Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork

Development of ECG-Based Machine Learning Methods for Early Detection of Neonatal Brain Injury

Thesis presented by

Kimia Rezaei, Electrical and Electronic Engineering

for the degree of

Doctor of Philosophy

University College Cork

School of Engineering

Head of School: Dr. Alan Morrison

Supervisors: Prof. Gordon Lightbody & Prof. William Marnane

January 2026

Abstract

Hypoxic-Ischemic Encephalopathy (HIE) is caused by inadequate blood flow and oxygen delivery to the brain, typically occurring around the time of birth and can lead to neurological impairments or death. HIE is the most common underlying cause of neonatal seizures, many of which are subclinical and difficult to detect. Early and accurate assessment of both HIE severity and seizure activity is essential in the neonatal intensive care unit (NICU) to support immediate diagnosis and treatment.

The increasing use of the Electroencephalogram (EEG) for supporting clinical decisions, such as initiating therapeutic hypothermia, seizure detection and prognostication, led to growing interest in developing automated systems to assist clinicians in grading HIE. Monitoring EEG requires specialist expertise and is resource-intensive, limiting timely diagnosis in many low-resource settings. The Electrocardiogram (ECG), however, is routinely monitored and is widely accessible.

This thesis investigates various approaches for the two tasks of HIE grading and seizure detection, utilizing a large ECG dataset from newborns. A traditional approach is implemented by deriving the heart rate (HR) from the ECG. The most effective HR variability (HRV) features are analysed, and the most powerful classification models are benchmarked.

In contrast, two new approaches are proposed. By eliminating the ECG preprocessing step and leveraging more robust deep learning models, computational overhead can be reduced, enabling real-time processing, and enhancing the practicality of HIE grading and seizure detection for clinical use. The first method introduced spectrogram calculated directly from the raw ECG data. Lightweight

convolutional neural network (CNN) architectures are designed, tailored to the input spectrogram to effectively capture hierarchical representations through progressively expanding receptive fields.

The second method proposes trend-augmented multiscale tokenization transformer (TMFormer), a novel transformer-based model that introduces a tokenization approach derived from raw ECG signal attributes, capturing spatial patterns at multiple scales and tracking the signal's long-term progression. Since seizure detection likely depends on short ECG segments, where derived representations such as spectrograms capture more informative patterns than raw signals, we developed a vision transformer with convolutional patch embedding (CNN–ViT). HIE grading results represent a promising step toward developing HIE assessment tools to support clinicians in timely decision-making for infant treatments even in resource-limited hospitals, However, Seizure detection remains challenging.

Acknowledgements

I would like to sincerely thank my supervisors, Prof. Liam Marnane and Prof. Gordon Lightbody, for giving me the opportunity to pursue this PhD. I am very grateful for their constant support throughout these years. Their technical expertise, their way of encouraging me to question every problem more deeply, and their consistent positivity helped me grow both personally and professionally.

I would like to thank Science Foundation Ireland for supporting my scholarship, as well as the INFANT Research Centre and University College Cork. I am also grateful to Prof. Geraldine Boylan and Dr. Sean Mathieson for their insightful feedback and assistance throughout my research. A special thanks to Jerry for his help with IT matters over the years. I am sincerely grateful to everyone working in the Electrical Engineering Building at UCC, in all capacities, for their support and kindness.

To my teammates, Shuwen and Kaiyu, thank you for making this PhD journey so enjoyable; without you, it would not have been the same. I especially appreciate the support you gave me when I first moved to Ireland and began my studies. I am also grateful to all the postgraduates I met throughout these years for the wonderful moments we shared. And to my best friend and housemate, Andriana, thank you for being by my side and for all the fun times we had together.

To my mom and dad, I can never thank you enough for your endless love and support over the years. To my sister Dena, always just a phone call away, making sure I never felt alone no matter the distance. And finally, to my wonderful husband Alex, meeting him at the beginning of my PhD journey was a gift, thank you for your constant support and for making this whole PhD journey feel lighter, warmer, and so much more lovely.

Declaration

I confirm that the work submitted is my own and has not been presented for another degree at University College Cork or elsewhere. All external sources are clearly cited. I have read and understood the University College Cork regulations on plagiarism and intellectual property.

Kimia Rezaei

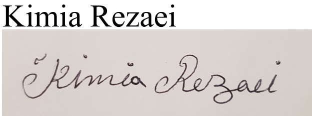


Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANSeR . . . . Algorithm for Neonatal Seizure Recognition

aEEG . . . . Amplitude-Integrated Electroencephalogram

AUC . . . . Area Under Curve

cEEG . . . . Continuous Electroencephalogram

CLS . . . . Class

CNN . . . . Convolutional neural network

CV . . . . Cross-Validation

ECG . . . . Electrocardiogram

EEG . . . . Electroencephalogram

FFN . . . . Feed-Forward Network

FN . . . . False Negative

FP . . . . False Positive

GANs . . . . Generative Adversarial Networks

HF . . . . High Frequency

HIE . . . . Hypoxic-Ischemic Encephalopathy

HR . . . . Heart Rate

HRV . . . . Heart Rate Variability

IBI . . . . Interburst Interval

SWC . . . . Sleep-Wake Cycling

LF .... Low Frequency

LOO .... Leave One Out

MCC .... Matthews Correlation Coefficient

MHSA .... Multi-Head Self-Attention

MLP .... Multi-layer perceptron

MRI .... Magnetic Resonance Imaging

NICU .... Neonatal Intensive Care Unit

NN .... Normalized RR interval

PSD .... Power Spectral Density

RBF .... Radial Basis Function

RF .... Random Forest

ROC .... Receiver Operating Curve RR

RR .... peak interval

SD .... Standard Deviation

SVM .... Support Vector Machine

tAPE .... time Absolute Position Encoding

TH .... Therapeutic Hypothermia

TINN .... Triangular Interpolation of NN interval

TN .... True Negative

TP .... True Positive

ViT .... Vision Transformer

VLF . . . . Very Low Frequency

XGBoost . . . . . . eXtreme Gradient Boosting

1 | Introduction

1.1 Introduction

To give every child the chance to survive and thrive, improving care during birth and the first week of life is critical. However, coverage of life-saving interventions remains uneven, and access to quality health services, such as essential neonate care is often limited [?]. According to the World Health Organization (2025) [?], neonate deaths represent 47% of all deaths among children under five, accounting for 2.4 million lives lost each year. Approximately one third of these deaths occur on the day of birth, and nearly three quarters take place within the first week of life. Most neonatal deaths are linked to inadequate quality of care at birth or insufficient skilled treatment immediately afterward, with the vast majority occurring in low and middle-income countries.

Neonatal encephalopathy is a clinical syndrome of neurological dysfunction in term and near-term infants. It is characterized by altered consciousness and abnormal tone and reflexes. When it is caused by a perinatal hypoxic-ischemic event, it is termed Hypoxic ischaemic encephalopathy (HIE). HIE is a leading cause of neonatal mortality and long-term complications. HIE can result in a variety of neurological and developmental conditions, including epilepsy, cerebral palsy, visual and hearing impairments, learning disabilities, and behavioral and emotional disorders such as attention-related difficulties and depression [?]. HIE is clinically graded as normal, mild, moderate, or severe shortly after birth using Sarnat scoring [?] [?]. Neonates with moderate to severe HIE can be treated with therapeutic hypothermia (TH), which improves neurological outcomes and reduces mortality in many infants [?]. TH is a time sensitive treatment in which the infant is cooled within six hours of birth to a target core body temperature of 33.5°C for 72 hours. After this period, the infant is gradually rewarmed to normal body temperature over 7 to 14 hours. Early and accurate assessment of HIE severity is essential in the Neonatal Intensive Care Unit (NICU) to support immediate diagnosis and treatment.

HIE is the most common underlying cause of seizures in the neonatal period. In affected infants, seizures typically begin within hours after birth, and their severity often reflects the extent of brain injury. Timely detection of seizures allows clinicians to identify significant problems early and enables NICU staff to start treatments promptly. A substantial proportion of neonatal seizures are subclinical, occurring without obvious clinical signs, meaning that infants may experience seizure activity without visible symptoms [**?**].

Electroencephalography (EEG) is the recording and monitoring of the brain's electrical activity. Its interpretation has several important diagnostic and therapeutic applications. EEG can detect focal abnormalities caused by primary brain lesions or secondary injuries, such as ischemia. It also allows analysis of background brain activity and its response to stimuli, which is useful for prognosis. Additionally, EEG can be used to assess the effects of medications on brain function [**?**]. Currently, continuous electroencephalography (cEEG) or amplitude-integrated EEG (aEEG) is commonly used to help assess HIE grade and monitor seizures. However, EEG monitoring presents significant challenges in the neonatal setting [**?**]. Interpreting neonatal EEG requires specialized expertise, which may not always be available in the NICU, and the high cost and complexity of EEG interpretation in the immediate postnatal period can hinder timely diagnosis. These combined limitations have driven ongoing research into automated and accessible methods for HIE grading and seizure detection.

During HIE, altered blood flow is hypothesized to affect the cardiac signal. Therefore, one potential alternative to EEG is the electrocardiogram (ECG), which is widely accessible, routinely monitored in neonates, and feasible even in smaller neonatal units and resource-limited settings. ECG alterations are less readily interpretable by neurophysiologists and typically require computational approaches, such as machine learning, to extract relevant information. Through analysis and processing of ECG signals, valuable diagnostic insights can be obtained. However, research on ECG-based HIE grading and seizure detection remains limited, highlighting the need for further investigation in this area.

1.2 Aims and contributions of this thesis

This thesis investigates the use of ECG signals as a more accessible and cost-effective alternative to EEG for HIE grading regarding requirement for TH and seizure detection in real-world clinical settings. The research aims to identify the most effective methodologies for this purpose, with particular attention to practical feasibility and clinical translation. The main contributions are as follows:

- ECG-based HIE and seizure detection: This work investigates the feasibility of using ECG signals as an alternative to EEG for grading HIE and detecting seizures in neonates.
- Improved accessibility and cost-effectiveness: By leveraging ECG, which is widely available in NICUs, this study addresses the high cost and limited accessibility of EEG, enabling more scalable neonatal monitoring.
- Large-scale neonatal dataset: Experiments were conducted on a substantial, multi-center neonatal dataset collected shortly after birth. The scale and diversity of this dataset add a unique dimension to the study and strengthen the generalizability of the findings.
- Novel signal representations: Multiple representations of the ECG were explored, and, for the first time, raw ECG waveforms were directly employed

for both seizure detection and HIE grading regarding requirement for TH. This approach enables real-time ECG analysis, even under noisy conditions.

- Advanced model design: A variety of machine learning approaches were benchmarked, culminating in the development of a tailored transformer-based model. Designed specifically to match the temporal and morphological characteristics of ECG signals, this model achieved state-of-the-art performance and outperformed traditional methods.

- Efficient Lightweight Deployment: The proposed lightweight model reduces computational overhead by minimizing the number of processing steps. As a result, the model is suitable for deployment on small devices such as smartphones, making it applicable in resource-limited hospital settings.

- Clinical interpretability: The model provides explainability techniques that help relate ECG patterns to established clinical markers.

- Validation and clinical relevance: The model's predictions were reviewed by clinical experts, highlighting areas of uncertainty in classification and reflecting real-world challenges in diagnosis.

- Comprehensive visualization and analysis: Results were presented across long-term recordings, providing insights into system behavior and clarifying the limitations of different methods in practical applications.

1.3 Thesis layout

The focus of this thesis is on two key aspects of neonatal health: HIE grading and seizure detection, both analyzed using ECG signals. For each task, three main investigations were conducted. The structure of this thesis reflects the dual focus of the work and is organized as follows:

Chapter 2 presents the necessary medical background for the work carried out in this thesis. It provides a brief explanation of HIE and seizures to highlight the importance of their treatment. Therapeutic hypothermia, which is used for neonatal treatment, is also explained, emphasizing the need for automated HIE grading and seizure detection to support clinicians. EEG signals and their application in HIE and seizure assessment are explained. ECG signals are discussed as a potential alternative to EEG. A review of previously developed automatic neonatal HIE grading and seizure detection algorithms based on both EEG and ECG is presented. A brief overview of the approaches that will be employed in this thesis is provided. Finally, the chapter provides details about the different datasets and their annotations used in this work.

Chapter 3 explains the process of extracting heart rate from EEG signals and the computation of features derived from heart rate variability (HRV). It discusses the relevance and effectiveness of these features for HIE grading and seizure detection using various machine learning models. Machine learning workflow, including data preprocessing, model selection, hyperparameter optimization, and performance evaluation techniques are explained. It investigates the practicality of these methods for clinical use, leveraging a large and diverse dataset (ANSeR) to assess their real-world applicability.

Chapter 4 describes the computation of ECG signal spectrograms and their usefulness for ECG signal analysis. It introduces Convolutional Neural Networks (CNNs) and explains the different CNN architectures designed in this work for HIE grading and seizure detection. The algorithms were optimized by exploring various spectrogram configurations, and data augmentation techniques were proposed to address class imbalance in the seizure detection dataset. Finally, the developed algorithms were applied to the ANSeR dataset for both tasks, and their performances were thoroughly evaluated.

Chapter 5 introduces transformer models and outlines their underlying principles. A transformer-based model is then proposed to process raw ECG signals directly, enabling real-time analysis and enhancing the practicality of the approach for clinical use. Additionally, the transformer model information flow from the input ECG signal is explained and visualized over the corresponding NN intervals. The proposed model is evaluated against traditional HRV-based approaches and CNN models using spectrograms for the HIE grading task. The performance and processing times of all methods are compared. Model outputs were reviewed by a clinical expert, who highlighted areas of uncertainty in classification, reflecting real-world challenges in diagnosis. Additionally, model predictions were visualized across entire recordings of up to three days. Finally, all models were combined in an ensemble to assess their complementary strengths and the overall impact on performance.

Chapter 6 provides a concluding summary of this work, reviewing the key findings from each chapter. It also outlines the main limitations of the study and discusses potential directions for future research.

1.4 Publication arising from this work

Journal publications Rezaei, K., Mathieson, S. R., Lightbody, G., Boylan, G. B., & Marnane, W. P. (2026, February). A Transformer-Based Model to Support Clinical Decision-making regarding Therapeutic Hypothermia Treatment of Infants with Hypoxic-Ischemic Encephalopathy. IEEE Access.

Conference publications Rezaei, K., Mathieson, S. R., Lightbody, G., Boylan, G. B., & Marnane, W. P. (2026, March). Investigating the Applicability of ECG Signals for Neonatal Seizure Detection Using ECG Spectrograms and CNN-ViT. Biosignals Conference.

Rezaei, K., Mathieson, S. R., Lightbody, G., Boylan, G. B., & Marnane, W. P. (2025, July). A Deep Learning Approach to Grading Neonatal Hypoxic-Ischemic Encephalopathy Using ECG Spectrograms. In 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-4). IEEE.

Rezaei, K., Yu, K., Mathieson, S. R., Flynn, A., Lightbody, G., Boylan, G. B., & Marnane, W. P. (2024, July). Assessing the Effectiveness of Heart Rate Variability as A Diagnostic Tool for Brain Injuries in Infants. In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1-4). IEEE.

2 | Automated HIE Grading and Seizure Detection

2.1 Introduction

In this chapter, essential medical background information relevant to this research is presented. First, key clinical concepts such as hypoxic–ischemic encephalopathy and seizures are introduced. The use of EEG signals for HIE grading and seizure detection is then discussed, followed by an overview of ECG signals as a potential alternative to EEG. Finally, therapeutic hypothermia, an established treatment for HIE, is described. The aim is to provide the reader with the necessary clinical context for understanding the problem domain explored in this work.

Furthermore, previously developed methods for automated neonatal HIE grading and seizure detection using both EEG and ECG are reviewed. In addition, a brief overview of the approaches which will be employed in this thesis for automatic HIE grading and seizure detection using ECG as an alternative to EEG is provided.

Finally, the chapter concludes with an explanation of the datasets and annotations employed in this thesis.

2.2 Hypoxic-Ischaemic Encephalopathy

Neonatal encephalopathy (NE) refers to dysfunction of the central nervous system occurring during the neonatal period [**?**]. It can arise from various causes, including hypoxia-ischaemia (HI), infections, and inborn metabolic disorders. HI is due to inadequate blood flow and oxygen delivery to the brain and other tissues and can occur before (antepartum), during (intrapartum), or after (postnatal) birth. An encephalopathy that develops as a result of HI is referred to as hypoxic-ischaemic encephalopathy (HIE) [**?**]. HIE is a global health concern and a leading cause of neonatal mortality and long-term morbidity. Its incidence varies significantly, ranging from 1.5–3 cases per 1,000 live births in high-income countries to as high as 14.9 per 1,000 in low-income settings [?]. Survivors are at risk of developing lifelong disabilities, including cerebral palsy, epilepsy, sensorineural hearing loss, visual impairment, and developmental delays.

