	-0.2602	14.637	0.8584
Sonified Signal	0.0600	2.985	224.3080
Sonified EEG CNN Feature Map	0.3457	0.800	166.8200
EEG Mel-Spectrogram CNN Feature Map	-0.0218	2.04	40.712

Evaluating the silhouette score first, the raw EEG data was -0.2602 while the FM/AM sonified signal achieved a silhouette score of 0.06. Although the silhouette score for the sonified signal still indicates poor clustering, with significant overlap in cluster boundaries as seen in Fig. 31, there is an improvement of 0.3202 over the raw EEG data, representing a relative increase in performance of 123.02%. The clustering performance

decreased for the feature map of the CNN trained on the EEG Mel Spectrogram images (-0.0218) compared to the FM/AM sonified EEG UMAP analysis. The UMAP of the feature map of the CNN model based on the FM/AM transformed signal shows an improved silhouette score of 0.3457. While this method shows an improved silhoutte score over all other models tested in this study, it is still lower than the ideal silhouette score of 1 for perfect separation. Evaluating this metric, in conjunction with visual analysis of the plot in Fig. 31, it is evident that while the samples are all clustered and showing clear trends seperating each grade, internal to each cluster the data points are dispersed and there is very limited distance seperating cluster boundaries. Given that the silhouette score represents the amount of seperation within the cluster, along with the strength of the cluster boundaries by definition, a lower silhouette score is expected.

The DBI for the UMAP embedding of the raw EEG signal was 14.637, compared to 2.985 for the sonified signal. A value of 0 indicates perfect clustering for the DBI and so the sonified signal shows a relative decrease of 79.6% over the embedding of the raw EEG signal. Despite this significant improvement, the DBI value still remains high, indicating the challenge of seperating the clusters. The feature map of the CNN when the EEG signal is plotted as a Mel Spectrogram achieved a DBI value of 2.04, a relative increase of 86.08% over the original EEG signal and 31.66% over the sonified EEG signal. The need for the CNN to perform seperation is evident from these results and, once again, further supports the use of deep learning in the novel methods proposed in this thesis. A further improvement was noted for the UMAP embedding of the feature map with the CNN trained on the spectrogram of the sonified EEG. A DBI value of 0.8 was noted, representing a relative decrease of 73.2% compared to the sonified signal, 94.43% relative to the raw EEG data and 60.78% compared to the UMAP embedding of the feature map with the CNN trained on the EEG Mel Spectrogram images. This further confirms the effectiveness in separating the HIE grades of the proposed method, highlighting the importance of both the pre-processing steps and the deep learning network.

A higher CHI value indicates better-defined clusters with a substantial improvement from 0.8584 to 224.308 for the raw EEG signal and the sonified signal, respectively. This improvement represents a 26,008% increase. It is worth noting that there is no defined range for the CHI, and relative improvement serves as the main measure for

evaluating the clustering performance of the embedding for the FM/AM sonified signal in comparison to the raw EEG UMAP embedding. Interestingly, the CHI decreased for the UMAP embedding of both the feature maps from the CNN models. Despite this, the feature map of the CNN trained on the spectrogram representation of the FM/AM sonified signal shows superior clustering performance to that of the feature map for the CNN trained on the EEG Mel spectrogram.

By performing combinatory analysis of these three quantiative metrics on the UMAP embedding a strong understanding of the signal seperability based on different data representations can be achieved. This allows for a comprehensive analysis of the clustering behaviour of the UMAP analysis, interpreting the cohesion and seperation of each point, identifying the separability of the cluster boundaries and, finally, assessing the cohesion and seperation on a global scale for the dataset. Considering all aspects of this, the novel method proposed in this Chapter, using 2D CNN to grade the signals based on the FM/AM transformation of the EEG signal, shows superior performance at seperating the four HIE grades. The need for innovative signal processing techniques to highlight the HIE behaviours within the EEG signal is obvious from this UMAP analysis, with the UMAP embedding of the sonified EEG performing better than the embedding of the original EEG signal across all metrics. While this pre-processing improved the performance, the need for a more effective method to further separate the signals is evident. The CNN performs exactly as desired for that task, with the embedding of the feature map dramatically improving the clustering behaviour.

The decisions behind the different phases in the proposed model are clearly supported through this UMAP analysis, where the proposed method provided increased ability to classify the HIE grades. Significant hyper-parameter tuning went into the development of this model to fully harness the power of deep learning. Exploration into the specific choices within the CNN model development are detailed in the next section, focusing on the effects of the hyper-parameter configurations, which focused on optimizing the model to further enhance classification performance.

2) Hyper-Parameter Tuning Comparative Analysis: While the UMAP analysis supports the broad architecture of the proposed method, support as to the decision of various hyper-parameters within the model architecture must also be provided. Discussion into the various hyper-parameter tunings performed is provided here and includes evaluation

of the image generation of the sonified EEG signal and the deep learning architecture parameters. The comparative analysis for hyper-parameter tuning used a fixed data split, with patient independence between folds being maintained and the distribution of the overall dataset was maintained across training, validation and test. This was done to ensure comparability and to be as reflective of the final test performance as possible.

a) Analysis of different Spectrogram Parameters: In order to define the optimal spectrogram parameters, the model performance was evaluated on a single channel basis with a fixed data split. This analysis aimed to assess the optimal grouping of the number of samples per FFTs, the hop-length and the window size for the spectrogram calculation, which was computed using the Librosa Python library package [88]. Late decision aggregation was implemented to combine the predictions for every channel of each patient data segment, with the overall grade calculated being the rounded average of these values.

Performing a grid search of these three parameters has allowed for a clear evaluation of their effect on the model performance metrics and for the determination of the optimal parameters to provide sufficient information as regards the signal features and its associated HIE grade. Fig. 34 shows the difference in performance results for the each spectrogram parameter configuration tested. The optimal spectrogram parameters were identified as 128 FFTs, 64 hop-length and 128 window size. The performance of the model using these parameters achieved a validation loss of 0.1255 and a validation accuracy of 96.15 %.

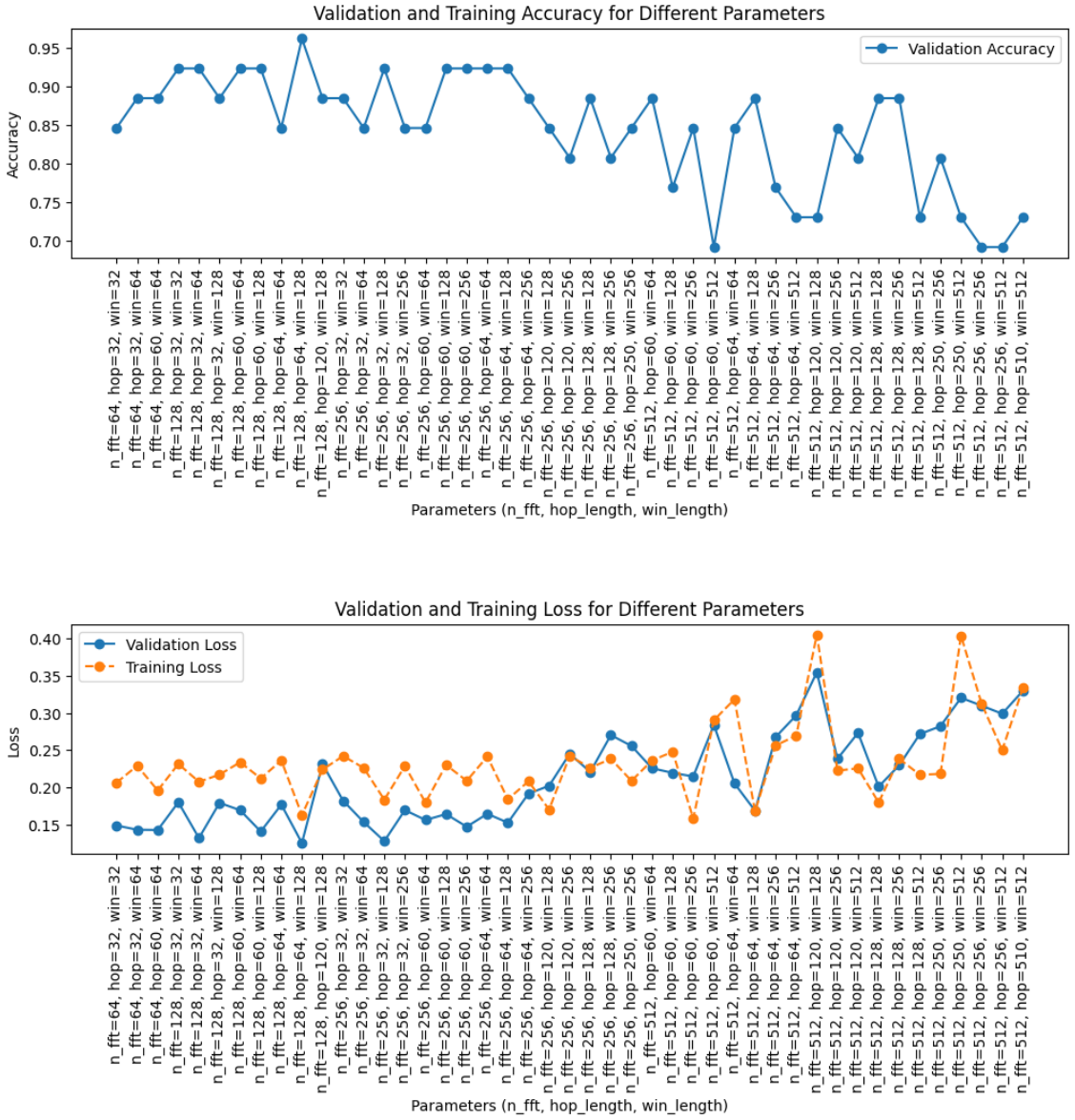


Fig. 34. Validation Accuracy and Loss for various spectrogram parameters under analysis

Given the optimal spectrogram parameters, a frequency resolution of 4 Hz was achieved. Every frequency bin of the spectrogram represents a 4 Hz frequency range, with the maximum frequency captured by the FFT being defined by Nyquist to be half the sampling frequency, i.e. 256 Hz. The time resolution, and so the duration between any two frames is 0.125 seconds, based on the hop length being 64 samples. Given that the window length is set at 128 samples, an overlap of 50% will be present between each window and so increasing the ability of the spectrogram to capture continuous changes in the signal. This is an important parameter as it has been seen to date that the HIE grade determined by visual representation is highly dependent on the continuity, or lack of continuity, present in the patient's EEG activity. With a sampling rate of the

signal defined as 512 Hz, each window represents 0.25 seconds of the signal and the resulting spectrogram will have 64 frequency bins spanning a range of 0 - 256 Hz and 1123 time frames for every 2 minute and 30 second audio sample.