2.3 Seizure

Seizures arise from excessive neuronal depolarization in the central nervous system, leading to a widespread, synchronized electrical discharge [?]. Within the neonatal period there is increased risk of seizures, which affects 1-5% of full-term infants. HIE is the most common underlying injury when seizures occur in the neonatal period, contributing to up to 53% of all neonatal seizures. In infants with HIE, seizures typically begin within hours after birth, with their severity reflecting the extent of underlying brain injury. They often subside within hours to days, regardless of treatment [?]. The impact of seizures on brain damage in HIE remains complex and poorly understood, partly because it is difficult to distinguish the direct effects of seizures from the injury already caused by hypoxia–ischaemia. Limited clinical evidence suggests that seizures may worsen brain injury in this context, emphasizing the importance of timely treatment of neonatal seizures.

2.4 EEG signal

The electroencephalogram (EEG) records the spontaneous extracellular electrical activity of the brain, generated by the action potentials of neurons [?]. Electrical potentials are detected through electrodes positioned on the scalp's surface. The recorded signals represent the combined synaptic activity of neurons within a particular brain region. Because the skull and soft tissues attenuate these signals, only the synchronized activity of large groups of neurons produces potentials strong enough to be measured at the scalp. More detailed information regarding EEG recording is provided in appendix.

EEG analysis emphasizes an overall assessment of the recording, taking into account the neonate's state of alertness and age-related patterns [?]. Early and accurate assessment of HIE severity and seizure is essential in the Neonatal Intensive Care Unit (NICU) to support immediate diagnosis and treatment. However, EEG monitoring presents significant challenges in the neonatal setting [?]. The execution and interpretation of neonatal EEG requires specialized expertise, which may not always be available in the NICU, and the high cost and complexity of EEG interpretation in the immediate postnatal period can hinder timely diagnosis.

2.4.1 HIE grading using EEG signal

Outcomes for infants with HIE depends on the severity of injury, which can vary from mild to severe, depending on the nature and timing of brain injury. Brain injury following hypoxic-ischemic injury evolves over hours to days and is reflected in changes in clinical and electroencephalography (EEG) or amplitude integrated EEG (aEEG) features [?]. HIE is clinically graded as normal, mild, moderate, or severe shortly after birth using Sarnat scoring [?] [?]. The Levene-modified Sarnat staging system is commonly used for grading HIE, but its accuracy is limited in the early postnatal period, as the most severe grade often becomes apparent only after 48 hours, too late for timely treatment [?]. Clinical assessment is further complicated by sedative and anti-epileptic medications [?]. The Thompson scoring system offers easier application and good predictive value, but its peak score, most reliable for outcome prediction, usually occurs on days 3–4, restricting its usefulness for early monitoring [?]. Additionally, unlike the Sarnat system, it has not been correlated with EEG grades of HIE. On the other hand continuous

Table 2.1: Classification of EEG Background Activity. IBI indicates interburst interval and SWC indicates Sleep-Wake Cycling.

Grade	Description
0	Normal: Continuous background pattern with normal physiologic features such as anterior slow waves
1	Normal/mild abnormalities: Continuous background pattern with slightly abnormal activity (e.g., mild asymmetry, mild voltage depression, or poorly defined SWC)
2	Moderate abnormalities: Discontinuous activity with IBI of <10 s, no clear SWC, or clear asymmetry or asynchrony
3	Major abnormalities: Discontinuous activity with IBI of 10–60 s, severe attenuation of background patterns, or no SWC
4	Inactive: Background activity of <10 μV or severe discontinuity with IBI of >60 s

multichannel EEG have been successfully used for early prediction of long-term neurodevelopmental outcomes in HIE [**?**]. Unlike clinical assessments, EEG can reveal the most severe grade within the first 6 hours after birth and enables continuous monitoring of encephalopathy progression. Grading is based on visual evaluation of background activity, considering factors such as signal continuity, interhemispheric symmetry, amplitude, frequency, and the presence of sleep–wake cycling.

Background EEG can be graded for severity, such as the scheme proposed by Murray et al. [**?**], EEG background is graded to 5 grades of normal, normal/mild, moderate, major and inactive as shown in Table 2.1.

Fig. 2.1 shows examples of EEG signal segments of the mentioned 4 grades of HIE.

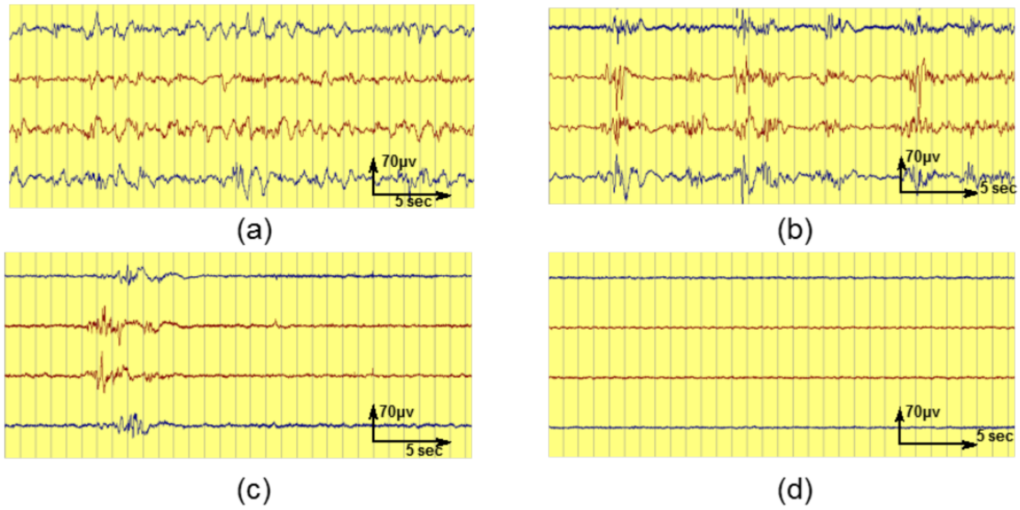


Figure 2.1: Examples of EEG waveforms in the 4 grades of HIE. (a) mild (b) moderate, (c) major, (d) inactive [**?**].

2.4.2 Seizure detection using EEG signal

The clinical presentation of seizures in neonates differs from that seen in children and adults. In HIE, most neonatal seizures are subclinical, showing no obvious motor signs. The distinct presentation of seizures in neonates is attributed to developmental, structural, and neurochemical variations [**?**]. Identifying seizures in the NICU is a crucial part of diagnosing brain injury. Neonatal seizures are difficult to detect, as only about a third show clinical signs. EEG is the gold standard for detection but requires specialist interpretation, while aEEG offers a simpler, more accessible alternative for non-specialists, though with reduced sensitivity, particularly for short or low-amplitude seizures. Neonatal seizures on EEG are defined as repetitive, evolving patterns exceeding 2 µV and lasting over 10 seconds. They appear as structured, rhythmic deviations from the normally pseudo-random background and can evolve in frequency, amplitude, and spatial distribution.

Seizure patterns on the EEG evolve and change over time. Neonatal seizures may be focal, multifocal, or generalized, meaning they can originate in a single brain

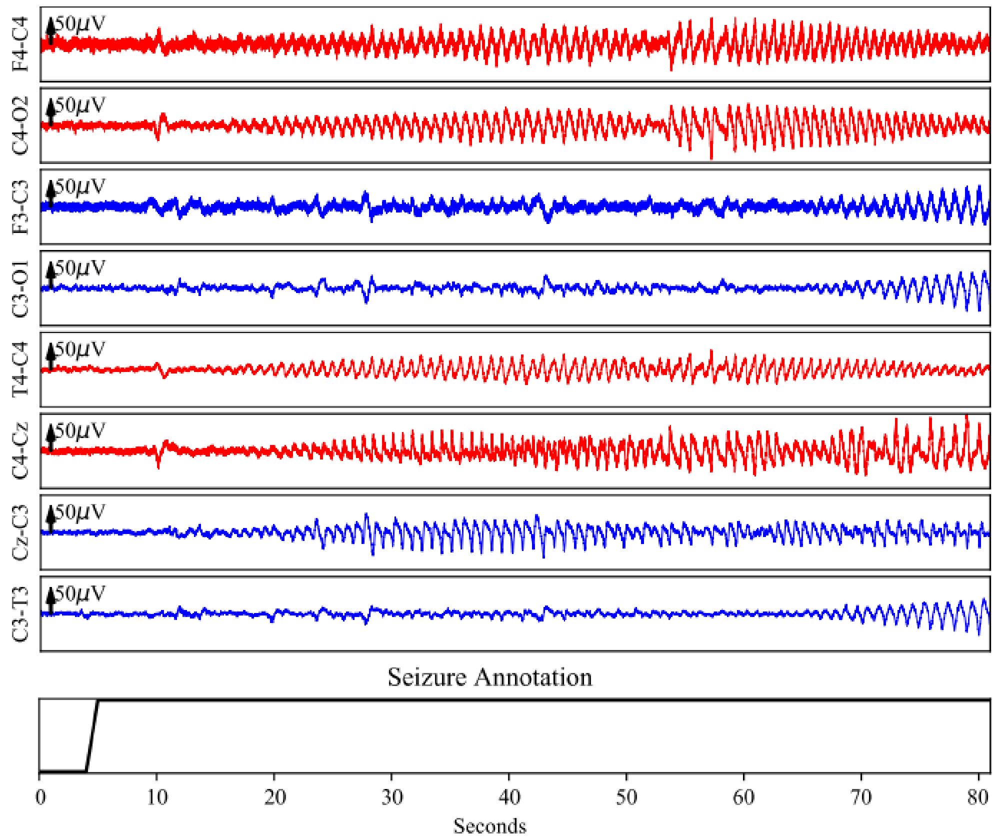


Figure 2.2: A seizure migrating across EEG channels. The red and blue traces represent activity from the right and left hemispheres of the brain, respectively. The repetitive seizure pattern originates in the right hemisphere and gradually spreads to the left over time. [**?**].

region, occur in multiple areas simultaneously, or spread from one region to another. Consequently, seizures may appear in one or several EEG channels. Figure 2.2 illustrates a seizure event that migrates across channels, beginning in the right hemisphere (red channels) and progressing to the left hemisphere (blue channels). This event represents a clear deviation from the background pattern, characterized by repetitive, high-energy activity [**?**].

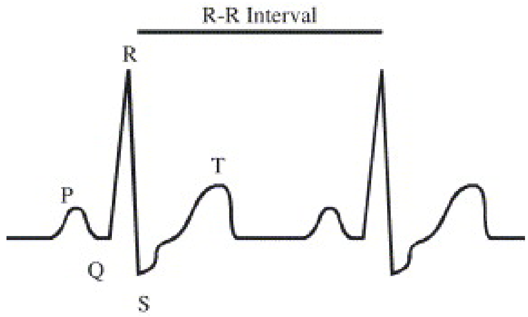


Figure 2.3: Ideal ECG signal: This figure depicts two idealized heart-beats. The R–R interval indicates the duration of a heartbeat. The major ECG complexes comprising one heartbeat are indicated by P, QRS, and T [**?**].

2.5 ECG signal

The ECG signal captures variations in electrical potential over time. Each heart-beat trace is characterized by three main complexes P, R, and T. Each identified by its peak, known as the fiducial point as shown in Fig. 2.3 [**?**].

Typically, ECG recordings in neonates with HIE are obtained using two electrodes placed on the right and left shoulders, connected to a bipolar channel on the EEG headbox. In this configuration, the right electrode served as the reference for the left, ensuring that no head electrodes were involved and that EEG signals did not contaminate the ECG trace.

ECG recordings are often affected by various sources of noise and artifacts. The purpose of preprocessing is to minimize these disturbances, accurately identify fiducial points (P, Q, R, S, T, including P-onset, P-peak, P-offset, QRS-onset, QRS-offset, T-onset, T-peak, and T-offset), and eliminate amplitude or offset variations so that signals from different patients can be compared reliably. Common noise types can be grouped as follows [**?**] [**?**]:

- Power line interference – a narrow-band signal at 50 or 60 Hz with a band-width below 1 Hz.

- Baseline wander – low-frequency noise (0.15–0.3 Hz) typically caused by respiration, leading to baseline shifts in the ECG trace.
- Electrode contact noise – generated when the electrode loses proper contact with the skin, temporarily disconnecting the subject from the recording system.
- Electrode motion artifacts – variations in electrode–skin impedance due to electrode movement.
- Muscle contraction noise (Electromyographic noise) – caused by electrical activity from skeletal muscles other than the heart lasting around 50 ms.
- Instrumentation noise – produced by the electronic components used in ECG acquisition.
- Electrosurgical noise – interference from other medical equipment present in the patient care environment, usually in the frequency range of 100 kHz to 1 MHz.
- Quantization noise and aliasing.
- Signal processing artifacts

Fig. 2.4 illustrate three common types of noise in ECG signal such as baseline wander, 50 Hz power line interference and muscle contraction noise.

2.6 Therapeutic Hypothermia

Therapeutic hypothermia (TH) is the only recommended treatment for moderate to severe HIE. It has been shown to reduce seizure burden and improve long-term

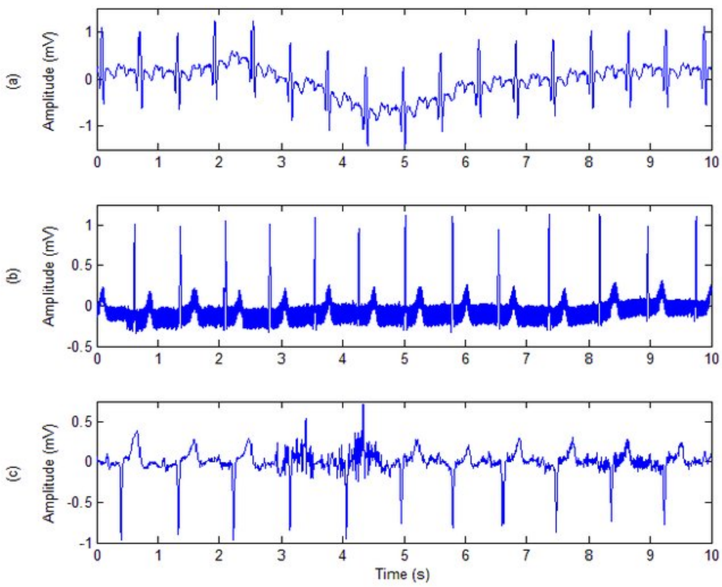


Figure 2.4: Common types of noise in ECG recordings. (a) Baseline wander, (b) 50 Hz power line interference, and (c) Muscle contraction noise. [**?**].

neurodevelopmental outcomes [**?**]. TH involves controlled cooling of the infant core temperature to 33.5 °C for 72 hours, followed by gradual rewarming over 7–14 hours. Its overall goal is to limit neuronal injury from secondary energy failure, a key contributor to adverse neurodevelopment. TH can be initiated any time within the first 6 hours after birth, though earlier initiation is associated with greater benefit [**?**]. Prompt and accurate assessment of HIE severity in the NICU is therefore critical to enable timely treatment. Intubation, a common medical procedure involving the insertion of a tube into the airway is frequently employed during therapeutic hypothermia in neonates, making it essential to critically evaluate its necessity and appropriateness.

The continuous EEG in grades normal and normal/mild (grades 0 and 1) generally relates to non-cooled infants with normal and mild Sarnat grades, while moderate, major and inactive EEG grades (grades 2, 3 and 4) relate to moderate to severe Sarnat grades requiring TH. However, these EEG and Sarnat grades are not strictly equivalent. This assumption is further supported by the fact that aEEG criteria were used in the TOBY cooling trial [?], where an abnormal aEEG (i.e., discontinuous activity, equivalent to moderate or worse EEG grade) was a requirement for cooling.

2.7 Automated HIE grading and seizure detection using the ECG signal (previous works)

Currently, EEG is commonly used to help assess HIE grade and monitor seizures. However, EEG monitoring presents significant challenges in the neonatal setting [?]. Interpreting neonatal EEG requires specialized expertise, which may not always be available in the NICU. EEG is not available in all hospitals or NICUs due to the high cost of the equipment. Additionally, the need to attach electrodes to neonate head and body and the complexity of EEG interpretation in the immediate postnatal period can hinder timely diagnosis. These combined limitations have driven ongoing research into automated and accessible methods for HIE grading and seizure detection. Previous research on automated methods has mostly used EEG signals. Further details on EEG-based automated methods for HIE grading and seizure detection are provided in the appendix.

One potential alternative is the electrocardiogram (ECG), which is widely accessible, routinely monitored in neonates, and feasible even in smaller neonatal units and resource-limited settings. By analysing and processing ECG signals, valuable diagnostic information can be extracted. However, research on ECG-based HIE grading and seizure detection remains limited, highlighting the need for further investigation in this area.

Several studies have utilized ECG signals for seizure detection in adults [?] [?

Table 2.2: Review of studies on HIE grading and seizure detection using ECG signals. Sens: Sensitivity, Spec: Specificity, AUC: Area under curve

Study	Problem	Number of Patients	Total Time (hours)	Result (%)
Ahmed et al.[?]	HIE	54	54	AUC of 81
Greene et al. [?]	Seizure	8	101	Accuracy of 70.5
Doyle et al. [?]	Seizure	14	1457	Sens 60, Spec 60
Olmi et al. [?]	Seizure	51	45.76	Sens 47, Spec 67
Lu et al. [?]	Seizure	16	35	AUC of 62.7

] [?] [?] [?], while others have explored multimodal approaches combining EEG and ECG signals of adults [?], [?] [?]. The heart rate in neonate is different from adults, therefore dedicated studies are needed to develop methods specifically tailored to neonatal heart signals.