The balance between time and frequency resolution is a key trade-off in any spectrogram computation and, when diagnosing the severity of HIE, both time and frequency domain features are important. The number of FFTs set at 128 appears to provide sufficient frequency resolution to identify subtle changes in the frequency component of the audio signal. This was achieved by modulation of the frequency of the raw EEG signal in the sonification process, which could be associated with brain dysfunction. The temporal continuity and the overlap between each window allow the model to see any transient or short-lived events in the signal that might correspond to abnormal brain activity, while also capturing any subtle shifts over time in the signal. Including this overlap in the spectrogram aids the CNN in identifying patterns across time by ensuring any key information is not fragmented across multiple windows and ensures a smoother transition between frames. This results in a more coherent feature map for the CNN to analyse. Each of these points being considered, the spectrogram parameters selected returned the best performance as they provided the CNN with the most informative and balanced representation of the features of the audio signal to determine the HIE grade.

b) Comparison of Regression Model vs Classification: In order to evaluate whether the hypothesis that a regression model would perform better than a classification model, given the nature of the grading task, a comparative analysis was performed with the results in Table VII. This hypothesis is based on the monotonic relationship between the four HIE grades and the brain activity expected in each grade. The four HIE grades are not independant of one another and so a Classification task would not accurately represent this.

TABLE VII

REGRESSION VS CLASSIFICATION MODEL PERFORMANCE

Model	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss	Test Accuracy
Regression Model	96.30%	0.0348	92.31%	0.0505	93.94%
Classification Model	94.55%	0.1168	88.46%	0.1656	90.91%

Comparing the results of the classification model in comparison to the regression

model an overall improvement in performance is noted for the regression model, thus aligning with the defined hypothesis in this test. This hypothesis is that a regression model best represents the grading task under evaluation due to the scaled nature of the HIE grades. The unseen test accuracy, while good for classification, at 93.94% is best for the regression task. Both the training and validation loss for the regression task, using Mean-Squared Error (MSE) Loss [64], are lower than the classification model which utilised Cross-Entropy loss [102]. In addition to the regression loss being smaller, the difference in the training and validation loss is smallest when a regression model is implemented. This indicates that the model is generalising well to the validation data. This experiment shows the regression model's ability to learn and leverage the monotonic relationship between the grades. Because of this, the regression model was selected as the final model for the grading task.

c) Data Augmentation: Evaluation of the impact of data augmentation on the model performance was tested for the the nested CV training scheme, with the aim of adding data augmentation to the training data being to reduce the difference between the training and validation loss. The Albumentations library [72] was used to apply the data augmentation before passing the image as input to the Pytorch model. The series of data augmentation configurations were defined as follows:

- **No Data Augmentation:** No data augmentation was applied; the original vertically stacked spectrogram images were used for training, validation, and testing.
- **Horizontal Flip:** Horizontal flip augmentation was then tested. Horizontal flip is applied to only a portion of the training data, with this portion chosen at random for each instantiation of the data loader based on the defined probability value.
- **Vertical Flip:** Augmentation of both horizontal flip and vertical flip applied to a portion of the training data only, once again based on the probability value.
- **Horizontal and Vertical Flip:** Training images were both horizontally and vertically flipped.
- **Channel Masking:** Vertical augmentation was applied to the stacked spectrogram image, with anything form 0% to 10% of the time axis being masked at a time.
- **Rotation:** All training images were randomly rotated within a limit of $\pm 10°$.
- **Random Brightness and Contrast:** Random adjustments of brightness and contrast were applied to the training images.

- **Gaussian Noise:** Gaussian noise with a variance between 10.0 and 50.0 was added.
- **Elastic Transformation:** Elastic transformations were applied.
- **Random Resized Crop:** Training images were randomly cropped and resized to 512×1024 pixels, with a scale factor between 0.8 and 1.0.

A representation of every augmentation scheme tested is provided in Fig. 35, where each augmentation is applied to the same data sample for comparability.

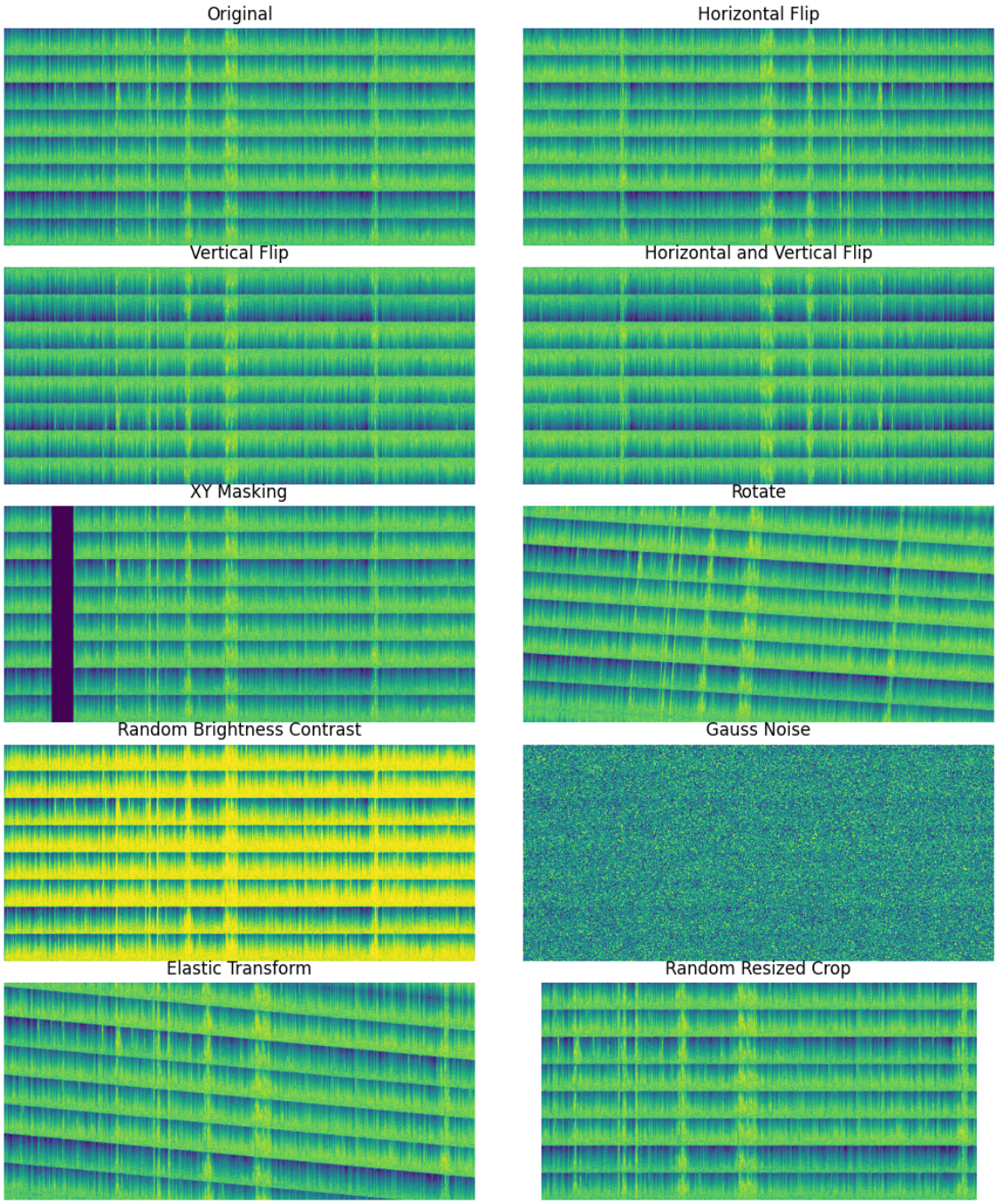


Fig. 35. Plot of a sample spectrogram for a chosen EEG signal with the data augmentation listed above applied.

The nested cross validation test accuracy for each data augmentation configuration

is provided in Table VIII, along with probability values of 10%, 30% and 50%. No augmentation applied to the training data performs the best compared to all other augmentation methods based on the results in Table VIII. This indicates that the model generalises well without requiring additional variations in data, potentially due to the inherent variability already present within the dataset.

TABLE VIII

TEST ACCURACY FOR EACH AUGMENTATION AND PROBABILITY

Augmentation	**Test Acc (Prob = 0.1)**	**Test Acc (Prob = 0.3)**	**Test Acc (Prob = 0.5)**
No Augmentation	0.89970828	0.89970828	0.89970828
Horizontal Flip	0.87573964	0.88165680	0.86982249
Vertical Flip	0.85798817	0.87573964	0.86390533
Horizontal and Vertical Flip	0.87573964	0.86390533	0.86982249
XY Masking	0.86982249	0.86982249	0.86390533
Rotate	0.85798817	0.85798817	0.86982249
Random Brightness Contrast	0.88165680	0.86390533	0.87573964
Gauss Noise	0.86390533	0.86390533	0.86390533
Elastic Transform	0.86982249	0.85207101	0.86982249
Random Resized Crop	0.88617845	0.84987651	0.84379722

The results in Table VIII highlight the data augmentations which were incompatible with the data. The Horizontal Flip or Vertical Flip augmentation, for example, reduced the test accuracy notably, suggesting that these augmentations added noise rather than beneficial variations, thus slightly hindering the model training. On the other hand, Random Brightness Contrast and XY Masking show relatively closer performance to the unaugmented model, suggesting some resilience to minor intensity and masking variations. However, even these did not improve upon the baseline accuracy, further supporting the assertion that augmentations may introduce changes that do not improve model learning for this task.

3) Analysis of EEG Input Window Size: The proposed model uses a 1-hour EEG segment as input to the CNN based on the FM/AM transformed signal plotted as a spectrogram image. Annotation of the dataset used in this study provided a single clinical grade for each 1 hour epoch of data and, so, the proposed model replicates this when defining the clinical grade, seeing the full duration of the single segment when making its prediction. As discussed previously, the input image representing the

full hour segment of EEG signal is highly rectangular in shape (image size of 65×1126). However, the ConvNextV1 Nano has a receptive field of 1328 however and so the full image is seen by the CNN model. While evaluating the performance of the 1-hour EEG segment inputted to the CNN, comparisons with different window sizes are necessary, particularly for implementation of this proposed method, in order to assess what is the minimum duration of EEG data which must be collected in order to provide high accuracy performance.

A window of 5 minutes, 10 minutes, 20 minutes and 40 minutes are all assessed in comparison to the 60 minute EEG input and the results are provided in Table IX.

TABLE IX

NESTED CROSS-VALIDATION PERFORMANCE FOR DIFFERENT INPUT WINDOW SIZE

Wndow Size	**Test Accuracy**
5 Minutes of EEG	53.2%
10 Minutes of EEG	74%
20 Minutes of EEG	83.46%
40 Minutes of EEG	86.42%
60 Minutes of EEG	89.97%

The trend from Table IX clearly shows that the larger EEG window greatly increases the test accuracy performance, with a minimal window of 20 minutes being required for acceptable nested cross-validation accuracy. This aligns with the findings given in [22], where a minimum of 20 minutes EEG signal is necessary for HIE diagnosis. The advantage of transforming the EEG signal to the audio domain based on the sonification process is the emphasis the transformation puts on the overall pattern and rythmic structure of the signal across the long window of the one hour EEG segment. Along with this, compressing the signal based on a speed of 20, reducing the signal from 60 minutes to 3 minutes of audio, allows for better evaluation of the full signal in a single image. As a result, the test accuracy for the 60 minute EEG signal shows high performance.

4) Discussion on Reduction in Performance for Nested Cross-Validation Outer Fold 4: A decrease in test accuracy was noted in outer fold 4 previously, which brought up a significant question as to what was causing such a large fluctuation from the baseline

performance of all other outer folds. A hypothesis as to why this has occurred is that the performance drop in outer fold 4 is likely due to the presence of particularly challenging grade 2 and 3 samples that exhibit outlying features, making them difficult for the model to classify.

Assessment of the training, validation and test data was performed in order to identify any outliers which were present. This step was pivotal in understanding if the error was caused due to the data the model was learning from, or whether the test data samples contained clear outliers from the dataset. With no significant decrease in performance for outer fold 4, in terms of the training and validation data it was assumed the problem was with the test data.

Initial analysis performed a direct comparison of the performance of outer fold 4 based on the EEG Mel Spectrogram model and the FM/AM transformed signal spectrogram method. A noticeable decrease in performance from the average accuracy was identified when the data was represented as a Mel Spectrogram image. This symmetry between the worsening performance for outer fold 4 based on both data representation proves that the sonification representation of the signal does not worsen the representation of the HIE features for the EEG signals in outer fold 4. This was an important evaluation as it verified that representing the EEG signal in the audio domain was beneficial for all EEG signals in the dataset and not just a subset of these signals.

No difference in data distribution was present in outer fold 4 in comparison to the other five folds and so the training and validation data used was reflective of the test dataset. In order debug this fluctuation, the accuracy per grade for each outer fold was calculated, where a clear drop in accuracy for grade 2 and 3 is present for outer fold 4 in comparison to the accuracy for predicting grade 2 and 3 samples of all other outer folds. A bar plot describing this accuracy per grade is seen in Fig. 36.

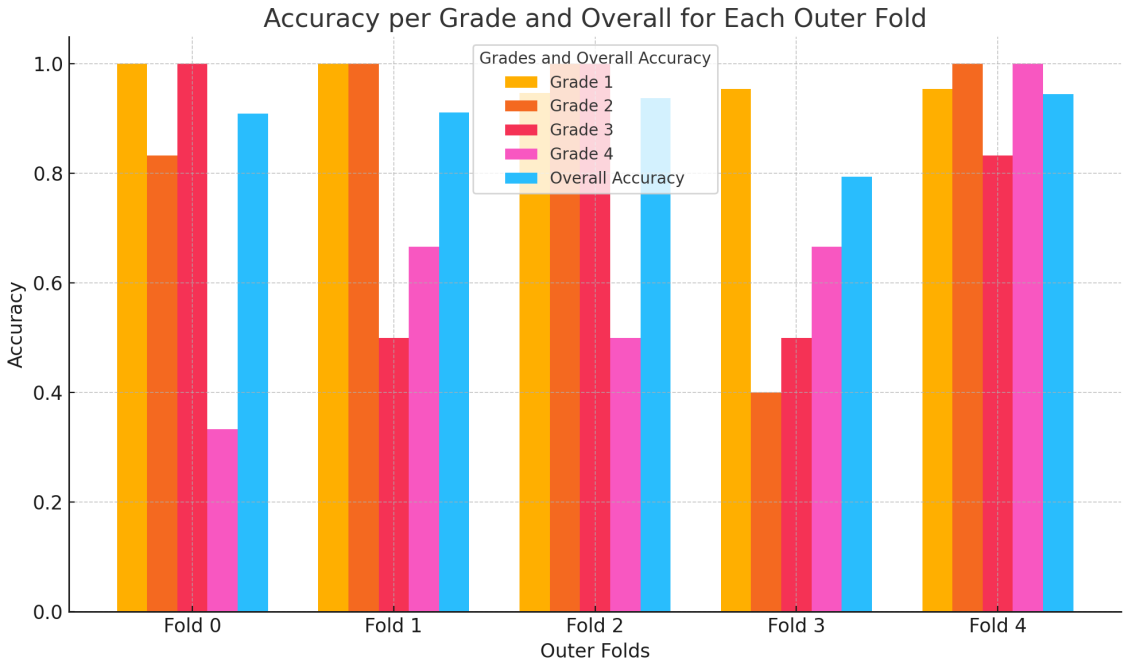


Fig. 36. Accuracy Per Grade per Fold for each Outer Fold

Delving deeper into why the grade 2 and 3 samples of outer fold 4 are performing very poorly, the incorrectly predicted samples in the test set of outer fold 4 were assessed. The predictions for these samples, when used as validation data and test data, are provided in Table X.

TABLE X

COMPARISON OF VALIDATION AND TEST PREDICTIONS FOR OUTLIER SAMPLES

File ID	True Value	Validation Predictions (Avg)	Test Predictions (Avg)
ID06_epoch4	4	3.5354, 3.2588, 3.7585, 2.9211 (3.3234)	3.35
ID03_epoch3	2	1.015, 0.982, 1.09, 1.078 (1.0418)	1.079
ID02_epoch1	1	2.0059, 1.0475, 1.7400, 2.22 (1.755)	1.63
ID20_epoch1	1	1.143, 2.4976, 2.387, 2.56 (2.14)	2.504
ID29_epoch1	3	1.0998, 1.625, 1.7237, 1.90576 (1.5887)	1.733
ID29_epoch2	3	2.589, 1.7465, 1.8679, 1.822 (2.00659)	1.746

It is obvious from the results in Table X that the model struggled to correctly predict these samples, even when used as validation samples, as both the validation predictions and the test predictions of these samples are incorrect. This lead to the hypothesis that there must be some outlying features in these samples which make identifying their HIE grade difficult to classify. While these samples are incorrectly predicted, all samples are within a tolerance of ± 1 of their true grade value, meaning that if a sample

was predicted as grade 1, the true value was either grade 1 or 2. This further supports the performance of the model, where the model is not simply predicting a random grade. The model accuracy, when accepting this tolerance in the model performance, is 98.82%. Thus, out of the overall test data, considering every point in the dataset to be an unseen test sample, only two samples were incorrectly predicted outside a ± 1 tolerance. This shows the high performance of the model, which is clearly learning the relationships between the features of the signal and the HIE grade, but identifying the perfect boundaries is a significantly difficult task. In addition, having access to the annotations of both annotators in the dataset, rather than the single consensus value, would have been beneficial to understand whether the samples proving difficult to grade were samples in which the annotators also had a discussion on the true grade present in the signal.

Given the significant drop in accuracy for outer fold 4, evaluation of the model performance excluding this fold was conducted. An average test accuracy of 92.59% was achieved with a 95% confidence interval of 91.81% to 93.38%, further highlighting the impact these outlier samples have on the overall model performance.

5) Comparison with 2D CNN Baseline: The initial implementation of HIE grading using 2D CNNs in this study was first introduced in Chapter 3 of this thesis and is used as the baseline evaluation for the following analysis. The goal of this analysis is to identify and quantify the benefit of representing the EEG signal in the audio domain has over the original EEG signal as input to the 2D CNN network. The method described in Chapter 3, which was trained on Mel Spectrogram representations of the EEG signal, will be denoted as the baseline performance for the following analysis. The novel method proposed in this chapter, using the FM/AM transformed spectrograms, will be referred to as the proposed method for ease of reading in the following section.

The baseline 2D CNN model proposed in Chapter 3 was then tested on the nested cross-validation evaluation framework detailed in this chapter. A test accuracy of 81.66% was achieved, which is a slight drop from the accuracy of this model measured on the fixed data split detailed in Chapter 3. The novel method proposed in this chapter achieves an overall test accuracy of 89.97%, when tested using the same evaluation framework. The model performance was assessed across a range of parameters with the results provided in Table XI. This Table compares the performance of the proposed method to

the baseline model developed in Chapter 3. The baseline model in Chapter 3 is tested on the evaluation framework proposed here in Chapter 4 for fair comparison between the two methods.

TABLE XI
COMPARISON OF MODEL PERFORMANCE METRICS FOR PROPOSED METHOD AND 2D BASELINE CNN MODEL

Metric	**2D CNN with Sonified EEG Spectrogram**	**2D CNN with EEG Mel-Spectrogram**
Precision	0.9079	0.7858
Recall	0.8994	0.7576
F1-Score	0.8985	0.7547
R-Squared	0.8507	0.7574
Cohen's Kappa Score	0.8219	0.5622

Performing the Wilcoxon signed-rank test [103] a p-value of 0.03125 was achieved, which is less than 0.05, thereby indicating that there is statisticaly significant difference between the two methods, with the novel method performing significantly better than that of the baseline method. The Wilcoxon signed-rank tests are recommended for comparison of two machine learning models due to their robustness and simplicity [103]. The proposed method shows a relative increase in performance of 10.18% over the baseline 2D CNN model. This highlights the benefit of the FM/AM transformation for enhancing the HIE features of the signal.

The proposed method shows very strong performance across all metrics, with the balance between sensitivity and specificity being well captured by the model based on the F1-score of 0.8985. The Cohen's Kappa score of 0.8219 further confirms the model's effectiveness, as it demonstrates a strong agreement beyond chance. The high precision score of 0.9079 reflects the model's strong performance when identifying true positives.

Comparing the proposed method to other autonoumous grading methods in the literature, there are a number of novelties in the method proposed here which should be highlighted. The first is the pre-processing phase. The proposed method uses the full 1 hour EEG segment as input to the model. Other work examined averaging the signal's features over the 1 hour window or segmenting the EEG signal into shorter windows of interest. In addition, late decision aggregation has been used in the literature for aggregating the window segments to output a single grade for each 1 hour segment.

By compressing the signal during the FM/AM transformation an improved frequency resolution across the full 1 hour window can be achieved at the input stage. This results in a optimal input to the model and allows a large window of interest to be seen, without overloading the computational demand of the model. In addition, the regression implementation of the proposed method showed higher performance over the classifcation task. By converting this problem to a regression task the relationships between the features can be clearly learnt. However, post-processing is needed to define the final grade. The proposed method shows a novel implementation of an optimised rounding function, which learns the relationships between the regression values and the true grades for the validation data. This can then be implemented on the test data to provide the final clinical grade. Given that different datasets and evaluation frameworks were used across the published literature, direct comparison between the methods stated in the literature and the method proposed in this thesis must be made with knowledge of the differences in frameworks.

A relative increase of 4.5% over the previous state-of-the-art, published by Yu et al [24], is achieved by the proposed method. This was tested on a large private dataset of 338 hours of EEG signal. In addition, a relative improvement of 1.2% is achieved over the leave-one-out performance, stated by Raurale et al, for CNN analysis of time-frequency representation of 5 minute segments of the EEG signal [31].

Based on this comparative analysis, the strength and importance of the proposed method for autonomous HIE grading is obvious, with the novel approach achieving superior performance across all metrics.