Research on neonatal HIE grading and seizure detection using ECG signals remains limited. Existing studies have focused on extracting the heart rate from ECG recordings and manually deriving time- and frequency-domain features from heart rate variability (HRV). These studies are few in number, generally constrained by small datasets. Table 2.2 provides a summary of studies on HIE grading and seizure detection using ECG signals, which are discussed in the following subsections.

2.7.1 HIE grading using ECG

Goulding et al. [?] studied seven HRV features, muNN, SDNN, TINN, LF/HF ratio, VLF power, LF power, and HF power (described in detail in Chapter 3) of neonates with HIE. They demonstrated that these features could distinguish between HIE grades. Specifically, all features except one significantly differed between mild and severe cases, while all seven features distinguished between mild and moderate HIE. Matic et al. [?] conducted continuous monitoring of

EEG and ECG in neonates divided by HIE severity. Their results indicated that HRV parameters were elevated in neonates with severe HIE compared to those with less severe conditions.

Ahmed et al. [**?**] performed a classification task involving two HIE severity levels. Using 54 one-hour ECG recordings, they extracted seven HRV statistical features to create supervectors. These were then classified using SVM, Gaussian Mixture Models, and supervector-based SVM. The highest accuracy was achieved with the supervector-SVM, yielding an AUC of 81%.

2.7.2 Seizure detection using ECG

The inconsistent performance of heart rate–based seizure detection may partly reflect physiological variability, as suggested by Goulding [**?**]. In their study, 26 neonates with HIE and seizures were analyzed. Seven HRV features, muNN, SDNN, TINN, LF/HF ratio, VLF power, LF power, and HF power (described in detail in Chapter 3), were examined across all participants. The results indicated that some features tended to increase during electro-clinical seizure segments, while one feature showed a trend toward a significant decrease in sub-clinical seizure episodes. Furthermore, several features were significantly elevated between baseline and seizure epochs in infants with a severe background EEG grade.

Heart rate–based approaches for neonatal seizure detection show low performance across standard evaluation metrics, including sensitivity, specificity, and overall accuracy. In the work by Greene et al. [**?**], a patient-independent linear discriminant classifier utilizing 41 HRV features was applied to 101 hours of recordings from eight neonates, all of whom experienced seizures. This approach achieved a sensitivity of 54.6% and a specificity of 77.3%. Other works used the SVM classifier and HRV features. Doyle et al. [**?**] explored a broader set of 62 time and frequency domain HRV features, extracted from 207.86 hours of recordings from 14 neonates with seizure activity. Their system, evaluated using Leave-One-Out (LOO) cross-validation with a SVM, yielded a sensitivity and specificity of 60%, However, the performance varied significantly among individual neonates; only two out of the 14 showed strong results, indicating the method's limited generalizability. More recent work by Olmi et al. [**?**] involved using 18 features fed into an SVM classifier. This study analyzed 45.76 hours of ECG data from 51 neonates, including 22 with seizures, and achieved a sensitivity of 47% and a specificity of 67%. Lu et al. [**?**] performed seizure detection on 16 neonates, using a total of 35 hours of ECG recordings. The method, based on an SVM with a linear kernel, achieved an AUC of 62.7%.

2.8 Automated HIE grading and seizure detection using ECG signal in this thesis

Previous studies have highlighted the potential of using ECG as an alternative to EEG for neonatal HIE grading and seizure detection, although the number of such studies remains limited.

To identify the most informative HRV features, this thesis examines the seven HRV features studied by Goulding et al. [**?**] [**?**], along with additional features reported in state-of-the-art studies, to determine and evaluate the most effective set of HRV features from the machine learning point of view.

Additionally, time–frequency representations of the ECG signal will be explored using ECG spectrograms and deep learning models, which have been effective in tasks such as heart rate estimation [**?**], cardiovascular disease classification [**?**], arrhythmia classification [**?**] [**?**], and assessing tetanus severity [**?**]. This eliminates the heart rate extraction and HRV feature calculation processes, addressing their limitations such as sensitivity to noise and computational inefficiency. This facilitates the use of modern deep learning architectures, including convolutional neural networks (CNNs) and vision transformer-based models.

This thesis will also evaluate the feasibility of using raw ECG signals with Transformer-based models, which have shown promising results in ECG applications including arrhythmia classification [?], [?] [?], obstructive sleep apnea detection [?], cardiac disease diagnostic [?], and stress detection [?]. The novel design proposed in this thesis preserves the temporal information of the data while reducing computational overhead, enabling real-time processing and enhancing its practicality for clinical use.

The main goal of this thesis is to leverage ECG as a more accessible and cost-effective alternative to EEG for developing supportive tools that assist clinicians in HIE grading and seizure detection in real-world clinical settings. By leveraging easily obtainable ECG signals, this methods reduces reliance on EEG-based assessments by neurophysiologists and could improve access to accurate HIE assessment and seizure detection in lower-resource healthcare settings.

Figure 2.5 illustrates all the methods used in this thesis for ECG processing and binary classification tasks, including HIE assessment for TH requirement (grade 0: TH not required, grade 1: TH required), as well as seizure detection (class 1: seizure, class 0: non-seizure). The first method (left) represents classical approaches based on HRV features combined with traditional machine learning models. The middle method illustrates an approach using ECG spectrograms with a CNN model. The final method (right) uses raw ECG signals directly as input to a transformer-based model, demonstrating how preprocessing steps are reduced.

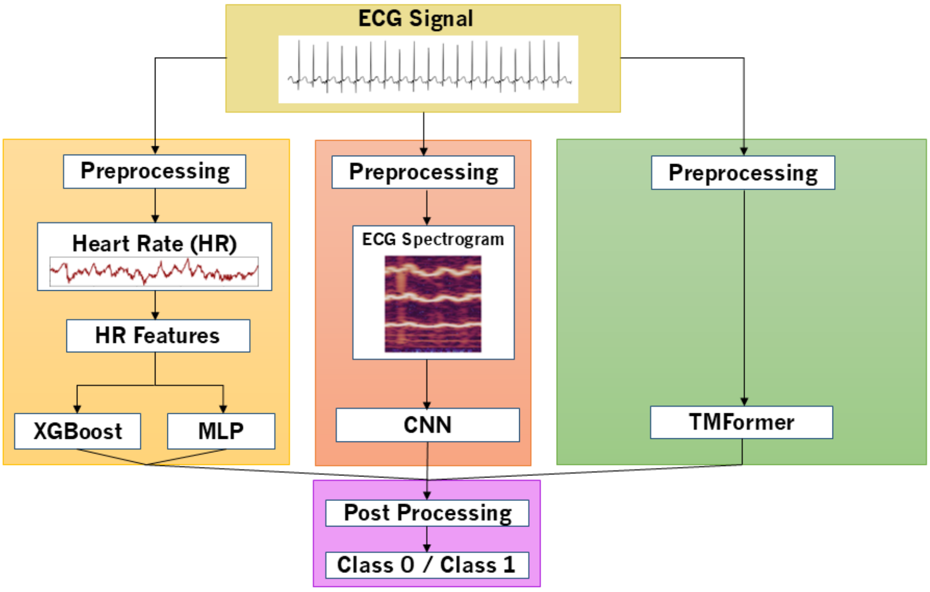


Figure 2.5: All the methods used in this thesis for ECG processing and binary classification tasks, including HIE assessment for TH requirement (grade 0: TH not required, grade 1: TH required), as well as seizure detection (class 1: seizure, class 0: non-seizure

2.9 Datasets and annotations

Various datasets were used in this study for the purpose of HIE grading and seizure detection using ECG signals. Overall information about the datasets is provided in Table 2.3, while detailed information is provided for each dataset in subsequent sections. The available ANSeR dataset for both HIE and seizure is very large and was used throughout the experimental chapters (Chapters 2, 4, and 5) of this thesis. In contrast, the EarlySet and Cork datasets are comparatively smaller and were used only in Chapter 3, as their size was insufficient for the deep learning methods applied in Chapters 4 and 5.

Table 2.3: Information about the datasets used for HIE grading and Seizure detection using ECG signal in this thesis.

Problem	Dataset name	Number of patients	Total time (hours)
HIE	ANSeR	58 ANSeR1, 76 ANSeR2	1756
HIE	EarlySet	120	120
Seizure	ANSeR	42 ANSeR1, 23 ANSeR2	2442
Seizure	Cork	14	1457

2.9.1 Annotations

2.9.1.1 HIE annotations

Typically, a neurophysiologist grades EEG signals into five categories: normal, normal/mild, moderate, major and inactive [**?**] as explained earlier in Table 2.1. Based on this HIE-EEG grading, decisions regarding treatment for the neonate can be made. In particular, neonates with moderate, major, or inactive HIE usually require therapeutic hypothermia (TH). In the NICU, EEG must be recorded and subsequently assessed by a neurophysiologist within six hours of birth to ensure timely and appropriate treatment decisions. Therefore, automatically classifying EEG signals into two categories could serve as a valuable decision-support tool for clinicians. Category 0 would include normal and mild cases, indicating that therapeutic hypothermia (TH) is generally not required, while Category 1 would include moderate, major, and inactive cases, suggesting that TH is needed. Accordingly, this study adopts a binary classification approach for HIE-EEG grading. However, instead of using EEG signals, ECG signals are utilized, which are more readily accessible and routinely monitored in neonates. Since the EEG and ECG recordings were time-synchronized in the datasets used in this thesis, the same EEG-based annotations were applied to the corresponding ECG segments.

2.9.1.2 Seizure annotations

For effective seizure detection, it is essential to distinguish between seizure and non-seizure (baseline) segments in the EEG recordings. Reliable identification of neonatal seizures requires specialized EEG monitoring equipment and expert neurophysiological interpretation, as many neonatal seizures are subclinical and lack obvious physical symptoms. In the dataset used for this thesis, neurophysiologists annotated the onset and offset times of seizure events in the EEG signals. To develope a clinical support tool for seizure detection, a binary classification approach should be applied, where 0 denotes non-seizure activity and 1 denotes seizure activity. Accordingly, this study adopts a binary classification approach for seizure detection.

2.9.2 Annotation limitations

This work benefits from expert EEG annotations provided by neonatal neurophysiology specialists. However, some HIE cases are borderline, leading to a degree of label uncertainty. Additionally, although the annotations are based on EEG recordings, the models in this thesis utilise the corresponding ECG signals. As a result, there may be a temporal mismatch between the EEG-based annotations and the ECG signal changes. In particular, ECG alterations related to seizure activity may occur slightly before or after the annotated EEG onset.

2.9.3 ANSeR dataset

In this study, two datasets called ANSeR1 and ANSeR2 were used [**?**] [**?**]. A multichannel EEG, as well as a single or two lead ECG where available, was recorded from neonates at risk of seizures shortly after birth and was continued for up to 72 hours where possible. The ECG sampling rate matched that of the EEG, at 256 Hz. All infants had a gestational age of at least 36 weeks. The signals were recorded in eight tertiary-level neonatal intensive care units across Europe (Ireland, United Kingdom, Sweden and Netherlands) using Nihon Kohden Neurofax (EEG-1200), NicoletOne ICU Monitor (Natus, Middleton, WI, USA) or XLTek (Natus).

2.9.3.1 HIE dataset

From the ANSeR1 and ANSeR2 datasets, a subset of EEG epochs from infants with HIE was graded at five different postnatal ages (6, 12, 24, 36, and 48 hours), with each epoch lasting 1 hour. A neurophysiologist graded the EEGs as normal, normal/mild, moderate, major or inactive [**?**]. For each neonate, there was more data available but this was unlabeled. To enhance model performance, additional 1-hour epochs were extracted from the intervals between these five time points and weakly labeled based on the preceding and following epoch classifications. Figure 2.6 illustrates the labeling process. As shown, EEG recordings at 6, 12, 24, 36, and 48 hours were annotated by a neurophysiologist. All of the intermediate epochs were labeled based on these reference annotations. However, epochs occurring during transitions between changing labels were excluded, as a definitive labels could not be reliably assigned.

The weak labeling was only employed on the ANSeR2 dataset which was used as a training dataset. The EEG-based labels were applied to the corresponding time segments of the ECG signal for model training. For binary classification, 1-hour ECG epochs were categorized into two groups. Category 0 includes cases of normal and mild HIE, which suggests that neonates do not require TH, and category 1 includes cases of moderate, major, and inactive HIE, which suggests that TH is required.

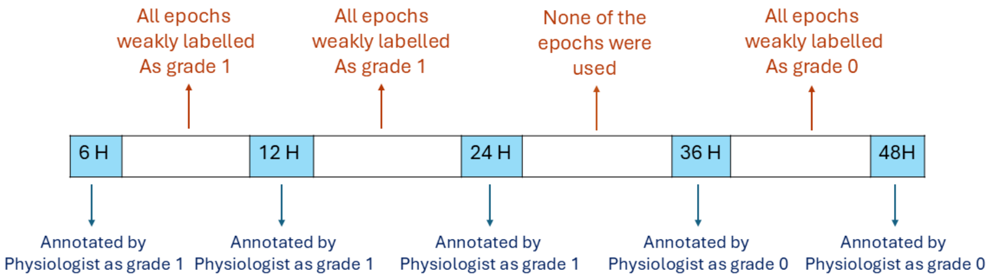


Figure 2.6: Labeling of an EEG recording from hour 6 to 48 of age. hours 6, 12, 24, 36, and 48 were labeled by neurophysiologist and the intermediate epochs were weekly labeled.

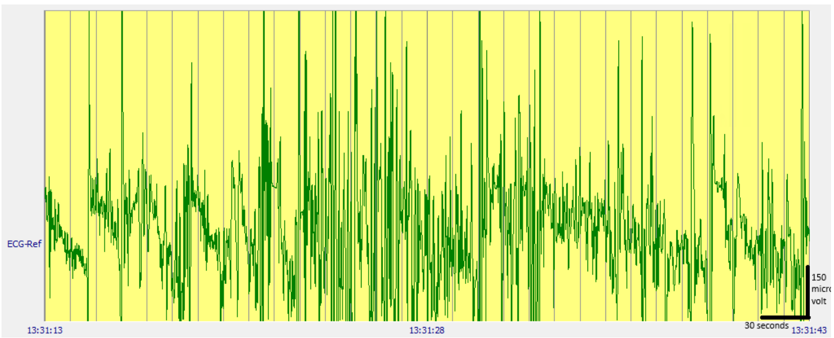
Recordings from patients without an ECG signal channel were entirely excluded from the analysis. For patients with available ECG data, 1-hour epochs that were heavily corrupted by noise were removed. Fig. 2.7 shows portions of a noisy ECG signal for which the entire recording was discarded.

Consequently, 58 out of 91 neonates from ANSeR1 and 73 out of 90 neonates from ANSeR2 were included in this study. The ANSeR1 dataset includes 216 1-hour epochs strongly labelled (130 class 0, 86 class 1). The ANSeR2 dataset comprises 254 strongly labelled and 1286 weakly labelled 1-hour epochs (792 category 0, 748 category 1).

2.9.3.2 EarlySet HIE dataset

Initially, a smaller subset from the ANSeR1 and ANSeR2 dataset was annotated. This dataset is used for the purpose of HIE grading only in chapter 3. It consists of 120 neonates, each with a recording duration of about 1 hour (66 category 1 and 54 category 0) [?]. In this thesis this dataset is referred to as the EarlySet dataset.

(a)



(b)

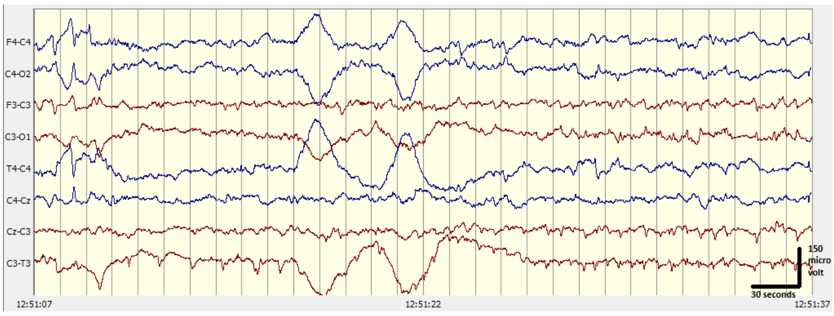
Figure 2.7: Two noisy ECG segments, with segment (b) being more noisy than segment (a). The entire recording was discarded due to high artifacts.

2.9.3.3 Seizure dataset

From the ANSeR1 and ANSeR2 datasets, only neonate recordings with seizure were used. For all the recordings, seizure start times and durations were provided by experts from the EEG recordings. Recordings without an ECG signal channel or those highly affected by noise were excluded. Out of 472 patient EEG recordings, 155 patients had seizure annotation, among which 65 patients had ECG channels which were used in this thesis. The ANSeR1 dataset consists of 42 patients and the ANSeR2 dataset includes 23 patients. In total, this dataset included approximately 2,442 hours of ECG recordings, comprising 1,569 seizure events with a combined duration of 82.89 hours. The seizure event durations ranged from a minimum of 2 minutes to a maximum of 116.3 minutes. The EEG-based labels were applied to the corresponding time segments of the ECG signal for model training. Fig. 2.8 presents a 30-second EEG segment exhibiting a seizure, annotated by a neurophysiologist, along with the corresponding ECG segment, which I annotated as a seizure.