F. Conclusion

By enhancing the representation of EEG signals through innovative sonification methods and optimising the 2D CNN pipeline, the proposed method achieves a new state-of-the-art accuracy of 89.97% with a Cohen's Kappa score of 0.82, indicating a very strong agreement beyond chance. Nested cross-validation was implemented and so a robust and generalisable model was successfully developed, ensuring state-of-the-art performance was attained [96].

The sonification process highlighted the continuity and discontinuity of EEG signals, crucial features for differentiating HIE grades. With comparative analysis performed

between the original EEG signal and its audio domain representation, the benefits of the sonification algorithm were identified, increasing the differentiability potential of the signals for each HIE grade. By optimising the spectrogram parameters to balance the time and frequency resolution, an informative representation of the sonified signal could be provided to the CNN.

Analysis of the decreased performance of outer fold 4 revealed the presence of challenging and outlying samples, particularly for grades 2 and 3. Despite this, the model maintained a high level of accuracy, showing resilience in its predictions. The efficacy of the proposed method was further proven by the UMAP analysis, which confirmed the model's ability to separate grades effectively.

Overall, this chapter presents a novel and effective solution for HIE grading, addressing the limitations of existing approaches. Deployment of this model on an edge device is the next step in the design process, once again further pushing the bounds in the field of autonomous HIE grading with the first deployment on an edge device in the literature to date. The goal is to evaluate the potential performance when running locally on a Snapdragon processor, equivalent to those available in mobile phones.

5. Model Quantisation and Implementation on Edge Device

EVALUATION of the model performance on an edge device is crucial for the deployment of the model in a clinical setting and so this chapter discusses the work implemented to deploy the proposed method on the Qualcomm RB3 Gen 2 edge device [104]. The model was trained and evaluated originally on NVIDIA Quadro GV100 GPU [105] and so access to very high computational power and memory capacity was available throughout the training process. However, this typically is not a realistic environment or configuration for the deployment in a clinical setting. This is important given that the primary motivation for this study is to aid medical professionals in determining the HIE grade of the EEG signal, and so deployment on an edge device is necessary. Implementing the HIE grading on an edge device is also necessary to retrofit the existing technology seemlessly, providing smart AI capabilities to the already existing technologies and not taking up a large amount of space in an already crowded environment. The performance of the grading algorithm on the edge device is evaluated and detailed throughout this chapter for the novel method proposed in Chapter 4. In addition, the model is quantized to 8-bit integer precision in order analyse the full potential of the Qualcomm RB3 Gen 2 vision device for AI inference, the details of which are provided in this chapter.

A. Details of the Target Device

The Qualcomm Robotics RB3 Gen 2 Vision Kit was selected as the target device for the deployment of the model, due to its powerful AI processor [104]. The RB3 is based on the QCS6490 Qualcomm processor. It has significant AI processing power, with high inferences per second, while maintaining energy efficiency [104]. In addition, this processor is found in many mobile devices and so this device shows significant potential for implementation as a decision support tool for medical professionals which would integrate seemlessly into their current workflow.

The Adreno 643 GPU is used in the RB3, running at 812 MHz [106]. The RB3 also contains the Octa-Core CPU, with one ARM Cortex-A78 clocked at 2.7 GHz, three ARM Cortex-A78 clocked at 2.4 GHz and four ARM Cortex-A55 clocked at 1.8 GHz [106]. In addition to the GPU and CPU, the RB3 gen 2 is designed with an AI-specific Digital Signal Processor (AI-cDSP), which can execute 12.5 trillion operations per

second (TOPS) for 8-bit integer operations [104]. This is a key metric in AI inference efficiency . The AI-cDSP is highly optimized for running AI workloads, in particular for inference tasks. By implementing quantization to 8-bit integer precision, the model can take advantage of the DSP for faster and more efficient inference times, in comparison to utilising the CPU or GPU alone.

B. Model Deployment

Conversion of the Pytorch model to Deep Learning Container (DLC) format is required in order to run the model on the Qualcomm RB3 [107]. DLC format is designed to maximize performance on Qualcomm's processors, and ensures low memory usage and fast inference times. The Snapdragon Neural Processing Engine (SNPE) SDK is utilised in order to convert the model to the DLC format [107].

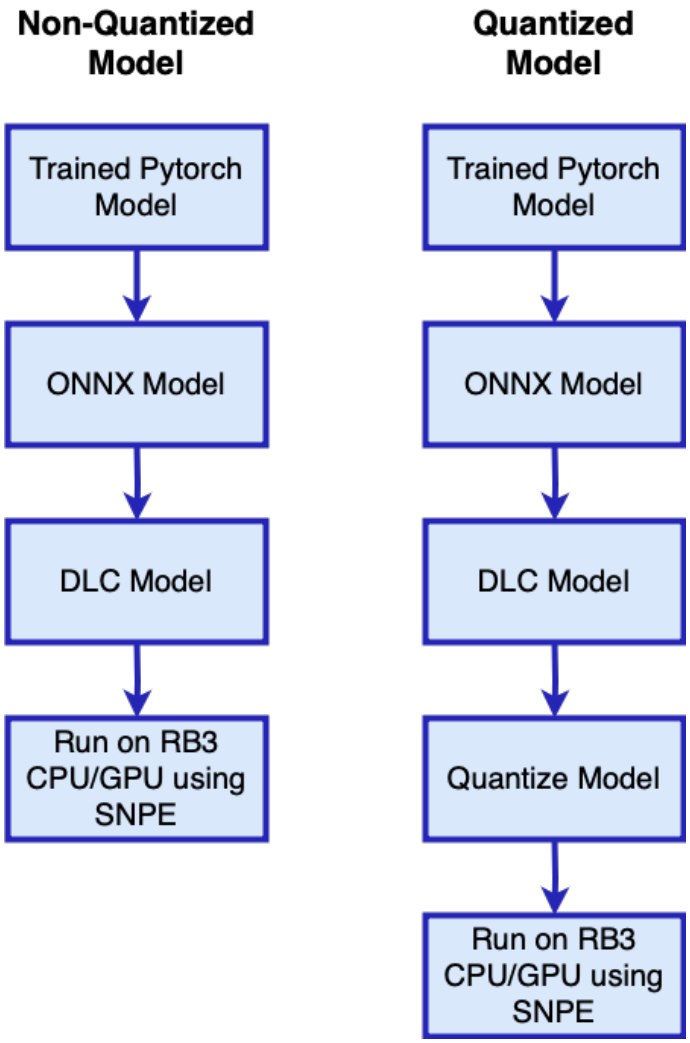


Fig. 37. Model Implementation and Deployment Workflow

In the conversion process of the model to DLC format, the SNPE library optimizes the computational graph and reduces unnecessary overhead in the model. In order to achieve this, the DLC converter often applies layer fusion, where separate layers are fused into a single operation, both reducing memory usage and improving the computational efficiency. Prior to applying DLC conversion, however, the model must first be converted to the Open Natural Network Exchange (ONNX) format, which is a standard format for representing deep learning models [108]. After conversion of the Pytorch model to ONNX, the model performance was tested to ensure no discrepancies were introduced nor that the model performance had decrease. The regression values of each model were assessed to evaluate whether any changes in the model weights or activations occurred, and, once this was evaluated and the response was satisfactory, the model was converted to the DLC format. Analysis of the model performance after DLC conversion was tested, with minor degradation in performance from 90.91% to 87.88%, due to the conversion process. An overview of the implementation workflow for both the non-quantized and the quantized model is provided in Fig. 37.

C. Model Quantisation

Automatic Mixed Precision (AMP) was then implemented to optimise the model training process. Mixed Precision (MP) is a technique used to reduce the memory footprint and increase the training speed of the model by using a combination of 16-bit (FP16) and 32-bit (FP32) floating point precision [109]. FP32 is the standard floating point precision used in deep learning models and is represented by 32-bits. FP16 represents each floating-point number by 16-bits and, while it significantly reduces the memory footprint, it often leads to numerical instability and underflow issues due to its reduced range, in comparison with FP32 single precision [109]. In addition to reducing the memory footprint of the model, the training and inference time of the model is dependent on the memory and computational bandwidth. Utilising FP16 reduces the number of bytes by 50% and, so, the model can be trained and evaluated faster [109]. This is of particular impact on NVIDIA GPUs which allow for eight times more arithmetic operations to be performed in parallel with FP16 in comparison to FP32 [109]. The NVIDIA Quadro GV100 GPU is used for all training throughout this study [105]. The benefits of half-precision are retained without the sacrifice of numerical

stability when using a mixed precision approach, thus the model performance is not compromised [110]. AMP dynamically determines the operations which can be reduced to FP16 and which operations are best left at FP32 single precision floating point in an automated process [111]. The precision for each operation is automatically handled by the autocast context manager of AMP, with the majority of operations set to FP16, while any sensitive operations remain at FP32. The loss value prior to backpropagation is scaled and, as a result, helps to prevent underflow and ensures no small gradient values become zero at FP16.

Given the use of AMP, and so the weights and activation functions of the model are either FP16 or FP32 (dependent on the sensitivity of the operation), the memory footprint of the model has been reduced in comparison to the single precision model. Although mixed precision for edge device implementation has been seen in recent literature to show very promising results [112], quantisation to 8-bit integer precision is essential for efficient use of the Snapdragon processor AI hardware.

Quantisation of the 2D CNN model was performed using Qualcomm's SNPE library [107]. The performance of the model decreases significantly after quantization, with the test accuracy dropping from 90.91% to 65.64%, and so clearly returning unsatisfactory results. This accuracy was calculated based on rounding the regression outputs using the optimised rounding thresholds for the full precision model, as described in Chapter 4. Retraining the Optimised Rounder on the validation outputs of the quantised model, resulted in increasing the test accuracy to 84.85%. This is approximately 2% below the state-of-the-art for single precision model, and so this is very satisfactory performance for the quantised model on edge hardware. The Optimized rounder was trained on the regression output values of the quantized model for both the training and validation dataset and then applied to the unseen test set. The performance for the training, validation and test accuracy are detailed in Table XII which shows the significant benefit of the optimized rounder has on the quantized model performance.

TABLE XII

QUANTIZED MODEL PERFORMANCE

Model	**Training Accuracy**	**Validation Accuracy**	**Test Accuracy**
Quantized Model	51.82%	51.82%	63.64%
Quantized Model with Optimized Rounder	85.45%	84.62%	84.85%

Identifying why the optimized rounder resulted in such a significant increase in the accuracy values of the model included an analysis of the regression values of the quantized model. The optimal thresholds calculated by the optimized rounder, based on the training and validation regression values of the quantized model, were 2.19945, 3.106767 and 3.42306. The violin plot is a combination of a box plot and density distribution plot. The median and interquartile range (IQR) are shown by the horizontal lines on the plot. The distribution of the data is then readable through the width of the plots. A wider area of the plot indicates a higher density data density for the respective vertical value, representing the predicted regression value here.

Looking at the violin plot for the regression values for each grade from the quantized model in Fig. 38, it is clear that the mean regression values for each grade are a lot closer together than the original model results. Therefore, standard rounding thresholds for this range of values grouped into four categories does not function well at splitting this regression data into its appropriate HIE grades.