2.9.4 Cork seizure dataset

In this study, a dataset of EEG recordings from 17 neonates admitted to the Neonatal Intensive Care Unit (NICU) at Cork University Maternity Hospital, Ireland, collected over a two-year period [**?**] was used. This dataset is referred to as Cork dataset. During this period, 55 full-term infants diagnosed with Hypoxic Ischaemic Encephalopathy (HIE) were recruited, and EEG monitoring was conducted for up to 72 hours. Recordings were obtained using a Viasys NicOne video EEG system at a sampling rate of 256 Hz, with electrodes placed according to the 10–20 system adapted for neonates. The Cork dataset consists of two different annotated sets, one for HIE grading and one for seizure detection. From this dataset,



(a) EEG

(b) ECG

Figure 2.8: (a) An EEG segment annotated by the neurophysiologist as a seizure. (b) Corresponding ECG segment, annotated as a seizure in this thesis.

ECG signals of 14 fullterm neonates with seizures were annotated by the neurophysiologist and subsequently used for the seizure detection task in this study. Since the EEG and ECG signals were synchronized in time, annotations were used for the ECG recordings for this study. In total, this dataset included approximately 1,457 hours of ECG recordings, incorporating 396 seizure events with a combined duration of 87.5 hours.

2.10 Implementation details

In this thesis, all classical machine learning experiments were conducted using Python 3, with libraries such as scikit-learn, NumPy, pandas, and pyHRV. These experiments were performed on a virtualized compute environment equipped with an Intel processor running at approximately 2.60 GHz and 64 GB of system memory.

All deep learning experiments were implemented using the PyTorch 2.3 framework in Python 3. Training was performed on a virtualized system equipped with a single NVIDIA Tesla V100-SXM2 (16 GB VRAM), with GPU acceleration supported by CUDA and cuDNN.

2.11 Conclusion

This chapter presented fundamental information about HIE and seizures, emphasizing the importance of their clinical treatment and the potential benefits of automatic systems for HIE grading and seizure detection in supporting clinicians. It also discussed the use of EEG signals for HIE grading and seizure detection, in addition to an overview of ECG signals as a potential alternative to EEG. Also, therapeutic hypothermia, an established treatment for HIE, was described.

This chapter summarized the literature on automated HIE grading and seizure detection methods, emphasizing the potential of using ECG as a more accessible alternative to EEG and underscoring the need for further research into its application for these tasks. Finally, the datasets and their corresponding annotations utilized in this study were described in detail.

3 | HIE Grading and Seizure Detection Using HRV Features

3.1 Introduction

In this chapter, the Pan-Tompkins method, which is widely used to detect QRS complexes and calculate the heart rate from ECG signals, will be described in detail. Following this, the preprocessing of heart rate variability (HRV) data and the statistical features that can be extracted from HRV will be discussed, highlighting how these features provide valuable insights into heart function based on the findings by Doyle et al. [**?**] and Goulding [**?**] [**?**].

Next, the ECG signals will be analyzed to extract HRV metrics and calculate their statistical features. The chapter will also evaluate the effectiveness of the clinical findings by Doyle et al. [**?**] and Goulding [**?**] [**?**] from the machine learning point of view for two applications: grading hypoxic-ischemic encephalopathy (HIE) and detecting seizures.

For this study, all datasets introduced in Chapter 2, section 2.9 will be utilized. Initially, smaller datasets (th eEarlySet HIE Dataset and the Cork seizure dataset) will be used to identify the most informative features and determine the optimal machine learning models. Subsequently, the large ANSeR dataset will be employed to assess the performance of these models and investigate how dataset size and diversity influence the effectiveness of machine learning in this context.

3.2 Heart rate extraction from ECG signal

Heart rate (HR) can be automatically derived from the ECG signal using a QRS detection algorithm. Several classical approaches exist, including Hamilton–Tompkins [**?**], Pan–Tompkins [**?**], the Christov algorithm [**?**], Shannon energy–based detectors [**?**], and Hilbert transform–based methods [**?**]. Many of these techniques share similar preprocessing steps, such as band-pass filtering and noise reduction, although their R-point detection and decision rules differ. More recent machine-learning-based detectors have also been developed, but they typically require higher computational resources. Studies have evaluated a range of classical and modern peak detection techniques on ECG signals. These methods exhibit reduced performance in detecting QRS complexes under various cardiac conditions [**?**]. Additionally, detection accuracy decreases significantly for poor quality ECG signals with high noise level [**?**].

The Pan–Tompkins algorithm [**?**], one of the most widely used QRS detection methods, is described in detail in the appendix of this thesis. The pan–Tompkins method is generally regarded as reliable; however, it is still susceptible to inaccuracies or missed detections during QRS complex identification. This limitation arises because the algorithm is tuned for normal heart rhythms, while real-world ECG recordings are frequently distorted by noise and artifacts, such as 50 Hz power line interference, baseline wander, and muscle activity. In clinical cases involving HIE or seizures, additional heart rate fluctuations can introduce further irregularities, making accurate peak detection even more challenging [**?**] [**?**]. These detection errors can subsequently compromise the accuracy of feature extraction from the NN interval. Fig. 3.1 demonstrates an example of peak detection on a noisy ECG segment using the Pan-Tompkins technique, where multiple peaks were either missed or falsely detected due to transient signal spikes.

HRV is a physiological measure which refers to the variation in the time intervals between successive heartbeats. This variability is typically quantified using the NN interval, which represents the time difference between two consecutive R-peaks (QRS complexes) in the ECG signal, as defined in Equation (3.1). Heart rate (HR), expressed in beats per minute (bpm), can then be calculated by taking the reciprocal of the NN interval, as shown in Equation (3.2).

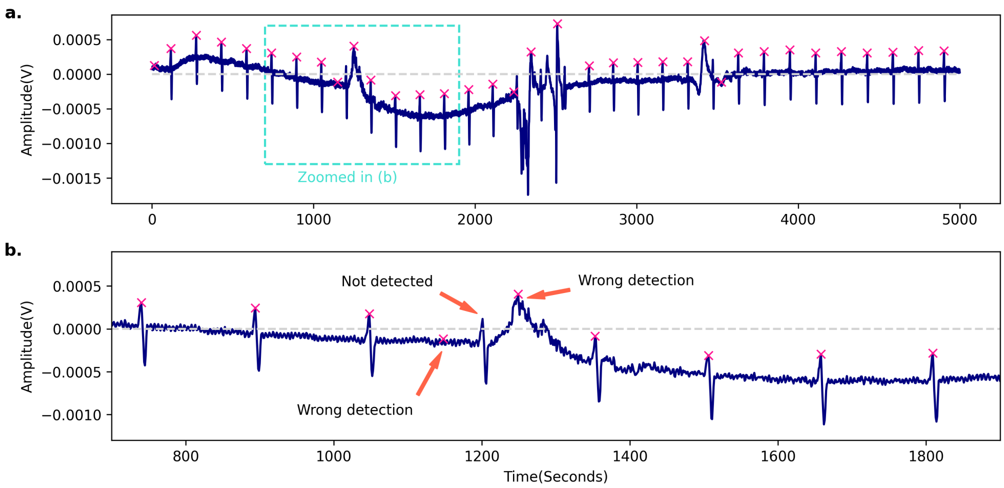


Figure 3.1: (a) An example of a noisy ECG signal where Pan-Tompkins missed some peaks and wrongly detected others. (b) A zoomed-in segment of (a) depicting missed and falsely detected peaks.

$$NN = peak_n - peak_{n-1} \tag{3.1}$$

$$HR\ (\text{bpm}) = \frac{60}{NN\ (\text{seconds})} \tag{3.2}$$

In this chapter, HRV software (HRV Analysis, University College Cork, Cork) was used to identify peaks within the ECG signal and and NN intervals are calculated from these timestamps using Python. Fig. 3.2 presents the heart rate calculation from an ECG signal using this software. The software implementation is based on the Pan–Tompkins algorithm.

3.3 Preprocessing of heart rate

To prepare the data for subsequent machine learning analysis, a series of preprocessing steps were applied. As outlined in section 3.2, the process began by ex-

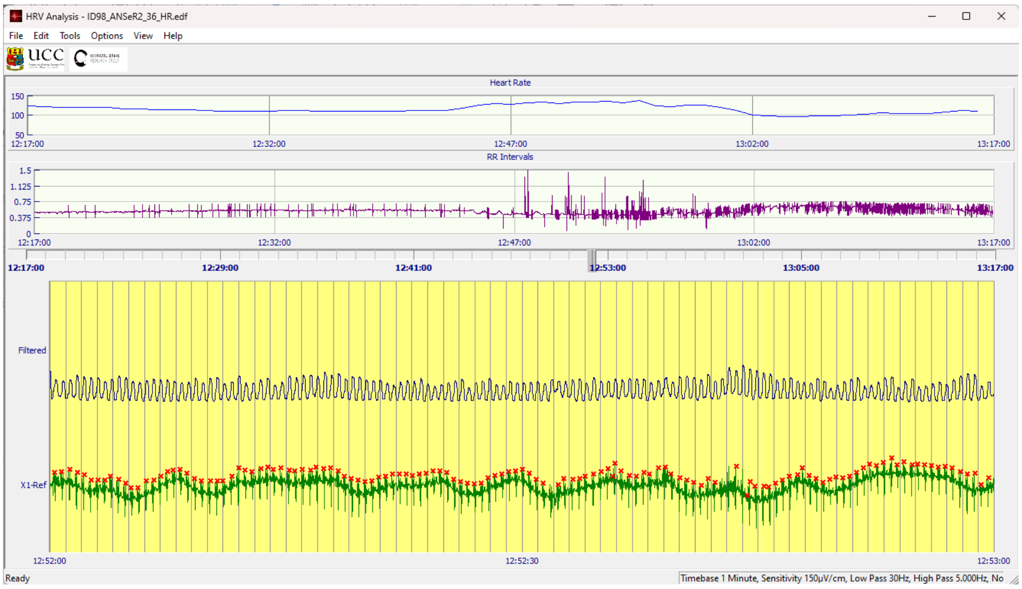
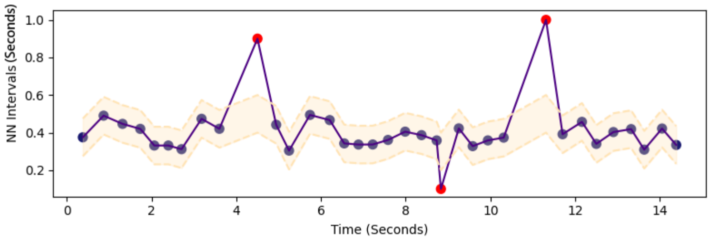


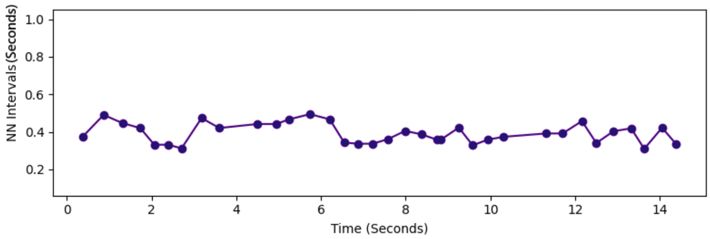
Figure 3.2: ECG signal peaks detected and NN intervals calculated by HRV software (HRV Analysis, University College Cork, Cork)

tracting normal-to-normal (NN) intervals from the raw ECG signal. The NN intervals were then cleaned using the Hampel filter [**?**], a robust method for outlier detection. The Hampel filter operates by applying a sliding window to the NN interval data, comparing each point to the median absolute deviation (MAD) of its local neighborhood. Points that significantly deviate from the local median are identified as outliers and replaced with the median value within the window. Fig.3.3 (a) shows the NN intervals of an infant. The normal range (typically 0.3–0.5 s) is indicated by the upper and lower borders. Points falling outside this range appear as red outliers. Fig.3.3 (b) demonstrates that, after applying the Hampel filter, these outliers are replaced by median-based estimates, resulting in a smoother signal.

In adults, the average resting heart rate typically ranges from 60 and 100 BPM, whereas neonates generally have a higher HR ranging from 70 to 190 BPM. For further HR processing, such as feature extraction and feeding into machine learn-



(a)



(b)

Figure 3.3: (a) NN intervals with outliers and borders representing higher and lower limits. (b) NN intervals after applying Hampel filter.

ing models, which will be discussed later in this chapter, the long recorded HR signals need to be trimmed into shorter segments. A segment length of two to five minutes is usually recommended as it balances the requirements for practical HR data analysis with the need to capture HIE and seizure related events and abnormalities within a clinically relevant time frame [?] [?]. Therefore, HR signals were split into shorter segments. To reduce information loss at the boundaries between segments and to capture more continuous signal dynamics, an overlap of 50% to 80% between consecutive segments was applied. Segments highly affected by noise and artifacts were discarded during preprocessing. Segment length and overlap were selected differently for each method and are detailed in the Experimental sections.

3.4 Heart rate variability features

From each clean HR segment, sixteen features were extracted from both the time and frequency domains. Among these, seven features previously examined by Goulding et al. [?] [?] were included, as his study investigated the statistical significance of these measures in neonates with seizures and varying degrees of HIE severity. These seven features consist of muNN, SDNN, TINN, LF/HF ratio, VLF power, LF power and HF power. Detailed descriptions of each feature calculation are in the appendix. The remaining features were selected based on a review of state-of-the-art research in heart rate variability analysis [?] [?] [?], with the goal of enhancing both the robustness and comprehensiveness of the feature set.

3.5 Machine learning models

In this study, five machine learning algorithms were employed for classification: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Isolation Forest (IF), Support Vector Machine (SVM), and the Multilayer Perceptron (MLP). Each algorithm was chosen to leverage its specific strengths. Tree-based ensemble models such as RF and XGBoost are used for handling complex nonlinear relationships and feature interactions, and IF for anomaly detection. Margin-based model such as SVM are used for robustness with high-dimensional data, and deep learning model such as MLP for capturing complex patterns through neural network modeling. Detailed descriptions for all used models can be found in the appendix.

3.6 Handling class imbalance

The seizure detection dataset is highly imbalanced, with the number of seizure events being significantly lower than non-seizure segments. To solve the imbalance problem, the minority class was over-sampled through Synthetic Minority Oversampling Technique (SMOTE) [**?**]. SMOTE works by selecting random instances in the minority class that are close in the feature space. Then, the synthetic instance is created by drawing a line between the examples with same label in the feature space and drawing a new sample at a point along that line. This, ensures that seizure samples contribute proportionally more to the loss function, helping the model avoid bias toward the majority non-seizure class.

In this chapter, SMOTE was applied using the minority sampling strategy, and only on the training data, leaving the test dataset unchanged. Consequently, the number of seizure segments (minority class) in each training set was increased through synthetic generation so that it equals the number of non-seizure segments (majority class).

SMOTE is applied to all supervised classifiers in this chapter since they rely on balanced labeled data to learn representative patterns from both classes. Isolation Forest is not included, as it is an unsupervised anomaly detection algorithm rather than a binary classifier and it operates naturally under highly imbalanced conditions.

3.7 Hyperparameter optimization

Hyper parameters are settings defined before training that control a machine learning model's behavior and performance. Choosing the best values, known as hyper parameter optimization, is crucial for good results. Grid search explores all possible combinations of predefined values for each hyper parameter, guaranteeing full coverage but at a high computational cost. To improve reliability and avoid over fitting, grid search is commonly combined with cross-validation, which evaluates models across multiple data splits. This must be done only on the training data to avoid influencing the test set, which should remain untouched as an unbiased benchmark of the model's true performance.

3.8 Performance evaluation

Cross-validation (CV) is a widely used technique for assessing the performance of machine learning models. By repeatedly splitting the dataset into training and testing subsets, it provides a more reliable estimate of how well a model generalizes to unseen data. In this thesis, all train and test subsets were patient independent (data from each patient is kept completely separate between training and testing). In K-Fold cross-validation, the dataset is divided into k equally sized folds (as evenly as possible). The model is trained on k–1 folds and tested on the remaining fold, with the process repeated until every fold has been used for testing, while all folds are patient independent. Patient independent Leave-One-Out (LOO) is a special case where each training set consists of all patients except one, and the left-out sample serves as the test set. This procedure repeats until each patient has been treated as the held-out test subject, therefore, the number of folds equals the number of patients in the data set.

These methods are useful as they allow the entire dataset to contribute to model training while ensuring that the test set remains unseen in each fold. Additionally, since each fold is tested on data from different patients, performance can be compared across diverse patient profiles, providing a more robust assessment. In this work, patient independent LOOCV was used only for seizure detection on the

Cork seizure dataset (described in Chapter 2, Section 2.9), as it includes data from only 14 patients. For all other experiments on larger datasets, patient independent k-fold cross-validation was employed, as it provides a more efficient and reliable estimate of model performance. In all folds, the training and test datasets are completely patient-independent thus ensuring that the test data remains unseen during training.