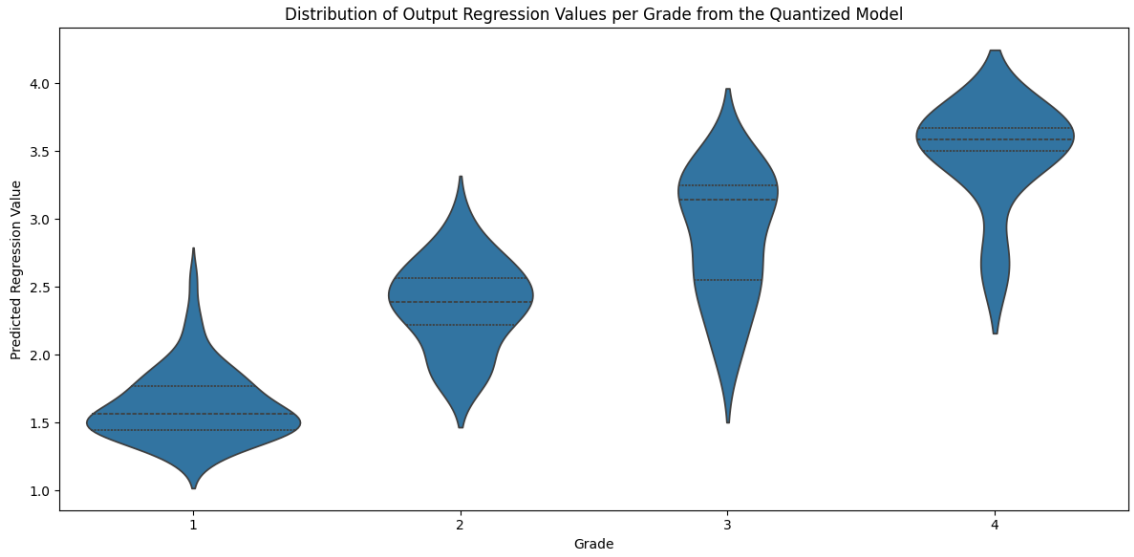


Fig. 38. Violin Plot of the Regression Values for the Quantized Model

The optimized rounder is able to identify the optimal thresholds for the regression values to be split into the four categories, and so the model performance is significantly improved. When quantizing the model, the weights and activations of the model are reduced from FP32 and FP16 precision to INT8. Analysing the output values of the quantized model showed that this reduction in precision caused a decrease in the range of output values. This increased the requirement for the optimized rounder to be applied in the post-processing step, to maintain satisfactory accuracy performance of the model.

D. Analysis of Model with AMP

AMP implementation was tested on the fixed data split model with the 8 channel vertically stacked spectrogram used as input to the model. This showed a significant improvement in the model speed and memory footprint, without sacrificing the model accuracy and loss performance. Table XIII below details performance with and without AMP applied.

TABLE XIII

TRAINING AND EVALUATION METRICS FOR COMPARISON OF AMP IMPLEMENTATION

AMP	Training Time	Inference Time	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Test Accuracy
Without AMP	23.29 s	0.29 ms	0.032	95.45%	0.077	92.31%	93.94%
With AMP	13.67 s	0.29 ms	0.017	98.18%	0.0662	96.15%	90.91%

The results of the model performance with AMP applied showed a slight decrease in the test accuracy performance from 93.94% to 90.91%. However, training time is reduced by 41.3%, from 23.29 seconds to 13.67 seconds. An improvement in the model memory footprint is also noted when AMP is applied, with the maximum memory for training being reduced from approximately 26 GB to 15 GB. This is expected due to the mixed precision approach reducing the required memory of the model to store intermediate activations, gradients and optimiser states. Stability of the memory usage across epochs was identified both with and without the use of AMP. This highlights the impact of AMP to reduce the training memory footprint of the model without introducing instability or excessive memory fluctuation while dynamically switching between FP16 and FP32 for different operations. The final memory of the model reduces from 549.05 MB to 547.79 MB when AMP is applied, which is a rather small change, indicating that

the major benefits of AMP for this model lie in reducing the intermediate memory usage during training as opposed to the static memory required to store the model weights. The combined benefits of improving both memory and computational efficiency make utilising AMP highly beneficial for this model.

E. Timing: Comparative Results

The goal of the developed model is that it be used as a post-acquisition tool for medical professionals, to aid in the diagnosis of HIE grade in neonatal patients. Consequently, real-time inference decision making is not a requirement of this study. The model is processing EEG data after the EEG signal has been aquired by the monitor for a sufficient duration, namely 1 hour, as this is the data used for this analysis. Therefore, an exceptionally fast inference time is not a requirement for successful deployment. However, inference time is still a key metric in analysing the model performance on the edge device. With a quicker response time for the decision, the higher the likelihood a medical professional will integrate this decision support tool into their clinical workflow. The ultimate goal is to design a solution that does not burden their current clinical workflow, but instead significantly enhances the information that they have available to them at the point of care.

Both the non-quantized and quantized models were evaluated based on timing metrics when deployed utilising the target device CPU, GPU. The quantized model was also evaluated on the DSP. The DSP is designed to run inference for 8-bit integer computations and so only the quantised model can be evaluated on the DSP. The inference time for each model on the RB3 was calculated based on the average inference time for all test samples which were stored locally on the target device. The inference times for both the non-quantized and quantized models are provided in Table XIV for each runtime configuration. It is important to note here that the pre-processing stage was not implemented on the RB3 device, and so the timing analysis excludes the time necessary to convert the EEG signal to the audio domain and generate the spectrogram image. This was due to the limitation of the RB3 device to run python scripts with the existing operating system. The goal of future work will aim to implement the pre-processing step on the snapdragon processer also.

TABLE XIV

INFERENCE TIME FOR NON-QUANTIZED AND QUANTIZED MODEL ON QUALCOMM RB3

Model	**CPU Inference Time (s)**	**GPU Inference Time (s)**	**DSP Inference Time (s)**
Non-Quantized Model	1.7405	0.4053	-
Quantized Model	1.7392	0.4051	0.062721

It is clear from Table XIV that the inference times utilising both the CPU and GPU separately for the non-quantized and quantized models, show no significant differences. This is expected given that the CPU and GPU hardware on the RB3 are not designed for optimal performance at 8-bit precision but, instead, are designed to handle FP32. The inference time utilising the CPU only was 1.74 seconds on average, which decreased to 0.405 seconds when the GPU was enabled. While the 1.74 seconds measured for the CPU alone is a relatively short time for a post-acquisition decision support tool, the improvement of 4.33 times when utilising the GPU is much more favourable.

An inference time of 62.721 ms was measured for the quantized model running on the DSP of the RB3 Gen2. This shows a significant improvement in the inference time in comparison to the CPU and GPU. The inference time of the model on the NVIDIA GPU the model was initially trained on was 0.29 ms, 216.24 times faster than when utilising the DSP on the RB3. This is to be expected as the NVIDIA GPU is designed for high performance computing and so is able to handle the model inference at a much faster rate than the DSP on the RB3. However, no clinical environment would have capacity to run inference on such a high performance GPU and so the DSP on the RB3 is a much more realistic environment for the model to deloyment. The inference time of 62.721 ms is still a very fast inference time for a model of this complexity and so is a very favourable result for the deployment of the model on the edge device. A breakdown of the execution statistics from the SNPE library for the quantized model implemented on the DSP is provided in Table XV.

TABLE XV

SNPE EXECUTE STATISTICS (AVERAGED)

SNPE Execute Statistics (Averaged)	
Total Inference Time:	62721 μs
Forward Propagate Time:	61321 μs
RPC Execute Time:	59426 μs
Snpe Accelerator Time:	57501 μs
Num times yield occurred:	0
Accelerator Time:	23247 μs

The forward propagation time is 97.76% of the inference time, and so stating that the majority of the inference time is spent on the forward propagation of the model and the remaining 2.24% of the inference time is spent on pre- and post-processing. This highlights that the computational load of the model is quite heavy, with the model complexity causing a delay in the inference time. Since the SNPE accelerator time is 57.501 ms, it is clear that the model is efficiently utilising the hardware of the target board. This highlights the benefit of quantizing the model to 8-bit integer precision for optimization of the model for the edge device.

The results of the throughput time of both the quantized and non-quantized models are detailed in XVI, where a batch size of 1 was utilised.

TABLE XVI

THROUGHPUT TIME FOR NON-QUANTIZED AND QUANTIZED MODEL ON QUALCOMM RB3

Model	**CPU Throughput Time (infs/s)**	**GPU Throughput Time (infs/s)**	**DSP Throughput Time (infs/s)**
Non-Quantized Model	0.5871	2.45943	-
Quantized Model	0.5864	2.45808	32.5476

The difference in throughput time for the CPU and GPU is negligible for the quantized and non-quantized models, as seen in Table XVI. The CPU has a throughput value of 0.587 infs/s and the GPU executes approximately 2.459 infs/s. The difference in throughput values is clearly noted when utilising the DSP in comparison to the CPU and GPU, with the DSP achieving a throughput of 32.5476 infs/s. This is 56.116 and 13.24 times faster than the CPU and GPU, respectively. This highlights the advantage associated with running the quantized model on the specialized hardware of the DSP

on the RB3 Gen2 device. The introduction of DSP execution in the quantized model enables substantial improvements in inference speed compared to using the CPU or GPU alone.

F. Memory Usage Analysis

Comparative analysis of the memory usage of both the non-quantized and quantized model on the RB3 Gen2 device, based on its configuration using the CPU, GPU and DSP runtime, was undertaken. The results of the memory usage analysis are detailed in Table XVII, where the Proportional Set Size (PSS) memory is provided for each model configuration.

TABLE XVII
MEMORY USAGE ON QUALCOMM RB3 GEN 2 VISION DEVELOPMENT KIT

Model	**CPU Memory Usage (MB)**	**GPU Memory Usage (MB)**	**DSP Memory Usage (MB)**
Non-Quantized Model	144.348	145.247	-
Quantized Model	110.906	161.009	292.453

A balanced memory footprint for the non-quantized model between the CPU and GPU during inference is seen in Table XVII. The non-quantized model maintains a similar average PSS memory on the CPU and GPU, with the same holding true for their peak memory usage at approximately 185,153 kB. The GPU's dirty memory, which is the memory allocated and written to by the application, averaged 65,599 kB with the clean memory averaging 79,086 kB. The clean memory is the memory shared across multiple processes. The CPU's dirty and clean memory averaged 64,728 kB and 79,058 kB, respectively.

The memory consumption for the quantized model was evaluated across the three compute units, and the results showed that a higher memory usage was recorded for the DSP runtime. This is primarily due to the memory allocation patterns typical of CNNs. The GPU can leverage shared memory and advanced memory management techniques which the DSP does not, which often results in allocation of larger intermediate buffers by the DSP to ensure performance optimization. Without this parallel threading capability and the optimized memory hierarchies of the GPU, the DSP relies on fixed memory allocation strategies and so the pre-allocation of a fixed amount results in a

larger memory footprint. However, this trade-off ensures performance optimization and deterministic execution, which is critical in real-time applications.

Comparing the results of the memory usage for the non-quantized and quantized models, a reduction in memory usage of 23.1% is achieved for the quantized model on the CPU. Despite the reduction of operations, from 32-bit and 16-bit in the mixed precision non-quantized model, to 8-bit integer precision in the quantized model, an increase in memory usage is noted as per Table XVII. An increase of 10.8% is seen which may be due to the overhead associated with managing quantized operations and, so, offsetting the memory saving by quantizing the model. Despite this increase in memory usage, the GPU benefits from faster execution times due to its optimized threading and memory hierarchy for the quantized model or the non-quantized model.

G. Conclusion

Quantization and implementation of the HIE grading algorithm highlighted the potential benefits of the proposed machine learning network detailed in this study. The model was successfully deployed on the Qualcomm Robotics RB3 Gen 2 Vision Kit, with the model quantized to 8-bit integer precision and implemented on the AI-cDSP. The model was able to achieve a test accuracy of 84.85% when the optimized rounder was applied to the quantized model. This is comparable with the current state-of-the-art without quantisation, with a relative decrease of only 1.44%. Implementation of the optimised rounder, with thresholds trained specifically for the quantised model,shows a significant improvement over the test accuracy of the model with the optimal rounding thresholds of the mixed precision model. This highlights the advantage of this novel post-processing step in enhancing the performance of the grading task, making it an essential tool in the machine learning pipeline of this task. Implementation of the quantized model resulted in large advantages, in terms of both the inference time and memory usage, for running the model inference. This is a key metric in the deployment of the model in a clinical setting, where the model must be able to run efficiently and effectively on the edge device. The inference time of the quantized model on the DSP was 62.721 ms, a significant improvement in comparison to the CPU and GPU runtimes, and so highlighting the advantage of the DSP hardware for running the model inference. To the best of the author's knowledge, this implementation is the only discussion of edge-device

AI for HIE grading in the literature to date and so provides a significant advancement in the field of autonmous HIE grading.

6. Conclusion and Future Work

THIS thesis describes the development of a novel approach to HIE grading in neonatal patients using advanced machine learning techniques and innovative signal processing methods. The primary objectives of this research were to analyse EEG signals to identify key features associated with HIE severity, develop effective grading algorithms, and evaluate the implementation of these models on edge devices for real-time diagnosis. This was achieved, with the successful exploration of two distinct approaches to automated HIE grading, namely a 2D CNN trained on Mel Spectrogram representations of EEG signals, and an innovative method that first converted EEG signals to the audio domain through sonification and then trained as 2D CNN on the spectrogram image of the audio signal. The second method was then deployed on a Qualcomm edge device to identify its potential for usage in a clinical setting, testing the performance on a Snapdragon processor. This research significantly contributes to the field by not only achieving a new state-of-the-art for HIE grading but, in addition, performed an in depth analysis of deployment of an automated HIE grading system on an edge device, both outcomes significantly advancing the current research in this field.

Previous findings have shown the potential for image based machine learning methods in the medical domain. This led to the motivation for using CNN models for automated HIE grading. The findings in this thesis demonstrate the significant potential of CNNs, with the CNN trained on the Mel Spectrogram images of the EEG signal achieving an accuracy of 84.85% on fixed-split unseen test dataset. This baseline model from Chapter 3, shows signs of overfitting to the train-validation set. While data augmentation did not enhance test performance, it narrowed the gap between train-validation and test results, which is not the expected outcome. Despite significant hyperparameter tuning and evaluating different learning schedulers no improvement was noted. Identifying the reasons for this behaviour might be an interesting avenue for future research.

Furthering this research, by converting the EEG signal to the audio domain using the FM/AM sonification algorithm, the extraction of rhythmic features relevant to each HIE grade was enhanced, thereby improving the overall test accuracy to 89.97%, with a 95% confidence interval of 89.15% to 90.73%. This marks a notable improvement over existing state-of-the-art methods, highlighting the benefits of novel signal processing techniques in clinical diagnosis. The motivation for the FM/AM transformation of the

EEG signal to the audio domain stemmed from the hypothesis that by transforming EEG signals to the audio domain using frequency and amplitude modulation, specific features of the signal relating to the grade of HIE are accentuated. This enhanced the ability to predict the HIE grade based on these representations. The increase in test accuracy for the proposed method over the Mel Spectrogram representation supports this statement, with a relative increase of over 10% for the same nested cross-validation frameworks. Further evaluation of the benefit of representing the EEG signal in the audio domain was assessed based on direct comparison of the UMAP analysis of the sonified EEG signal in comparison to the raw EEG signal. An improvement across all three metrics, namely silhouette score, DBI and CHI, further highlights the advantage of this data representation. Both temporal and spectral features associated with the HIE grade are unveiled when converting the EEG signal to the audio domain, by applying the FM/AM sonification algorithm, due to the EEG signal being the modulating signal for both the amplitude and frequency modulation. By enhancing the features for differentiability between the HIE grades the potential of auditory processing techniques to reveal subtle patterns in the data, which may be less apparent in traditional processing methods, is further supported. The best results of this proposed method, when the nested cross-validation evaluation framework is implemented achieved a test accuracy of 89.97%. This state-of-the-art result will improve early and accurate identification of HIE severity, which is critical for early intervention and tratement planning.

Beyond increasing the state-of-the-art accuracy for automated HIE grading, this work shows the deployment of the proposed method on a Qualcomm RB3 Gen2 edge device. An inference time of 62.7 milliseconds for the quantized model was achieved. This result proves the viability of using such models for real-time clinical decision support systems, a factor which has not yet been considered in the literature. This further underscores the benefit of this research for the field of HIE grading, highlighting the practical significance of this work in the healthcare setting where near-instantaneous decision-making is crucial in urgent neonatal care situations. This will allow for the medical professional to identify urgent cases rapidly, reducing the burden on these clinicians to quickly evaluate all HIE grading cases with an increased diagnostic confidence by combining the predictive output and the clinicans professional expertise. In addition, having fast diagnosis capabilities will allow for rapid treatment intervention, which is

crucial to improving the patient outcome. The ability to process EEG data efficiently on an edge device also minimizes reliance on cloud computing, reducing latency and ensuring data privacy, which is a critical concern in medical applications. By integrating such AI-driven models into existing neonatal monitoring systems, this work paves the way for more accessible, automated, and objective assessments, reducing inter-rater variability and aiding medical professionals in high-pressure environments

While the research in this thesis has contributed to the research in HIE grading using automated methods by both furthering the state-of-the-art performance and assessing the potential for edge device implementation, there are a number of avenues where this research can be further expanded.

- *Testing on Expansive Dataset:* Given that a significant limitation to this study is the size of the publicly available dataset, the research can be further developed by testing the proposed method's performance on a larger dataset. There are currently no such existing public datasets and so prior to this work, the data would have to first be acquired and annotated by trained medical professionals.
- *Optimization of Quantized Model Performance:* The proposed method showed a decrease in test accuracy to 84.85% for the quantized model. This was further reduced to 63.64% without the trained optimised rounder function used in the post-processing. This shows there is a possible potential for optimization in the quantization of the model weights and activation functions when reduced to 8-bit integer precision. This is a possible avenue for future work which would allow for further investigation into the integration of these models into real-world clinical settings.
- *Analysis of other Clinical Conditions:* Sonification of the EEG signal showed promising improvement for differentiating the different HIE grades. Future work could be expanded into the assessment on this method to analyse other brain conditions, where rhythmic patterns in the signal are highly indicative of the disease under evaluation. An example of the conditions could be sleep disorders, such as Sleep Apnea, where disruptions in the sleep cycles of the EEG waveform are present.

In conclusion, the research presented in this thesis significantly advances the field of automated HIE grading by combining novel signal processing and deep learning

techniques. The innovative method of highlighting the features of the HIE grades by representing the EEG signal based on the FM/AM transformation increased the ability to classify the HIE grades. This resulted in a new state-of-the-art accuracy of 89.97% for automated HIE grading. Furthermore, the successful deployment of the proposed method on edge devices, such as mobile platforms used by medical professionals, demonstrates its potential for real-time clinical use. This ease of integration into existing medical workflows underscores the practical benefits of this approach in improving early diagnosis and treatment outcomes for neonatal patients. This research successfully achieves the overall goal of the study by developing novel methods for HIE grading, that surpassed previous standards, which can lead to better patient outcomes.

REFERENCES

[1] P. Greco, G. Nencini, I. Piva, M. Scioscia, C. A. Volta, S. Spadaro, M. Neri, G. Bonaccorsi, F. Greco, I. Cocco, F. Sorrentino, F. D'Antonio, and L. Nappi, "Pathophysiology of hypoxic–ischemic encephalopathy: A review of the past and a view on the future," *Acta Neurologica Belgica*, vol. 120, no. 2, pp. 277–288, Apr. 2020.

[2] V. Riljak, J. Kraf, A. Daryanani, P. Jiruška, and J. Otáhal, "Pathophysiology of Perinatal Hypoxic-Ischemic Encephalopathy – Biomarkers, Animal Models and Treatment Perspectives," *Physiological Research*, pp. S533–S545, Dec. 2016.

[3] L. N., C. N., and M. B., "Immediate complications of hypoxic-ischemic encephalopathy in term neonates with resistive index as prognostic factor," *International Journal of Contemporary Pediatrics*, vol. 8, p. 711, Mar. 2021.

[4] J. S. Dumbuya, L. Chen, J.-Y. Wu, and B. Wang, "The role of G-CSF neuroprotective effects in neonatal hypoxic-ischemic encephalopathy (HIE): Current status," *Journal of Neuroinflammation*, vol. 18, no. 1, p. 55, Feb. 2021.

[5] D.-Y. Chen, I.-C. Lee, X.-A. Wang, and S.-H. Wong, "Early Biomarkers and Hearing Impairments in Patients with Neonatal Hypoxic–Ischemic Encephalopathy," *Diagnostics*, vol. 11, no. 11, p. 2056, Nov. 2021.

[6] C. J. Edmonds, S. K. Helps, D. Hart, A. Zatorska, N. Gupta, R. Cianfaglione, and B. Vollmer, "Minor neurological signs and behavioural function at age 2 years in neonatal hypoxic ischaemic encephalopathy (HIE)," *European Journal of Paediatric Neurology*, vol. 27, pp. 78–85, Jul. 2020.

[7] E. Okulu, İ. Hirfanoğlu, M. Satar, O. Erdeve, E. Koc, F. Ozlu, M. Gokce, D. Armangil, G. Tunç, N. Demirel, S. Unal, R. Ozdemir, M. F. Deveci, M. Akar, M. K. Demirel, M. Çetinkaya, H. S. C. Buker, B. Karagöl, D. Yaprak, A. Akcan, A. Anik, F. Narter, S. Arayici, E. Yildirim, I. M. Akin, O. Sahin, O. E. B. Ozdemir, F. Ovalı, M. A. Akin, Y. Çelik, A. Orman, S. Uslu, H. Ozkan, N. Koksal, A. Taştekin, M. Gunduz, A. Arısoy, R. Gurpinar, R. Ors, H. Altunhan, R. Keçeci, H. Yıldızdaş, D. Terek, M. Ates, Ş. Kader, M. Mutlu, K. Çelik, E. Yucesoy, M. K. Mert, S. Gulası, K. Kucuktasci, D. Arman, B. Hekimoğlu, N. D. Gultekin, H. Çelik, D. Kahvecioglu, C. Akyildiz, E. Taşkin, N. A. Çiftdemir, S. Uygun, T. Kaya, A. Akdağ, and A. Yilmaz, "An observational, multicenter, registry-based cohort study of Turkish Neonatal Society in neonates with Hypoxic ischemic encephalopathy," *PLOS ONE*, vol. 18, 2023.