The classification performance of the model on the test set of each fold was evaluated using several standard metrics. First a confusion matrix was generated for each fold, detailing the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) produced by the classifier. This matrix enables a direct assessment of the types of errors produced by the model, which is particularly valuable in clinical applications where different error types have different implications. From the confusion matrix, specificity, and sensitivity were derived. Specificity (3.3) quantifies the model's ability to correctly identify class 0 segments, which is particularly important in imbalanced datasets where negative samples dominate. Sensitivity (3.4), on the other hand, measures the proportion of class 1 segments correctly detected, reflecting the model's effectiveness in identifying clinically relevant positive cases. Taken together, these metrics offer a balanced evaluation that considers overall correctness while addressing the need to minimize both missed detections and false alarms.

$$Specificity = \frac{TN}{TN + FP} \tag{3.3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.4}$$

In addition to these metrics, the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) curve was also used to assess the model's performance. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) across all thresholds. The AUC quantifies the overall ability of the model to discriminate between the two classes. AUC is less sensitive to class imbalance than accuracy as it evaluates the ranking of predictions rather than the absolute number of correct classifications. A higher AUC value (close to 1.0) indicates better discriminatory performance, which is essential in clinical applications where false negatives can have serious consequences.

The Matthews Correlation Coefficient (MCC) [**?**] was also reported, as it provides a balanced measure that takes into account true and false positives and negatives, making it especially informative for datasets with unequal class distribution. MCC provides a value between -1 and 1. 1 indicates perfect prediction, 0 indicates no better than random and -1 indicates total disagreement. MCC is defined as follows:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.5}$$

Accuracy measures the overall proportion of correctly classified instances (both positive and negative) out of all predictions. It gives a general idea of how often the model is correct. Accuracy is included in the radar chart alongside the other evaluation metrics. Accuracy is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

Impurity-based feature importance was calculated for the tree-based models. This metric reflects the mean and standard deviation of the reduction in impurity (a measure of how mixed or disorderly the target values are in a node) accumulated across all trees in the model. Higher importance values indicate features that contribute more to the model's predictions, while lower values indicate less influential features. The values are normalized to sum to one, providing a clear picture of each feature's proportional contribution to the model's overall predictive performance.

To visually explore the distribution of data across different classes, Kernel Density Estimation (KDE) [**?**] [**?**] plots were generated for all features. These plots provide a clear depiction of how each feature's values are distributed within each class, offering insight into their potential discriminative power.

Additionally, Uniform Manifold Approximation and Projection (UMAP), a dimensionality reduction technique, is conducted. UMAP was used to project the high-dimensional data into a 2D space, enabling visualization of the data distribution and facilitating the inspection of class separation within the large dataset. In UMAP, good class separation appears as clusters of different colors being spatially apart, whereas poor separation is indicated by clusters of different colors intermingling.

By jointly considering all the mentioned measures, a comprehensive view of the model's performance in the context of HIE grading and seizure detection was obtained.

3.9 Evaluate the effectiveness of HRV features from the machine learning point of view

Heart rate-based detection of neonatal seizures has produced mixed outcomes across various studies. The inconsistent performance in heart rate-based seizure detection may partly stem from physiological variability, as suggested by Goulding [**?**

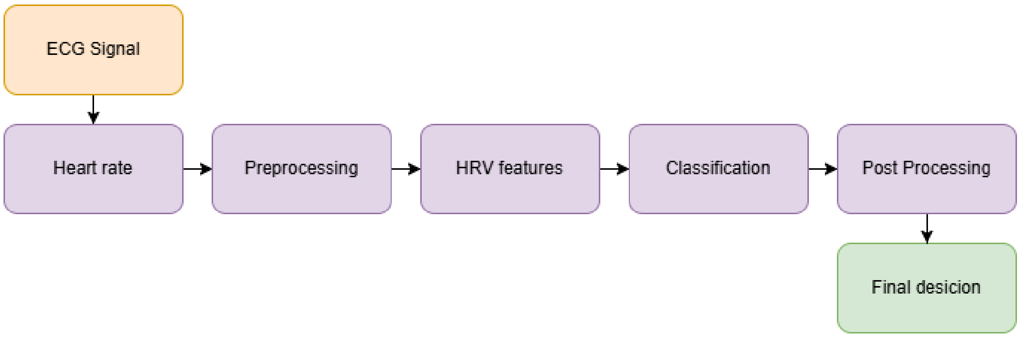


Figure 3.4: The flowchart of proposed method for HIE grading and seizure detection using heart rate variability features.

]. Their analysis focused on seven HRV features extracted from ECG recordings of 26 neonates with HIE and seizures. Based on this study, TINN, LF, and HF tended to increase during electro-clinical seizure segments, while muNN showed a trend toward a significant decrease in sub-clinical seizure segments. Additionally, TINN, VLF, LF, and HF were significantly elevated between baseline and seizure epochs in infants with a severe baseline EEG grade. Additionally, Goulding et al. [**?**] demonstrated that the same features could distinguish between HIE grades. Specifically, all features except LF/HF ratio significantly differed between mild and severe cases, while all seven features distinguished between mild and moderate HIE.

In this section, the effectiveness of this clinical finding is evaluated from a machine learning perspective. Various methodological approaches will also be explored to identify the most effective pipeline for HIE grading and seizure detection, including preprocessing, feature engineering, modeling, and post-processing. The overall flowchart of the proposed model is illustrated in Fig. 3.4.

3.9.1 Experimental setup

For this experiment, the EarlySet HIE dataset and the Cork seizure dataset mentioned in chapter 2, section 2.9 were used. As explained in Section 3.3, a segment length of two to five minutes balances the requirements for practical HR data analysis. Since the datasets are small and the seizure dataset is imbalanced, for both HIE grading and seizure detection, the HR signals were segmented into 2-minutes segments with an overlap of 50% to have sufficient samples for analysis using the models.

For HIE grading, 120 patients each with about 1-hour ECG epoch were used. Among the 120 patients, 66 patients were category 1 (TH required) and 54 were category 0 (TH not required). From these 120 patients, 7,493 two-minute segments were utilized, and 39 segments were removed due to artifacts.

For seizure detection, 28,485 segments (26735 non-seizure and 1750 seizure) were employed, with 655 epochs being excluded due to noise. 1750 seizure segments were extracted from 396 seizure events. Here, a seizure event refers to the full annotated seizure event from the long recording, not the 2-minute segmented seizures. Since seizure labeling by experts was performed on a per-second basis, each two-minute segment was labeled as a seizure if at least two-thirds of its duration contained seizure activity.

For HIE grading, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) with a radial basis function (RBF) kernel were employed, as these methods are well-suited for structured data and provide strong baseline performance in classification tasks. For seizure detection, the same three models were used in addition to Isolation Forest (IF), which was specifically included due to its suitability for anomaly detection, a relevant approach given the relative rarity of seizure events. The Multilayer Perceptron (MLP) was not considered in this stage because of the limited dataset size, as neural networks typically require larger datasets to generalize effectively; however, it will be explored further on the larger ANSeR dataset.

Here the data was cleaned using the Hampel filter, however, it was excluded from the preprocessing pipeline for data fed into the IF model, since it focuses on detecting outliers. Further, the data was standardized before training the SVM classifier to ensure all features had zero mean and unit variance. This improves the performance and convergence of SVM, which is sensitive to the scale of input features.

Hyperparameter optimization for all models was conducted independently using grid search to identify the best parameter settings.

For HIE grading, a binary classification is conducted to distinguish between ECG patterns indicating whether TH is not required (Category 0) or required (Category 1). The aim is to support clinicians in decision-making regarding requirement of TH for neonates with HIE. Finally, for each 1-hour epoch, the predicted probabilities of all test samples were averaged, and this mean probability determined the category 0 (no need for TH) or 1 (TH is required) for that epoch. Therefore, comparison with expert annotations for each 1-hour ECG epoch will be feasible.

For the seizure detection task, the models classified each segment as seizure or non-seizure.

For HIE grading, patient independent 10-fold cross-validation was performed, with each fold comprising 12 patients, to evaluate the model on 120 patients,

For seizure detection, a patient-independent leave-one-out cross validation was employed on all 14 patients to evaluate the performance of the classification algorithm on the test set (13 patients).

Table 3.1: Optimized hyperparameters of the RF model for HIE grading, determined through grid search. Sqrt is square root.

Parameter	Value	Parameter	Value
max features	sqrt	Number of estimators	1000
Max depth	7	Min sample split	10
mean sample leaf	5	criterion	gini

Table 3.2: Optimized hyperparameters of the XGBoost model for HIE grading, determined through grid search.

Parameter	Value	Parameter	Value
max depth	7	Number of estimators	800
colsample bytree	0.9	Subsample	0.07
min child weight	7	tree method	hist

3.9.1.1 HIE grading

The optimal hyperparameters identified for all folds for each classifier based on the training dataset (108 patients) are presented in Table 3.1, Table 3.2 and Table 3.3.

3.9.1.2 Seizure detection

Here, seizure segments were over-sampled using SMOTE to help solve the imbalance data issue for all classification models, but not IF since it is an outlier detection model. Therefore, for the training dataset, the SMOTE technique was applied to oversample seizure segments, ensuring that their count became equal to that of the non-seizure segments.

Table 3.3: Optimized hyperparameters of the SVM model for HIE grading, determined through grid search.

Parameter	Value	Parameter	Value	Parameter	Value
Kernel	rbf	Gamma	90	C	700

Table 3.4: Optimized hyperparameters of the RF model for seizure detection, determined through grid search.

Parameter	Value	Parameter	Value
max features	sqrt	Number of estimators	1000
Max depth	10	Min sample split	10
mean sample leaf	5	criterion	gini

Table 3.5: Optimized hyperparameters of the XGBoost model for seizure detection, determined through grid search.

Parameter	Value	Parameter	Value
max depth	6	Number of estimators	700
colsample bytree	0.7	Subsample	0.7
min child weight	3	tree method	hist

The optimal hyperparameters identified for each classifier using grid search on the train dataset (13 patients) are presented in Table 3.4, Table 3.5, Table 3.6 and Table 3.7.

Since the annotations are based on EEG signals and the labels are approximated for each segment, it is important to account for signal changes occurring both before and after each segment. To address this, a post processing step is applied using a moving average filter on the model's predicted class probabilities for seizure detection. The moving average works by replacing each prediction with the average of predictions over a short temporal window around it, effectively smoothing rapid fluctuations and emphasizing consistent trends in the signal. This helps reduce noise, account for gradual changes, and provide more reliable detection of

Table 3.6: Optimized hyperparameters of the IF model for seizure detection, determined through grid search

Parameter	Value	Parameter	Value
max features	10	Number of estimators	1000
Max samples	2000	boot strap	True

Table 3.7: Optimized hyperparameters of the SVM model for seizure detection, determined through grid search

Parameter	Value	Parameter	Value	Parameter	Value
Kernel	rbf	Gamma	0.06	C	100

seizure events across consecutive segments.

3.9.2 Results

3.9.2.1 HIE grading

The performance of all classifiers was evaluated using 10-fold cross-validation, and the results were averaged across all folds. The final performance metrics are presented in Table 3.8. Among the classifiers, XGBoost with all sixteen features attained the highest AUC of 77.75%. Since XGBoost achieved the best performance, it was also applied to the subset of seven features (referred to as XGBoost2) previously analyzed by Goulding [?]. XGBoost2 attained a lower AUC of 76.36%, indicating that the additional features selected in this thesis, contribute to improved model performance. Random Forest achieved an AUC of 77.13% (slightly below XGBoost), while SVM produced the lowest AUC at 75.69%. This suggests that tree-based ensemble methods (Random Forest and XGBoost) better capture the complex, nonlinear HRV patterns and feature interactions relevant to seizure detection than the SVM in this setting.

Comparing the results of this thesis with AUC of 81% reported by Ahmed et al. [?], the AUCs are lower. However, their study was conducted on a much smaller dataset (54 patients compared to 120 in this thesis) and relied on leave-one-out cross-validation, which is known to produce optimistically biased estimates. In addition, their approach used Gaussian Mixture Models and Supervectors, whereas this thesis work employs features grounded in a range of state-of-the-art studies.

Table 3.8: Hypoxic-Ischemic Encephalopathy grading results of applied method.

Classifier	Sensitivity(%)	Specificity(%)	AUC(%)
Random Forest	75.38	59.62	77.13
Support Vector Machine	65.05	69.58	75.69
XGBoost	74	72.42	77.75
XGBoost2	76.77	63.67	76.36

These differences in dataset size, validation strategy, and feature design likely contribute to the variation in performance.

Since XGBoost achieved the best performance, its feature importance was computed and visualized. The feature importance values remained largely consistent across folds; therefore, the results from a single representative fold are presented in Fig. 3.5. As depicted, muNN emerged as the most important feature, followed by TriIndex, CSI and LF power. In the study by Goulding et al. [**?**], which evaluated seven HRV features, MuNN, LF power, and TINN were highlighted as particularly important for distinguishing between mild (category 0, TH not required) and moderate (category 1, TH required) HIE severity. In the present work, TriIndex, and SDI is introduced as additional features, both of which also demonstrate notable importance for classification.

Fig. 3.6 presents KDE plots for each feature, comparing samples labeled as category 0 and category 1 in the HIE grading dataset. Among these, muNN, identified as the most important feature in the classification process 3.5, shows the most separation between the two classes, as its KDE curves overlap less compared to other features. Similarly, in the study by Goulding et al. [**?**], the muNN feature was identified as particularly important for distinguishing between mild (category 0, TH not required) and moderate (category 1, TH required) HIE, as well as between mild (category 0, TH not required) and severe (category 1, TH required) HIE. In addition, the CVI and ApEn features introduced in this thesis also show modest

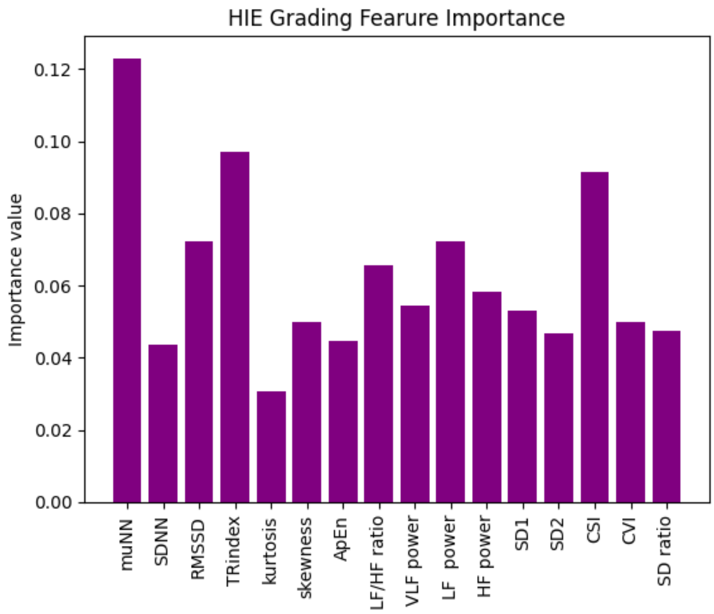


Figure 3.5: Feature importance of XGBoost classifier regarding HIE grading.

separation between the corresponding severity classes.

Fig. 3.7 presents the UMAP visualization of the test dataset, with eight 2D projection. The goal is to examine how well the full feature set separates the two classes. As seen, the separation between the classes remains visible, indicating that the extracted features provide meaningful discriminatory information. The visible class separation in the UMAP plot supports the KDE findings by showing that, although some features individually provide only modest discrimination, their combined structure yields a clearer distinction between category 0 and category 1.