[8] C. Vega-Del-Val, J. Arnaez, S. Caserío, E. P. Gutiérrez, M. Benito, L. Castañón, A. Garcia-Alix, and on behalf of the IC-HIE Study Group, "Temporal Trends in the Severity and Mortality of Neonatal Hypoxic-Ischemic Encephalopathy in the Era of Hypothermia," *Neonatology*, vol. 118, no. 6, pp. 685–692, 2021.

[9] A. C. Lee, N. Kozuki, H. Blencowe, T. Vos, A. Bahalim, G. L. Darmstadt, S. Niermeyer, M. Ellis, N. J. Robertson, S. Cousens, and J. E. Lawn, "Intrapartum-related neonatal encephalopathy incidence and impairment at regional and global levels for 2010 with trends from 1990," *Pediatric Research*, vol. 74, no. 1, pp. 50–72, Dec. 2013.

[10] Q. Wang, H. Lv, L. Lu, P. Ren, and L. Li, "Neonatal hypoxic–ischemic encephalopathy: Emerging therapeutic strategies based on pathophysiologic phases of the injury," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 32, no. 21, pp. 3685–3692, Nov. 2019.

[11] S. Ristovska, O. Stomnaroska, and D. Danilovski, "Hypoxic Ischemic Encephalopathy (HIE) in Term and Preterm Infants," *PRILOZI*, vol. 43, no. 1, pp. 77–84, Mar. 2022.

[12] T. Malai, V. Khuwuthyakorn, S. Kosarat, W. Tantiprabha, S. Manopunya, M. Pomrop, K. Katanyuwong, C. Saguensermsri, and N. Wiwattanadittakul, "Short-term outcome of perinatal hypoxic-ischaemic en-

cephalopathy at Chiang Mai University Hospital, Thailand: A 15-year retrospective study," *Paediatrics and International Child Health*, vol. 42, no. 3-4, pp. 109–116, Oct. 2022.

[13] W. Jia, X. Lei, W. Dong, and Q. Li, "Benefits of starting hypothermia treatment within 6 h vs. 6–12 h in newborns with moderate neonatal hypoxic-ischemic encephalopathy," *BMC Pediatrics*, vol. 18, no. 1, p. 50, Feb. 2018.

[14] "[PDF] Prognostic Tests in Term Neonates With Hypoxic-Ischemic Encephalopathy: A Systematic Review | Semantic Scholar," https://www.semanticscholar.org/paper/Prognostic-Tests-in-Term-Neonates-With-A-Systematic-Laerhoven-Haan/eed45b2aea6ce8b4d11457b59a09cf66122fc68c.

[15] J. M. O'Toole, S. R. Mathieson, S. A. Raurale, F. Magarelli, W. P. Marnane, G. Lightbody, and G. B. Boylan, "Neonatal EEG graded for severity of background abnormalities in hypoxic-ischaemic encephalopathy," *Scientific Data*, vol. 10, no. 1, p. 129, Mar. 2023.

[16] N. S. Abend and D. J. Licht, "Predicting outcome in children with hypoxic ischemic encephalopathy:," *Pediatric Critical Care Medicine*, vol. 8, no. 6, pp. 1–8, Nov. 2007.

[17] H. van Laerhoven, T. R. de Haan, M. Offringa, B. Post, and J. H. van der Lee, "Prognostic tests in term neonates with hypoxic-ischemic encephalopathy: A systematic review," *Pediatrics*, vol. 131, no. 1, pp. 88–98, Jan. 2013.

[18] M. Douglas-Escobar and M. D. Weiss, "Hypoxic-ischemic encephalopathy: A review for the clinician," *JAMA pediatrics*, vol. 169, no. 4, pp. 397–403, Apr. 2015.

[19] M. Thoresen, L. Hellström-Westas, X. Liu, and L. S. de Vries, "Effect of Hypothermia on Amplitude-Integrated Electroencephalogram in Infants With Asphyxia," *Pediatrics*, vol. 126, no. 1, pp. e131–e139, Jul. 2010.

[20] G. Boylan, L. Burgoyne, C. Moore, B. O'Flaherty, and J. Rennie, "An international survey of EEG use in the neonatal intensive care unit," *Acta Paediatrica (Oslo, Norway: 1992)*, vol. 99, no. 8, pp. 1150–1155, Aug. 2010.

[21] J. Brogger, T. Eichele, E. Aanestad, H. Olberg, I. Hjelland, and H. Aurlien, "Visual EEG reviewing times with SCORE EEG," *Clinical Neurophysiology Practice*, vol. 3, pp. 59–64, Jan. 2018.

[22] X. Wang, H. Liu, E. B. Ortigoza, S. Kota, Y. Liu, R. Zhang, and L. F. Chalak, "Feasibility of EEG Phase-Amplitude Coupling to Stratify Encephalopathy Severity in Neonatal HIE Using Short Time Window," *Brain Sciences*, vol. 12, no. 7, p. 854, Jul. 2022.

[23] A. Tuiskula, A. S. Pospelov, P. Nevalainen, S. Montazeri, M. Metsäranta, L. Haataja, N. Stevenson, A. Tokariev, and S. Vanhatalo, "Quantitative EEG features during the first day correlate to clinical outcome in perinatal asphyxia," *Pediatric Research*, pp. 1–7, May 2024.

[24] S. Yu, W. P. Marnane, G. B. Boylan, and G. Lightbody, "Neonatal Hypoxic-Ischemic Encephalopathy Grading from Multi-Channel EEG Time-Series Data Using a Fully Convolutional Neural Network," *Technologies*, vol. 11, no. 6, p. 151, Dec. 2023.

[25] E. Biagioni, L. Bartalena, A. Boldrini, R. Pieri, and G. Cioni, "Constantly discontinuous EEG patterns in full-term neonates with hypoxic-ischaemic encephalopathy," *Clinical Neurophysiology*, vol. 110, no. 9, pp. 1510–1515, Sep. 1999.

[26] L. Lacan, N. Betrouni, M.-D. Lamblin, L. Chaton, A. Delval, J.-L. Bourriez, L. Storme, P. Derambure, and S. NguyenThe Tich, "Quantitative approach to early neonatal EEG visual analysis in hypoxic-ischemic encephalopathy severity: Bridging the gap between eyes and machine," *Neurophysiologie Clinique*, vol. 51, no. 2, pp. 121–131, Mar. 2021.

[27] I. Korotchikova, N. J. Stevenson, B. H. Walsh, D. M. Murray, and G. B. Boylan, "Quantitative EEG analysis

in neonatal hypoxic ischaemic encephalopathy," *Clinical Neurophysiology*, vol. 122, no. 8, pp. 1671–1678, Aug. 2011.

[28] J. Guo, X. Cheng, and D. Wu, "Grading Method for Hypoxic-Ischemic Encephalopathy Based on Neonatal EEG," *Computer Modeling in Engineering & Sciences*, vol. 122, no. 2, pp. 721–742, Feb. 2020.

[29] N. J. Stevenson, I. Korotchikova, A. Temko, G. Lightbody, W. P. Marnane, and G. B. Boylan, "An automated system for grading EEG abnormality in term neonates with hypoxic-ischaemic encephalopathy," *Annals of Biomedical Engineering*, vol. 41, no. 4, pp. 775–785, Apr. 2013.

[30] R. Ahmed, A. Temko, W. Marnane, G. Lightbody, and G. Boylan, "Grading hypoxic–ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine," *Clinical Neurophysiology*, vol. 127, no. 1, pp. 297–309, Jan. 2016.

[31] S. A. Raurale, G. B. Boylan, S. R. Mathieson, W. P. Marnane, G. Lightbody, and J. M. O'Toole, "Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time–frequency distributions," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046007, Mar. 2021.

[32] A. L. Murray, D. S. O'Boyle, B. H. Walsh, and D. M. Murray, "Validation of a machine learning algorithm for identifying infants at risk of hypoxic ischaemic encephalopathy in a large unseen data set," *Archives of Disease in Childhood - Fetal and Neonatal Edition*, Sep. 2024.

[33] J. Huang, B. Chen, B. Yao, and W. He, "ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network," *IEEE Access*, vol. PP, pp. 1–1, Jul. 2019.

[34] G. Vrbancic and V. Podgorelec, "Automatic Classification of Motor Impairment Neural Disorders from EEG Signals Using Deep Convolutional Neural Networks," *Elektronika ir Elektrotechnika*, vol. 24, no. 4, pp. 3–7, Aug. 2018.

[35] B. Mandhouj, M. A. Cherni, and M. Sayadi, "An automated classification of EEG signals based on spectrogram and CNN for epilepsy diagnosis," *Analog Integrated Circuits and Signal Processing*, vol. 108, no. 1, pp. 101–110, Jul. 2021.

[36] P. Z. Yan, F. Wang, N. Kwok, B. B. Allen, S. Keros, and Z. Grinspan, "Automated spectrographic seizure detection using convolutional neural networks," *Seizure - European Journal of Epilepsy*, vol. 71, pp. 124–131, Oct. 2019.

[37] J. Hu, C. Wang, Q. Jia, Q. Bu, R. Sutcliffe, and J. Feng, "ScalingNet: Extracting features from raw EEG data for emotion recognition," *Neurocomputing*, vol. 463, pp. 177–184, Nov. 2021.

[38] M. N. A. Tawhid, S. Siuly, K. Wang, and H. Wang, "Automatic and Efficient Framework for Identifying Multiple Neurological Disorders From EEG Signals," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 76–86, Mar. 2023.

[39] J. Torres, J. D. Buldain Pérez, and J. R. Beltrán Blázquez, "Arrhythmia Detection Using Convolutional Neural Models," Jun. 2019, pp. 120–127.

[40] M. D. Le, V. Singh Rathour, Q. S. Truong, Q. Mai, P. Brijesh, and N. Le, "Multi-module Recurrent Convolutional Neural Network with Transformer Encoder for ECG Arrhythmia Classification," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. Athens, Greece: IEEE, Jul. 2021, pp. 1–5.

[41] S. Gomez, M. O'Sullivan, E. Popovici, S. Mathieson, G. Boylan, and A. Temko, "On sound-based interpretation of neonatal EEG," Jun. 2018.

[42] A. Väljamäe, T. Steffert, S. Holland, X. Marimon, R. Benitez, S. Mealla, A. Oliveira, and S. Jordà, "A review

of real-time EEG sonification research," in *International Conference on Auditory Display 2013 (ICAD 2013)*, Lodz, Poland, Jul. 2013, pp. 85–93.