3.9.2.2 Seizure detection

The performance of all classifiers was evaluated using LOOCV, and the results were averaged across all iterations. The final performance metrics are presented in Table 3.8. Among the models, the Random Forest (RF) using all 16 features achieved the highest AUC (69.21%). For comparison with the study by Goulding

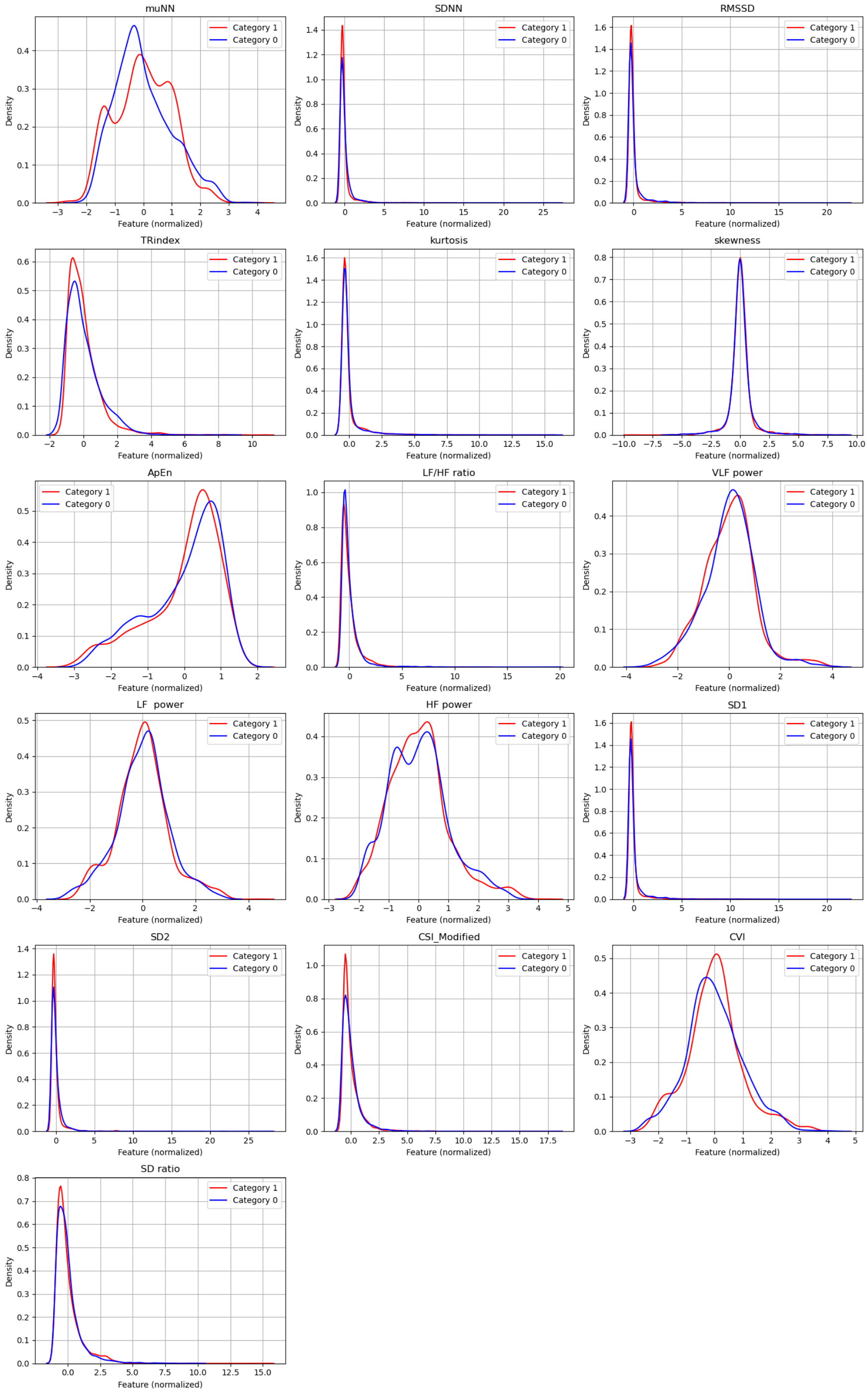


Figure 3.6: KDE plot of HIE category 0 and category 1 features.

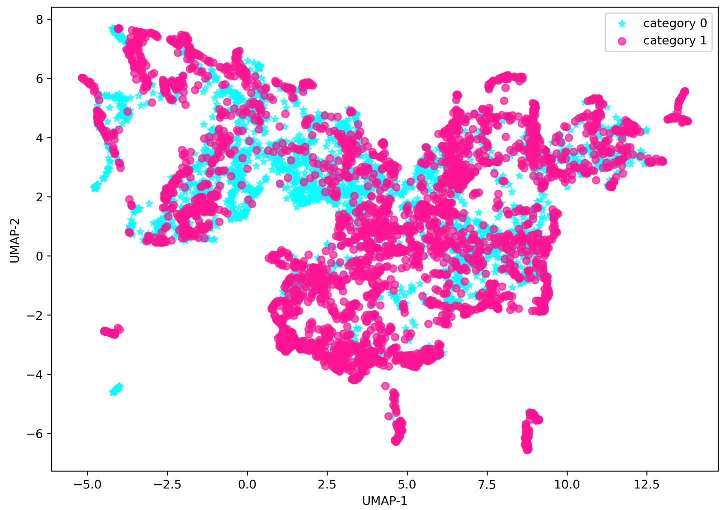


Figure 3.7: UMAP 2D projection for two classes of HIE grading test dataset.

et al. [?], the RF model was also trained using only the seven features examined in their work (referred to as RF2), resulting in a lower AUC of 66.13%. XGBoost and SVM achieved lower AUCs of 64.22% and 62.98%, respectively. The Isolation Forest achieved the lowest AUC (53.16). Although it showed very high specificity, its sensitivity was notably low. This reflects a high false-negative rate, suggesting that seizure segments were not sufficiently distinct from non-seizure segments for the model to reliably identify them as outliers.

Additionally, the results after post-processing (as described in the Experiment Seizure section) show an overall improvement in performance, with the SVM achieving the highest AUC of 71.3%. However, sensitivity decreased across all models. The AUCs for both SVM and XGBoost increased compared to the results obtained without post-processing. However, the overall reduction in sensitivity is mainly attributable to the high number of false positives generated by these models, particularly by the XGBoost classifier.

The results are comparable to other state-of-the-art methods [?] [?] [?], with AUCs generally in the 60%+ range and low sensitivity.

Table 3.9: Seizure detection performance of all classifiers using LOOCV, shown before and after post-processing.

Evaluation			
Classifier	Sensitivity (%)	Specificity (%)	AUC (%)
Random Forest	48.34	82.42	69.21
Random Forest2	57.2	70.29	66.13
Support Vector Machine	36.34	82.84	62.98
Isolation Forest	11.79	90.62	53.16
XGBoost	65.03	58.47	64.22
Evaluation After Post Processing			
Classifier	Sensitivity (%)	Specificity (%)	AUC (%)
Random Forest	44.15	89.23	70.94
Random Forest2	54.29	75.79	67.97
Support Vector Machine	26.11	92.06	71.3
Isolation Forest	3.8	94.52	53.08
XGBoost	4.23	99.15	67.74

Since RF had the best result, the seizure detection result of each patient is provided in Table 3.10. The results show variability in model performance across patients: 7 out of 14 achieved an AUC above 60%. Similar to a previous study by Doyle [**?**], performance also varied across individual patients.

Again, since RF achieved the best performance, its feature importance was computed and visualized. The feature importance values remained largely consistent across folds; therefore, the results from a single representative fold are presented in Fig. 3.8. As shown, LF Power holds the highest feature importance for seizure detection, followed by the LF/HF ratio, muNN, ApEn, CVI, and SD ratio. Interestingly, LF Power also showed the lowest P-value (0.055) among the seven features evaluated by Goulding [**?**] during both seizure and non-seizure conditions, underscoring its statistical relevance. Additionally, muNN showed a trend toward a significant decrease (p = 0.059) during subclinical seizures (seizures without observable clinical symptoms) compared to the baseline in the same study. Moreover, three of the newly introduced features in this study, ApEn, CVI, and SD

Table 3.10: Random forest seizure detection result of LOO per patient reported as True negative, false positive, false negative, true positive, sensitivity, specificity and AUC.

Patient number	Confusion Matrix Results				Performance Metrics (%)		
	TN	FP	FN	TP	Sens (%)	Spec (%)	AUC (%)
1	4368	897	27	6	18.18	82.96	47.93
2	3774	480	18	30	62.5	88.72	80.56
3	7215	576	177	33	15.71	92.61	61.1
4	4008	195	111	0	0	95.36	48.73
5	2217	243	33	12	26.67	90.12	49.73
6	8601	1440	135	93	40.79	85.66	67
7	8190	288	201	69	25.56	96.6	81.65
8	5769	3666	900	1734	65.83	61.14	69.66
9	3522	174	219	0	0	95.29	34.57
10	3972	1110	138	75	35.21	78.16	65.74
11	7467	1260	369	129	25.9	85.56	54.22
12	2145	1350	144	339	70.19	61.37	68.93
13	1194	810	24	9	27.27	59.58	31.38
14	3660	1614	216	9	4	69.4	21.41

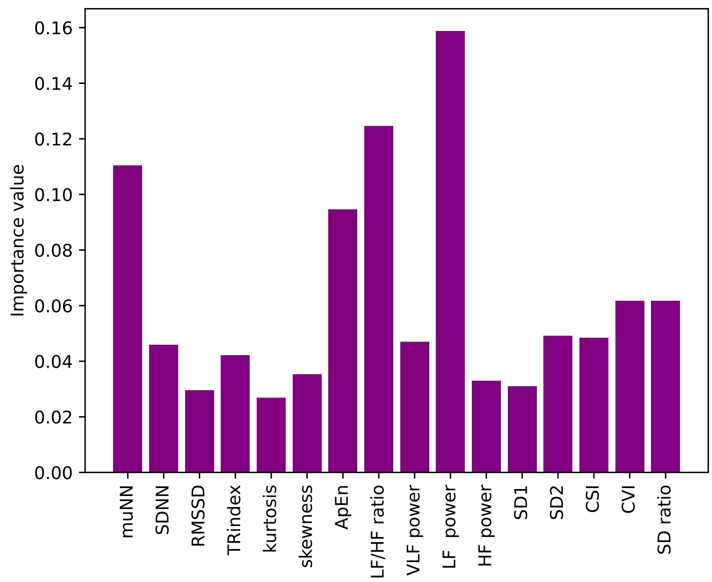


Figure 3.8: Feature importance of Random Forest classifier regarding seizure detection.

ratio, rank among the top six in importance, contributing to the RF classification task.

Fig. 3.9 illustrates KDE plots for each feature, comparing samples labeled as baseline (class 0) and seizure (class 1) in the seizure detection dataset. As shown, the features with higher importance—such as ApEn, CVI, muNN, and LF Power—exhibit less overlap between seizure and non-seizure classes, indicating better separability. Among these, muNN and LF Power were also identified as significant for distinguishing seizure from non-seizure segments in the study by Goulding [**?**], while the other two features, introduced in this thesis, also appear effective for the seizure detection task. This visual trend supports their stronger contribution to the classification task. However, the KDE plots also reveal that, overall, there is substantial overlap in the feature distributions between the two classes. This appears to contribute to the relatively low AUC values observed in the results.

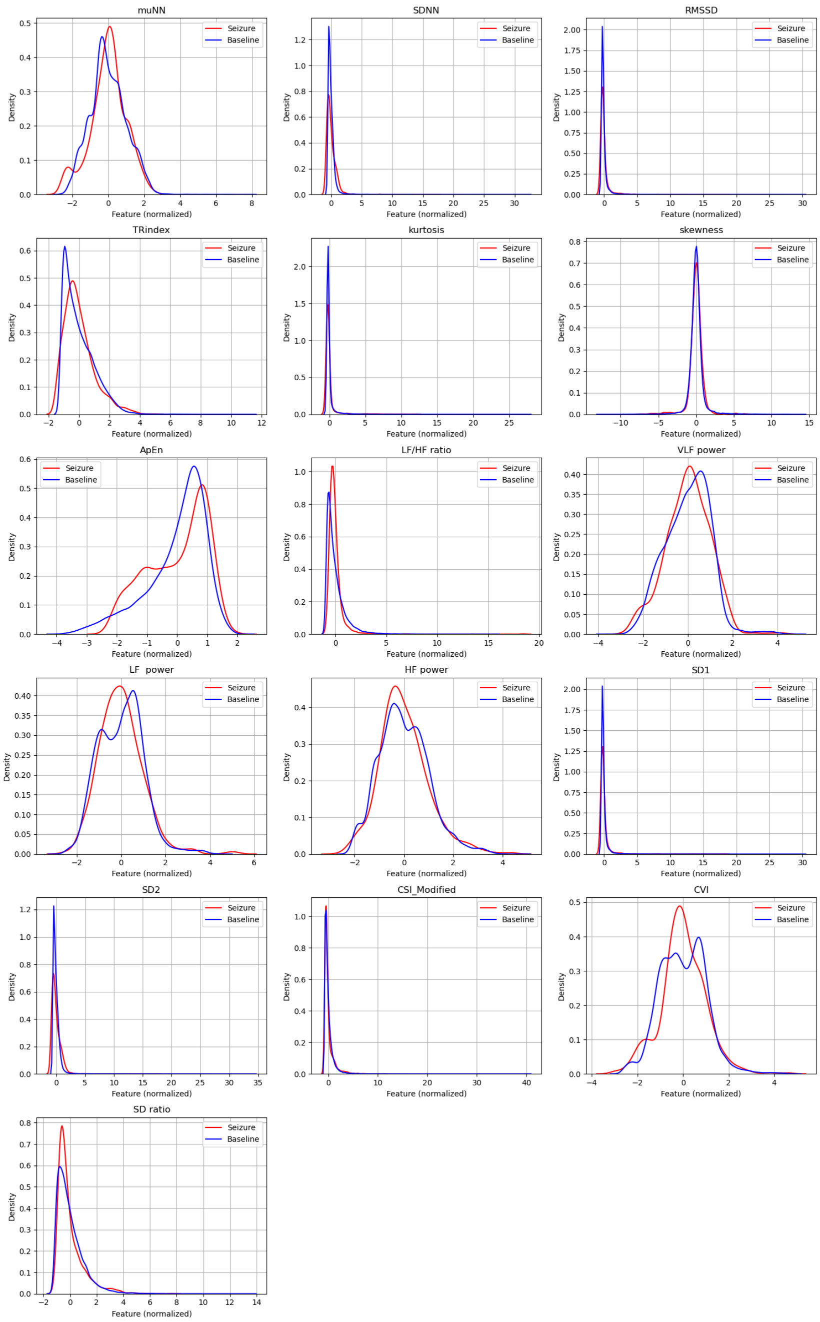


Figure 3.9: KDE plot of seizure and non-seizure features.

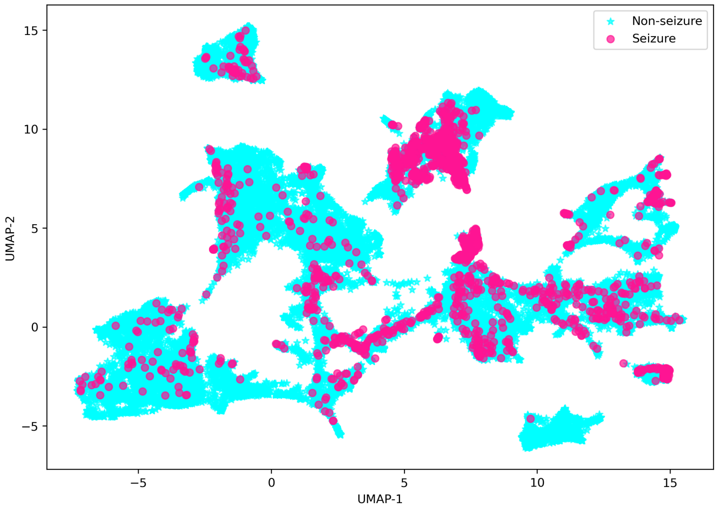


Figure 3.10: UMAP 2D projection of seizure detection dataset.

Fig. 3.10 shows a 2D UMAP visualization of the entire seizure dataset. As seen, there is no clear or consistent separation between the seizure and non-seizure (baseline) classes, both are heavily overlapped in the embedding space. The limited class separation, combined with a pronounced class imbalance, underscores the difficulty of distinguishing between the two conditions using the available features and helps explain the limited performance of the classification models on this dataset.

3.10 Using a larger dataset (ANSeR Dataset)

This time, the analysis utilizes a very large and diverse dataset (ANSeR) which was described in detail in Chapter 2, section 2.9, to investigate how dataset size and diversity influence the effectiveness of machine learning in this context. A larger dataset would also enable the use of an MLP, a deep learning model, to evaluate its effectiveness in this task.

3.10.1 Experimental setup

Since the ANSeR HIE dataset was substantially larger than the EarlySet HIE dataset used in Section 3.9, a four-minute segment length was selected (longer than the two-minute segments used for EarlySet HIE dataset) to capture greater variability in the HRV signal. For the seizure detection task, where seizure events were short and less frequent than non-seizure events, a two-minute segment length was used, matching the segment length of the Cork seizure dataset used in Section 3.9, to ensure a higher number of seizure events could be included. To better capture continuous signal dynamics, a higher overlap of 80% between consecutive segments was applied. Similar to Section 3.9, each two-minute segment was labeled as a seizure if at least two-thirds of its duration contained seizure activity. However, since ANSeR dataset used in this section is larger, Segments with less than two-thirds seizures activity were discarded as such mixed patterns (seizure and non-seizure) in one segment may not represent pure seizure patterns and reduce the model's ability to learn distinct seizure-specific HRV features. While, those with no seizure activity were labeled as non-seizure.

Since XGBoost achieved the best performance for HIE grading and RF for seizure detection, they were applied to the respective tasks here as well. Additionally, an MLP was included to evaluate the potential benefits of deep learning, which is feasible given the sufficiently large dataset.

In all cases, the data was preprocessed using a Hampel filter. Further the data was standardized before training the MLP classifier to ensure all features had zero mean and unit variance.

As before, the XGBoost and RF hyperparameter optimization was performed separately using grid search. The MLP model weights were initialized using the Xavier uniform method. The model was trained using the cross-entropy loss function with a batch size of 1024 for a maximum of 150 iterations. Optimization was carried out using the Adam optimizer with an initial learning rate of 0.0005, and a weight decay of 0.001 was applied for regularization. Internal cross-validation was performed to monitor performance and implement early stopping.

Since the datasets were large and diverse, model performance was evaluated in a patient-independent manner without extensive cross-validation, as a limited number of folds was sufficient to obtain reliable estimates.

For HIE grading, classifiers were trained on the ANSeR2 dataset (73 patients) and tested on the completely patient independent ANSeR1 dataset (58 patients). A binary classification is conducted to distinguish between ECG patterns indicating whether TH is not required (Category 0) or required (Category 1). Finally, for each 1-hour epoch, the predicted probabilities of all test samples were averaged, and this mean probability determined the category 0 (no need for TH) or 1 (TH is required) for that epoch. In total, 111,835 4-minute segments were extracted from ANSeR2, of which 54,394 were category 1. For ANSeR1, 15,688 segments were extracted, with 6,222 classified as category 1.

For seizure detection, models were trained on ANSeR1 (42 patients) and tested on the separate ANSeR2 dataset (23 patients), ensuring that all test data remained entirely unseen and patient independent during training. A total of 366,222 segments were extracted, of which 12,434 corresponded to seizures.

In the seizure detection task, model performance on the ANSeR2 dataset, which included 23 patients, was suboptimal. To address this, a 5-fold patient-independent cross-validation across both ANSeR1 and ANSeR2 datasets with an 80/20 split was performed for both models, with each fold containing 32 patients for training and 13 patients for testing. This approach increased the amount of data avail-

Table 3.11: Optimized hyperparameters of the XGBoost model for HIE grading, determined through grid search

Parameter	Value	Parameter	Value
Max depth	4	Number of estimators	200
colsample bytree	0.7	Subsample	0.7
min child weight	5	tree method	hist

Table 3.12: Optimized hyperparameters of the RF model for seizure detection, determined through grid search

Parameter	Value	Parameter	Value
max features	sqrt	Number of estimators	1000
Max depth	10	Min sample split	5
mean sample leaf	5	criterion	gini

able for training while still preserving a sufficient test set. The goal was to assess whether training on more data and incorporating additional seizure events could improve performance, although the dataset remained imbalanced. All folds were evaluated and compared to ensure a comprehensive and fair analysis.

3.10.1.1 HIE grading

The final selected hyperparameters for XGBoost using grid search method are presented in Table 3.11.

3.10.1.2 Seizure detection

For the training dataset, the SMOTE technique was applied to oversample seizure segments, ensuring that their count became equal to that of the non-seizure segments.

Selected hyperparameters for the RF model using randomized grid search are presented in Table 3.12.

Table 3.13: Hypoxic-Ischemic Encephalopathy grading results of applied method.

Classifier	Sensitivity(%)	Specificity(%)	AUC(%)
XGBoost	86.05	68.46	79.26
MLP	82.56	65.39	79.34

3.10.2 Results

3.10.2.1 HIE grading

The performance of both classifiers was evaluated for HIE grading and the results are provided in Table 3.13. THE AUC for XGBoost and MLP are 79.26% and 79.34% respectively which are quite close whereas MLP performs slightly better. The results show that the model's performance improves on the larger ANSeR dataset compared to the EarlySet HIE dataset experiment conducted in section 3.9.2.1 which achieved an AUC of 77.75. As before, results are based on more variable and larger dataset comparing to result by Ahmed et al. [**?**].

For both models, the confusion matrix is calculated and illustrated in Fig. 3.11. The confusion matrix values show that XGBoost outperforms MLP, with higher true positives, true negatives, and lower false positives and false negatives across all metrics.

3.10.2.2 Seizure detection

RF and MLP models were trained on ANSeR1 dataset and tested on ANSeR2 dataset. The evaluation results on the test set are shown in Table 3.14. The performance on the ANSeR large dataset is lower than on the Cork seizure dataset because, in the Cork seizure dataset, LOOCV tests the model on a single patient per fold (since the dataset size was smaller). With very few seizure events per patient, this can inflate performance metrics. In contrast, the larger datasets, with

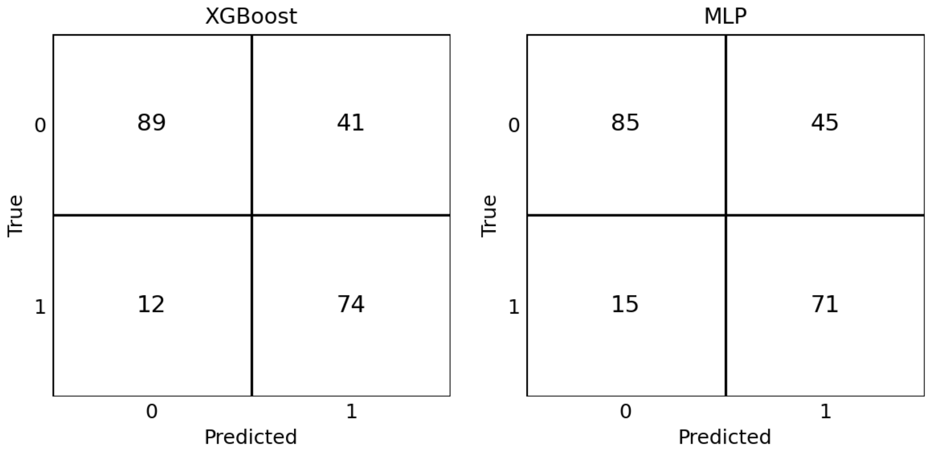


Figure 3.11: Confusion matrix of XGBoost and MLP classifiers regarding HIE grading.

Table 3.14: Test results on ANSeR2 dataset for seizure detection using RF and MLP.

Model	Sensitivity (%)	Specificity (%)	AUC (%)
RF	50.5	64.09	60.04
MLP	30.69	78.35	55.93

two separate patient independent train (ANSeR1) and test (ANSeR2) datasets, provides a more realistic evaluation across many patients, revealing the model's true generalization ability, especially in highly imbalanced data.

For both models, the confusion matrix is calculated and presented in Fig. 3.12

Further, seizure detection has been applied on all the datasets using 5-Fold patient independent cross validation. The result of all folds for both RF and MLP are shown in Tables 3.15 and 3.16. The results indicate the performance of the models varied across different splits; however, the overall seizure detection outcomes were in the same range as the ANSeR1 and ANSeR2 split.

Figure. 3.13 is a visualization of the evaluation result using a radar chart. The results show that the MCC metric is consistently low across folds, which reflects

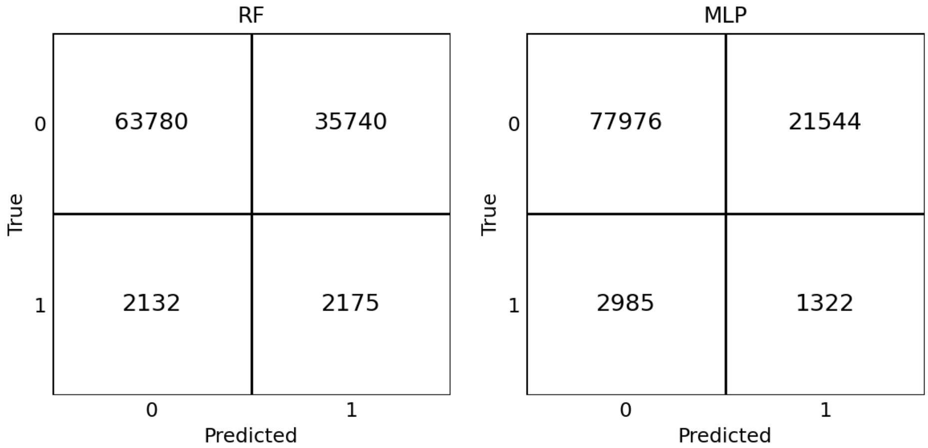


Figure 3.12: Confusion matrix of RF and MLP classifiers regarding seizure detection.

Table 3.15: 5-Fold cross validation result for seizure detection using RF.

Fold	TN	FP	FN	TP	Sensitivity (%)	Specificity (%)	AUC (%)
0	34636	20615	322	434	57.41	62.69	65.84
1	76618	21695	2229	1753	44.02	77.93	67.23
2	32698	22982	962	874	47.6	58.72	54.35
3	40663	22224	1013	1104	52.15	64.66	63.54
4	60697	20960	2508	1235	32.99	74.33	55.91
Average	**49062**	**21695**	**1406**	**1080**	**46.43**	**67.27**	**61.77**

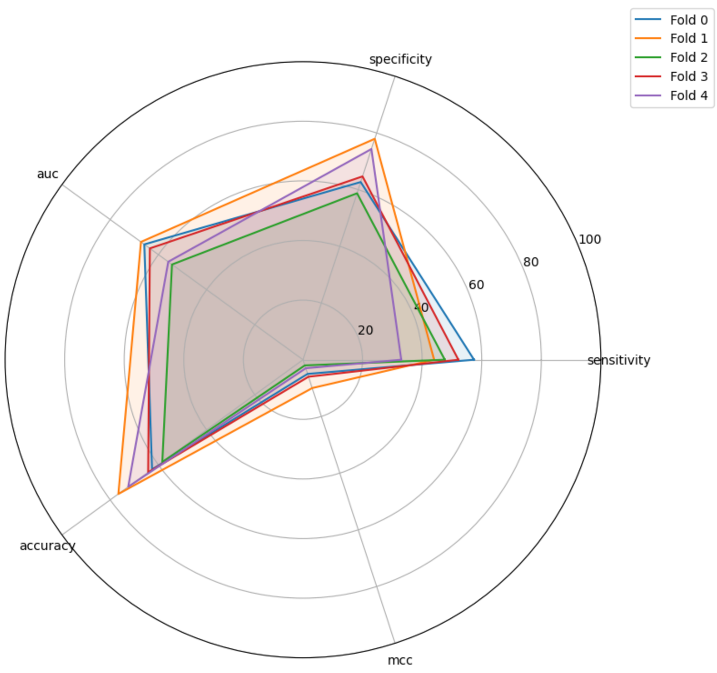
Table 3.16: 5-Fold cross-validation results for seizure detection using MLP.

Fold	TN	FP	FN	TP	Sensitivity (%)	Specificity (%)	AUC (%)
0	40395	14856	441	315	41.67	73.11	60.37
1	75448	22865	2350	1632	40.98	76.74	63.18
2	38291	17389	1139	697	37.96	68.77	54.11
3	45449	17438	1164	953	45.02	72.27	62.66
4	59177	22480	2546	1197	31.98	72.47	52.52
Average	**51732**	**19026**	**1528**	**958**	**39.92**	**72.47**	**58.57**

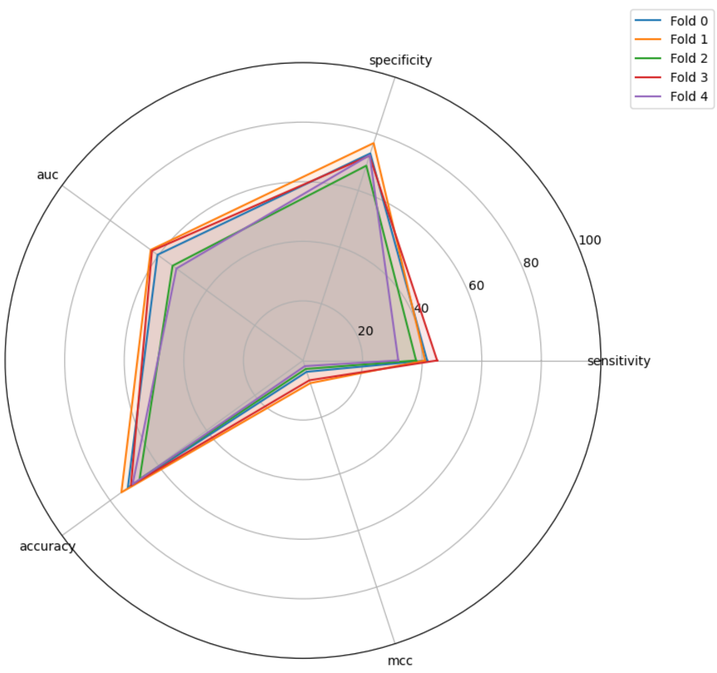
the strong class imbalance in the test dataset. Additionally, sensitivity is notably lower than specificity. This indicates that the model is better at correctly identifying non-seizure segments than detecting seizure segments. These patterns reflect the underlying class imbalance and suggest that the decision boundary may be biased toward the majority (non-seizure) class.

To better understand the performance of the RF and MLP models in relation to dataset diversity and class imbalance, the AUC of each model in relation to the number and length of seizure events in each fold's test set was analyzed. Here, a seizure event refers to the full seizure segment from the long recording, not the 2-minute segmented seizures. For each fold, the average of the maximum seizure event lengths per patient and the number of events longer than 10 minutes were calculated. The duration of a seizure event is important because prolonged seizures are associated with more pronounced alterations in cerebral blood flow and autonomic regulation. As a result, longer seizures can have a greater impact on heart rate.

As shown in Table 3.17, fold 1, which achieved the highest AUC for both models, also had the longest seizure events on average. This fold includes two of the longest seizure events in the entire dataset (116 minutes and 44 minutes). Long seizure events capture more pronounced changes in heart rate, making it easier for the models to detect these patterns. Conversely, fold 2 had the lowest AUC for both models, suggesting that shorter seizure events make it harder for the models to detect meaningful changes. Fold 4 illustrates another scenario: although its average seizure length is not particularly low, the high number of seizure events reduces the length of individual events, again making detection more difficult for the models.



(a) RF



(b) MLP

Figure 3.13: Radar chart comparing the evaluation results of RF and MLP regarding seizure detection for all 5 folds.

Table 3.17: Relationship between seizure event characteristics and seizure detection model performance across folds.

Fold	Model AUC (%)		Seizure Event Analysis	
	MLP	RF	average of max seizure event length per patient (min)	number of events longer than 10 mins
0	60.37	65.84	11.26	7
1	**63.18**	**67.23**	**25.6**	23
2	54.11	54.35	10.12	13
3	62.66	63.54	18.32	24
4	52.52	55.91	14.86	52

3.11 Conclusion

The experimental results highlight the critical role of specific HRV features in both HIE grading and seizure detection. In particular, muNN was confirmed as an important feature across both tasks. Its relevance is supported by previous findings: Goulding et al. [?] demonstrated its utility in distinguishing between mild and moderate HIE severity, while a later study by Goulding [?] reported a trend toward a significant decrease (p = 0.059) in muNN during subclinical seizures compared to baseline. As muNN represents the mean of RR intervals, serving as the primary indicator of patient-specific heart rate—its importance in classification tasks is well justified.

Similarly, LF power showed strong contributions to classification in both HIE grading and seizure detection. It was previously identified as a discriminative feature for separating mild from moderate HIE severity [?], and in Goulding's subsequent study [?], it yielded the lowest p-value (0.055) among the seven features evaluated under both seizure and non-seizure conditions, underscoring its statistical significance. These findings emphasize that substantial information is contained within the low-frequency components of the HRV signal, making them particularly effective for machine learning–based classification.

Beyond muNN and LF power, additional features introduced in this study, such as TriIndex, CSI and ApEn, also proved effective in improving classifier decisions. Previous studies also support their usefulness, showing that these features are broadly relevant in ECG analysis.

HIE grading performance on a completely unseen, patient-independent test set increased from 77.75% to 79.34% when trained on a larger dataset (ANSeR dataset). Compared to the previous work by Ahmed et al. [?], which reported an AUC of 81%, the results in this thesis are slightly lower but likely more reliable. This is because the present study was conducted on a larger dataset and uses a more robust evaluation strategy, whereas the LOOCV approach used by Ahmed et al. is known to produce biased and overly optimistic performance estimates. Importantly, this represents a promising step toward developing HIE assessment tools to support clinicians in decision-making. By leveraging ECG signals, such tools could be deployed even in resource-limited hospitals, enabling timely treatment for affected infants.

Seizure detection performance reduced from 69.21% on the smaller dataset to 60.04% on the larger one, suggesting that the results obtained using Leave-One-patient-Out Cross-Validation on the small dataset may not generalize well to unseen data. Results of this thesis are slightly higher than those reported in other state-of-the-art studies [?] [?] [?]. Moreover, the use of a larger and more diverse dataset in this thesis, reduces the risk of overfitting and makes the performance estimates more representative of real-world conditions.

Further analyses on alternative test sets with different patients using 5-fold cross validation, revealed considerable variability across datasets, underscoring uncertainty in the model's performance. This variability is likely influenced by several factors:

- Class imbalance: The low number of seizure segments compared to non-seizure segments limits the model's ability to generalize, even after oversampling. The model may overfit to non-seizure features since they dominate the training set.

- Event duration: Short seizure segments often do not exhibit enough distinct changes in the signal for effective model training, whereas the model performs better for longer seizure events.

- Label alignment: The seizure labels were annotated by experts based on EEG recordings, whereas the model used in this thesis is trained and evaluated on ECG signals. It is possible that seizure-related changes in ECG occur earlier or later than the annotated onset in the EEG. This can contribute to false positives and reduce accuracy.

- Limited class separation in feature space: Data visualizations, including KDE plots and UMAP projections, show limited separation between seizure and non-seizure segments, highlighting the inherent difficulty of this task. However, some of this overlap may also be attributed to the mismatch between EEG-based labels and ECG signal changes.

- Seizure pattern variability: Variations in seizure types such as local or global can lead to different signal patterns, making model detection less consistent.

Overall, while HRV-based HIE grading shows strong potential, seizure detection remains challenging due to data imbalance, variable event lengths, and the inherent differences between ECG and EEG signals.