[43] G. Baier, T. Hermann, and U. Stephani, "Event-based sonification of EEG rhythms in real time," *Clinical Neurophysiology*, vol. 118, no. 6, pp. 1377–1386, Jun. 2007.

[44] P. Loui, M. Koplin-Green, M. Frick, and M. Massone, "Rapidly Learned Identification of Epileptic Seizures from Sonified EEG," *Frontiers in Human Neuroscience*, vol. 8, p. 820, Oct. 2014.

[45] A. M. Pavel, J. M. Rennie, L. S. de Vries, M. Blennow, A. Foran, D. K. Shah, R. M. Pressler, O. Kapellou, E. M. Dempsey, S. R. Mathieson, E. Pavlidis, L. C. Weeke, V. Livingstone, D. M. Murray, W. P. Marnane, and G. B. Boylan, "Neonatal Seizure Management: Is the Timing of Treatment Critical?" *The Journal of Pediatrics*, vol. 243, pp. 61–68.e2, Apr. 2022.

[46] D. Murray, G. Boylan, A. C. Ryan, and S. Connolly, "Early EEG Findings in Hypoxic-Ischemic Encephalopathy Predict Outcomes at 2 Years," *Pediatrics*, vol. 124, pp. e459–67, Sep. 2009.

[47] S. Di Matteo, N. M. Viall, and L. Kepko, "Power Spectral Density Background Estimate and Signal Detection via the Multitaper Method," *Journal of Geophysical Research: Space Physics*, vol. 126, no. 2, p. e2020JA028748, 2021.

[48] G. Rivera-Lillo, E. A. Stamatakis, T. A. Bekinschtein, D. K. Menon, and S. Chennu, "Delta band activity contributes to the identification of command following in disorder of consciousness," *Scientific Reports*, vol. 11, no. 1, p. 16267, Aug. 2021.

[49] J. R. Castro Conde, D. González Barrios, C. González Campo, N. L. González González, B. Reyes Millán, and A. J. Sosa, "Visual and Quantitative Electroencephalographic Analysis in Healthy Term Neonates Within the First Six Hours and the Third Day of Life," *Pediatric Neurology*, vol. 77, pp. 54–60.e1, Dec. 2017.

[50] F. Pedregosa, F. Pedregosa, G. Varoquaux, G. Varoquaux, N. Org, A. Gramfort, A. Gramfort, V. Michel, V. Michel, L. Fr, B. Thirion, B. Thirion, O. Grisel, O. Grisel, M. Blondel, P. Prettenhofer, P. Prettenhofer, R. Weiss, V. Dubourg, V. Dubourg, J. Vanderplas, A. Passos, A. Tp, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*.

[51] S. Viznyuk, "Shannon's entropy revisited," Mar. 2015.

[52] "SelectKBest," https://scikit-learn/stable/modules/generated/sklearn.feature_selection.SelectKBest.html.

[53] "When to use the Bonferroni correction - PubMed," https://pubmed.ncbi.nlm.nih.gov/24697967/.

[54] "Autoregressive Estimation of Short Segment Spectra for Computerized EEG Analysis | IEEE Journals & Magazine | IEEE Xplore," https://ieeexplore.ieee.org/document/4121292.

[55] Y. Zhang, S. Yang, Y. Liu, Y. Zhang, B. Han, and F. Zhou, "Integration of 24 Feature Types to Accurately Detect and Predict Seizures Using Scalp EEG Signals," *Sensors*, vol. 18, no. 5, p. 1372, May 2018.

[56] S. Talebi, J. Waczak, B. A. Fernando, A. Sridhar, and D. J. Lary, "Data-Driven EEG Band Discovery with Decision Trees," *Sensors*, vol. 22, no. 8, p. 3048, Jan. 2022.

[57] S. Kubicki, W. M. Herrmann, K. Fichte, and G. Freund, "Reflections on the topics: EEG frequency bands and regulation of vigilance," *Pharmakopsychiatrie, Neuro-Psychopharmakologie*, vol. 12, no. 2, pp. 237–245, Mar. 1979.

[58] R. Cassani, M. Estarellas, R. San-Martin, F. J. Fraga, and T. H. Falk, "Systematic Review on Resting-State EEG for Alzheimer's Disease Diagnosis and Progression Assessment," *Disease Markers*, vol. 2018, no. 1, p. 5174815, 2018.

[59] J. J. Newson and T. C. Thiagarajan, "EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies," *Frontiers in Human Neuroscience*, vol. 12, Jan. 2019.

[60] M. Li and B.-L. Lu, "Emotion classification based on gamma-band EEG," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sep. 2009, pp. 1223–1226.

[61] F. Lotte, "A new feature and associated optimal spatial filter for EEG signal classification: Waveform Length," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov. 2012, pp. 1302–1305.

[62] "HMS - Harmful Brain Activity Classification," https://kaggle.com/competitions/hms-harmful-brain-activity-classification.

[63] J. Volkmann, S. Stevens, and E. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America*, vol. 8, pp. 208–208, Jun. 2005.

[64] "Mean Squared Error," in *The Concise Encyclopedia of Statistics*. New York, NY: Springer, 2008, pp. 337–339.

[65] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Oct. 2020.

[66] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[67] Y. Sasaki, "The truth of the F-measure," *Teach Tutor Mater*, Jan. 2007.

[68] N. R. Draper and H. Smith, "On Worthwhile Regressions, Big F's, and R2," in *Applied Regression Analysis*. John Wiley & Sons, Ltd, 1998, ch. 11, pp. 243–250.

[69] "A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960," https://journals.sagepub.com/doi/10.1177/001316446002000104.

[70] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," Mar. 2022.

[71] R. Wightman, N. Raw, A. Soare, A. Arora, C. Ha, C. Reich, F. Guan, J. Kaczmarzyk, MrT23, Mike, SeeFun, Contrastive, M. Rizin, Hyeongchan Kim, C. Kertész, Dushyant Mehta, G. Cucurull, Kushajveer Singh, Hankyul, Y. Tatsunami, A. Lavin, Juntang Zhuang, M. Hollemans, M. Rashad, S. Sameni, V. Shults, Lucain, X. Wang, Yonghye Kwon, and Y. Uchida, "Rwightman/pytorch-image-models: V0.8.10dev0 Release," Zenodo, Feb. 2023.

[72] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *ArXiv e-prints*, vol. 1809.06839, 2018.

[73] A. Temko, W. Marnane, G. Boylan, and G. Lightbody, "Clinical implementation of a neonatal seizure detection algorithm," *Decision Support Systems*, vol. 70, pp. 86–96, Feb. 2015.

[74] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020.

[75] J. Hamill, G. E. Caldwell, and T. R. Derrick, "Reconstructing Digital Signals Using Shannon's Sampling Theorem," *Journal of Applied Biomechanics*, vol. 13, no. 2, pp. 226–238, May 1997.

[76] "SPECTRAL AUDIO SIGNAL PROCESSING," https://ccrma.stanford.edu/~jos/sasp/.

[77] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[78] "Hamming Window," https://ccrma.stanford.edu/~jos/sasp/Hamming_Window.html.

[79] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," May 2017.

[80] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging Weights Leads to Wider Optima and Better Generalization," Feb. 2019.

[81] "PetFinder.my Adoption Prediction," https://kaggle.com/competitions/petfinder-adoption-prediction.

[82] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, Jan. 1965.

[83] S. El Omda and S. R. Sergent, "Standard Deviation," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024.

[84] S. H. K. M., "Standard Deviation," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer, 2011, pp. 1378–1379.

[85] S. T. S and C. Lingam, "Speaker based Language Independent Isolated Speech Recognition System," in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, Jan. 2015, pp. 1–7.

[86] T. L. Narayana, A. Venkatesh, B. R. Reddy, G. B. Babu, and U. M. S. Lohith, "A Novel Approach for Speech Recognition," *International Journal of Engineering Research*, vol. 4, no. 18, 2016.

[87] D. Bansal, N. Turk, and S. Mendiratta, "Automatic speech recognition by cuckoo search optimization based artificial neural network classifier," in *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, Oct. 2015, pp. 29–34.

[88] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and Music Signal Analysis in Python," in *Python in Science Conference*, Austin, Texas, 2015, pp. 18–24.

[89] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[90] K. Rao and M. k e, *Speech Recognition Using Articulatory and Excitation Source Features*, Jan. 2017.

[91] H. A. Feldman, N. Kaiser, and J. A. Peacock, "Power Spectrum Analysis of Three-Dimensional Redshift Surveys," *The Astrophysical Journal*, vol. 426, p. 23, May 1994.

[92] N. Ahmed, T. Natarajan, and K. Rao, "Discrete Cosine Transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[93] E. Gasca, J. S. Sánchez, and R. Alonso, "Eliminating redundancy and irrelevance using a new MLP-based feature selection method," *Pattern Recognition*, vol. 39, no. 2, pp. 313–315, Feb. 2006.

[94] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[95] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," Sep. 2020.

[96] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, Nov. 2019.

[97] E. A, Z. M, M. E, S. Y, D. H, S. H, and B. Me, "Impact of Hypoxia-Ischemia on Neurogenesis and Structural and Functional Outcomes in a Mild-Moderate Neonatal Hypoxia-Ischemia Brain Injury Model," *Life (Basel, Switzerland)*, vol. 12, no. 8, Jul. 2022.

[98] A. Hazra, "Using the confidence interval confidently," *Journal of Thoracic Disease*, vol. 9, no. 10, pp. 4125–4130, Oct. 2017.

[99] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 747–748.

[100] A. Singh, S. Mittal, P. Malhotra, and Y. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," Mar. 2020, pp. 306–310.

[101] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conference Series: Materials Science and Engineering*, vol. 569, no. 5, p. 052024, Jul. 2019.

[102] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," Jun. 2023.

[103] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[104] "RB3 Gen 2 Quick Start Guide," https://docs.qualcomm.com/bundle/publicresource/topics/80-70014-253/set_up_the_device.html.

[105] "NVIDIA QUADRO GV100 Datasheet," Jun. 2018.

[106] "Qualcomm® Application Processors Selector Guide."

[107] "Snapdragon Neural Processing Engine SDK," https://docs.qualcomm.com/bundle/publicresource/topics/80-63442-2/introduction.html.

[108] J. Bai, F. Lu, and K. Zhang, "ONNX: Open Neural Network Exchange," *Github Repository*, 2019.

[109] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed Precision Training," Feb. 2018.

[110] M. Rakka, M. E. Fouda, P. Khargonekar, and F. Kurdahi, "Mixed-Precision Neural Networks: A Survey," 2022.

[111] "Automatic Mixed Precision — PyTorch Tutorials 2.4.0+cu121 documentation," https://pytorch.org/tutorials/recipes/recipes/amp_recipe.html.

[112] M. Rakka, M. E. Fouda, P. Khargonekar, and F. Kurdahi, "A Review of State-of-the-Art Mixed-Precision Neural Network Frameworks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024.