4 | HIE Grading and Seizure Detection Using ECG Spectrogram

4.1 Introduction

In this chapter, an alternative way of analysing the ECG signal is introduced for neonates with different grades of HIE and seizure. Here, the ECG signal is investigated using ECG spectrogram and Convolutional Neural Network for both HIE grading and seizure detection. This method is proposed to overcome the limitations and shortcomings of the heart rate calculation, feature extraction and classical classification approaches. In this method, the spectrogram is used to directly compute the time-frequency information of the ECG signal. Additionally, spectrogram-based representations enable the application of deep learning models like CNNs, which offer notable advantages over classical methods in terms of accuracy. Unlike classical methods that depend on hand-crafted features, CNNs automatically extract meaningful patterns directly from raw data. Their architecture is particularly well-suited for identifying spatial features, and their built-in translation invariance enhances robustness by enabling consistent recognition even when input signals shift or contain noise.

No previous studies appear to have investigated ECG time-frequency representations in relation to HIE. To the best of my knowledge, this is the first study utilizing ECG spectrograms for neonatal HIE grading.

This chapter includes sections explaining spectrograms and CNNs. Furthermore, the computation of ECG spectrograms and the proposed CNN architectures for HIE grading and seizure detection are described. Each task is addressed separately, with independently designed CNN models. The methods are then evaluated through experiments on two distinct ANSeR datasets: the HIE ANSeR dataset and the seizure ANSeR dataset, as introduced in chapter 2, section 2.9.

4.2 Spectrogram

Spectrograms are commonly used for signal representation, capturing the changes in a nonstationary signal's frequency content over time using consecutive Fourier transforms particularly in domains such as audio and speech recognition [**?**]. ECG spectrograms have proven to be useful in tasks such as estimating heart frequency [**?**], arrhythmia classification [**?**] and classifying the severity of tetanus [**?**]. Infections and arrhythmia detection may represent a relatively simpler classification challenge using ECG compared to the detection of neurological conditions.

A spectrogram provides a visual representation of how a signal's frequency content evolves over time. Typically, it is displayed as a two-dimensional plot where time is shown on one axis and frequency on the other. The amplitude or power of each frequency at a given moment is conveyed through color or intensity, adding a third dimension to the visualization.

The signal is divided into overlapping frames to minimize boundary artifacts. Each frame is windowed and transformed using the discrete Fourier transform, producing complex values representing magnitude and phase across time and frequency. This process is defined as the short-time Fourier transform (STFT):

$$\text{STFT}[m, \omega] = \sum_{n=0}^{N-1} x[n]\, w[n-m]\, e^{-j\omega n} \tag{4.1}$$

where $x[n]$ is the discrete ECG signal, $w[n]$ is the window function (a Hamming window is used in this thesis), m indexes the time frames, and ω represents the continuous angular frequency variable. The spectrogram, which estimates the power spectral density, is defined as the squared magnitude of the STFT:

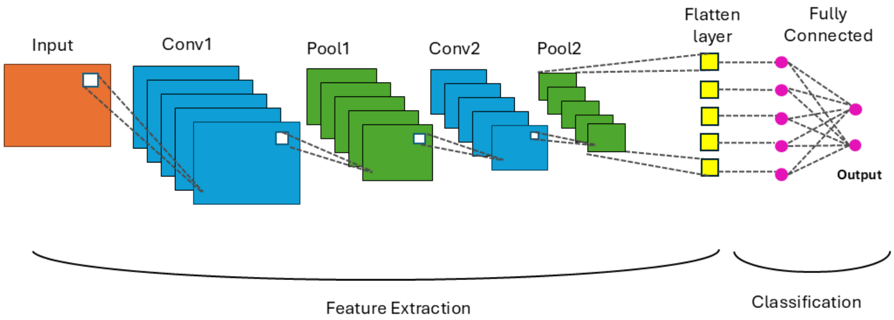


Figure 4.1: Architecture of a simple CNN model for image classification. Feature extraction is performed using two convolutional layers and two pooling layers. The extracted features are then flattened and passed through a fully connected network to produce a two-class output.

$$\text{Spectrogram}(m, \omega) = |\text{STFT}[m, \omega]|^2 \tag{4.2}$$

In this chapter, spectrograms were used as input representations to the CNN classifier, capturing the time–frequency characteristics of ECG signals. The spectrogram transforms a one-dimensional ECG signal into a two-dimensional matrix, where each element represents the signal's power spectral density (PSD) at a specific frequency and time point.

4.3 Convolutional Neural network

Convolutional Neural Networks (CNNs) are a specialized class of neural networks designed to efficiently extract features from data with a grid-like structure, such as two-dimensional or three-dimensional images. CNNs are composed of multiple layers, including an input layer, convolutional layers, pooling layers, and fully connected layers. Fig. 4.1 shows the architecture of a simple CNN model.

The core operation of a CNN is convolution, which involves sliding small trainable

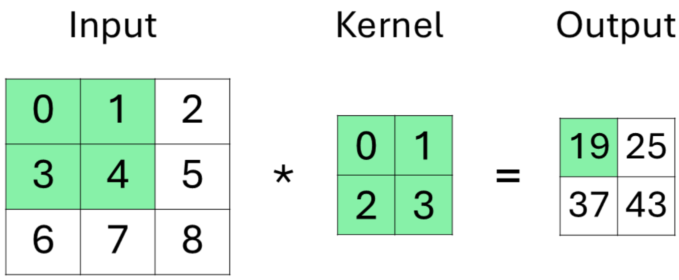


Figure 4.2: A basic example with a 2×2 kernel.

kernels (filters) across the input matrix to extract features. Each kernel performs an element-wise multiplication with the portion of the input it covers. For multi-channel inputs (e.g., RGB images), each kernel has the same depth as the input, and the element-wise multiplication is performed across all channels, with the results summed to produce a single scalar. A layer can have multiple kernels (filters), each generating a separate feature map; the number of filters determines the depth of the layer's output. The movement of the kernel across the input is controlled by the stride, which determines how many pixels the filter moves at each step, and padding, which determines whether borders are preserved or trimmed during the convolution.

Fig. 4.2 illustrates a simple convolution operation using a two-dimensional input of size 3×3 and a kernel of size 2×2. The convolution window is initially placed at the upper-left corner of the input. It then slides across the input from left to right and from top to bottom. At each position, the values within the current window are multiplied element-wise with the kernel, and the resulting products are summed to produce a single scalar value. This scalar corresponds to the output value at the associated spatial location in the feature map.

To reduce the spatial resolution while retaining essential information, max pooling is applied after the convolutional layer. Pooling layers downsample the feature

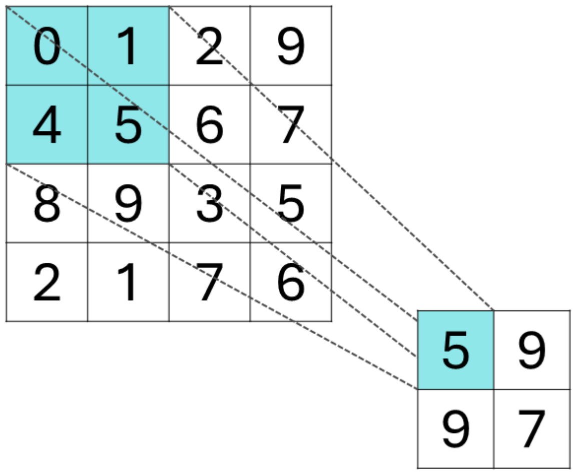


Figure 4.3: An example of max pooling, which reduces the feature map from 4×4 to 2×2.

maps, making the model more computationally efficient and translation invariant. Fig.4.3 shows an example of max pooling, which reduces the feature map from 4×4 to 2×2.

An essential element in CNNs is the activation function, which introduces non-linearity to allow the network to model complex relationships. The Rectified Linear Unit (ReLU) is the most commonly used activation function and is applied here after each convolution. As shown in Fig. 4.4, ReLU sets all negative values to zero while leaving positive values unchanged, enabling sparse and efficient representations. The ReLU function is provided below.

$$g(z) = \max(0, z) \tag{4.3}$$

Batch normalization is a technique used to improve the stability and speed of neural network training by normalizing layer activation using statistics computed over each mini-batch. It is typically applied after the linear or convolutional operation

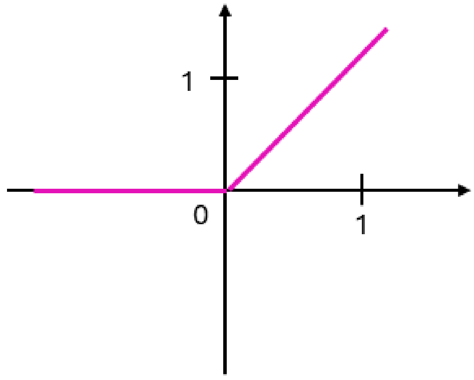


Figure 4.4: Plot of the Rectified Linear Unit (ReLU) activation function.

and before the nonlinear activation function. In fully connected layers, batch normalization normalizes the activation of each neuron across the mini-batch, while in convolutional layers it is applied on a per-channel basis, computing the mean and variance across both the batch and spatial dimensions of each feature map. Batch normalization enables faster convergence, allows higher learning rates, and often improves generalization.

The concept of the receptive field is critical in understanding CNN behavior. It refers to the region in the input that influences a single unit in a feature map. As more convolution and pooling layers are added, the receptive field increases, allowing deeper layers to capture more global contextual information. This gradual expansion of the receptive field is especially valuable for spectrogram analysis, where both local and long-range patterns may be relevant for classification.

After the earlier layers extract and refine the features, the model flattens all feature maps into a single long vector, which is then fed into a dense (fully connected) layer. Each unit in this layer connects to every element of the vector, allowing the model to learn how different features relate to one another. Once the fully connected layer processes these features, the output layer transforms the results into class probabilities, indicating the category the model predicts for the input image.

4.4 Handling class imbalance

Since the seizure detection dataset is highly imbalanced, with the number of seizure events being significantly lower than non-seizure segments, two strategies were employed to address this issue: class weighting and data augmentation.

A common approach to make neural networks cost-sensitive is to modify the training process such that higher-cost (minority) examples have a greater impact on weight updates. This can be achieved by adjusting the learning rate, applying class-specific weights in the loss function, or directly minimizing a misclassification cost instead of the standard loss [**?**] [**?**]. In this thesis, class weights were assigned to penalize misclassification of the minority class (seizure) more heavily during training. The weights were computed as follows:

$$w_c = \frac{N}{K \cdot n_c} \tag{4.4}$$

where:

- w_c is the weight assigned to class c,
- N is the total number of samples in the dataset,
- K is the total number of classes,
- n_c is the number of samples belonging to class c.

This, ensures that seizure samples contribute proportionally more to the loss function, helping the model avoid bias toward the majority non-seizure class.

Data augmentation is a widely used technique to reduce overfitting and enhance the generalization ability of deep learning models. By generating synthetic data, it increases both the size and diversity of a dataset, allowing models to learn more robust and transferable representations. A key application of data augmentation, especially in domains such as image, audio, and text processing, is addressing class imbalance. By creating additional samples for the minority class, augmentation can help balance the dataset and improve the effectiveness of the learning process [?] [?]. A broad range of augmentation strategies have been developed. Traditional methods include adding noise, applying color jitter, or using patch-based approaches such as CutOut and CutMix, which have achieved notable success in various fields. More recently, generative approaches leveraging models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models have emerged, enabling the synthesis of entirely new and realistic samples [?].

Here, data augmentation is of interest to increase the number of seizure samples for the seizure detection task. For ECG signal spectrograms, where temporal dynamics are critical, techniques like Mixup or CutMix may disrupt time–frequency relationships, limiting their suitability. Similarly, although generative models hold promise for producing synthetic ECG data, their effectiveness is constrained by the scarcity of seizure events needed for training. Since the lack of seizure data is the fundamental challenge, training such models becomes impractical. In this context, adding Gaussian noise remains one of the most practical and effective augmentation methods for ECG signals [?]. By introducing small, random perturbations into the signal values, this technique increases dataset variability while preserving temporal structure—making it particularly well-suited for time-sensitive data such as ECG spectrograms.

4.5 Proposed method using ECG spectrograms and CNN models

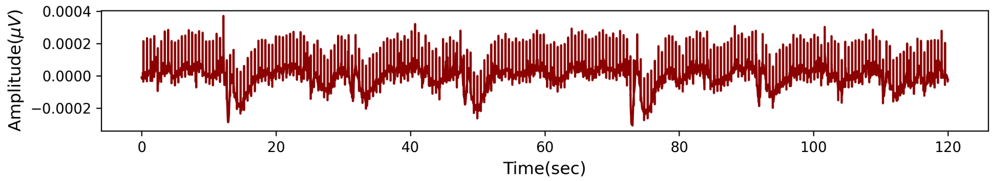
4.5.1 ECG preprocessing and spectrogram calculation

To eliminate power line interference, a notch filter was applied to the ECG signals. Baseline wander was removed by applying a second-order high-pass Butterworth filter with a cutoff frequency of 0.5 Hz to suppress low-frequency components.

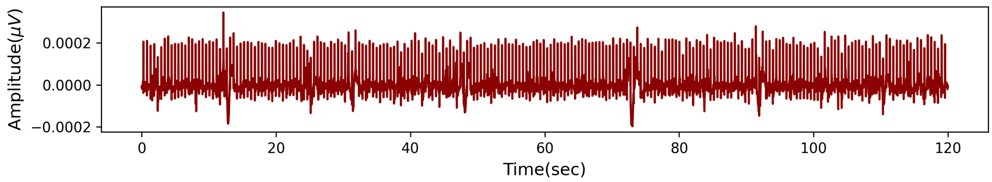
The filtered ECG signal were divided into four-minute segments for HIE grading and two-minutes segments for seizure detection and spectrograms were calculated for each segment with 80% overlap, consistent with the approach used in chapter 3 for heart rate feature extraction using the ANSeR dataset. Similar segment lengths were used because they capture sufficient abnormalities and changes in the signal, and they provide an adequate amount of data for analysis. For seizure detection, this segment length also offers a suitable window to include enough seizure events. In addition, using comparable segment lengths allows the results in this chapter to be directly compared with those presented in Chapter 3. Overlap was employed to reduce information loss at segment boundaries and provide a more comprehensive representation of the ECG signal patterns. Fig. 4.5 presents a noisy raw ECG signal and the filtered version, illustrating a 2-minute segment of the recording.

All spectrograms were computed using a Hamming window. Other window types, such as Tukey and Hanning, were also tested; however, they did not produce substantial differences in the visual appearance of the spectrograms or in the model's performance when evaluated on a subset of the dataset.

For the HIE grading spectrograms, a 28-second Hamming window with a 27-second overlap was used. A 28-second window contains a sufficient number of



(a)



(b)

Figure 4.5: Visualization of 2-minute segments of ECG signal: (a) Raw ECG signal; (b) ECG signal after filtering with a notch filter and a high-pass Butterworth filter.

ECG QRS peaks to capture meaningful signal variations, and the resulting spectrograms were visually inspected to confirm their suitability. In addition, the resulting spectrogram dimensions were well-suited for the CNN architecture used in this study. For the seizure detection spectrograms, a 2-minute ECG segment was used with a 16-second Hamming window and a 15-second overlap. These settings follow the same criteria as those used for HIE grading, providing enough QRS complexes to capture meaningful signal changes, producing visually acceptable spectrograms, and yielding spectrogram dimensions that are appropriate for the CNN architecture. All spectrogram power spectral densities were converted to a logarithmic scale to compress the dynamic range and enhance the visibility of low-power spectral components.

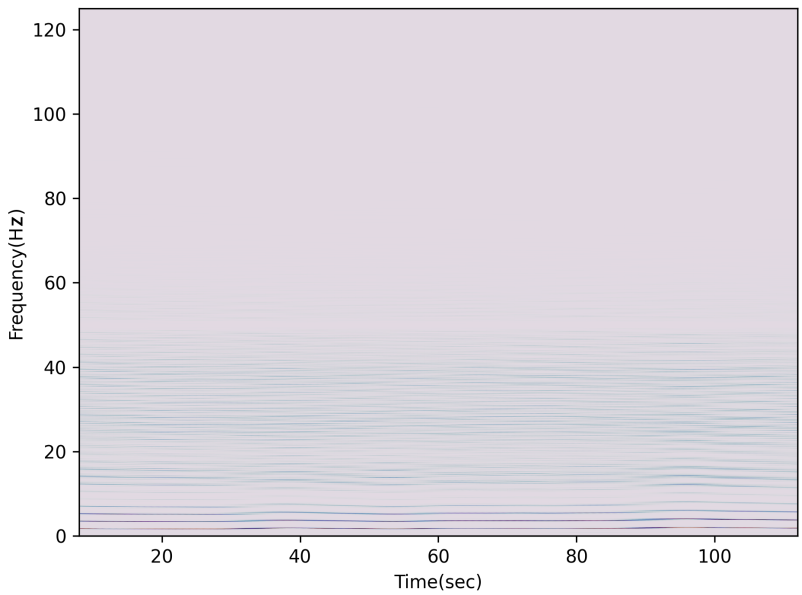
Fig. 4.6 (a) illustrates the spectrogram of a 2-minute ECG segment with a 16-second Hamming window and a 15-second overlap, calculated for the seizure detection task. As shown, the spectrogram appears mostly gray, with a few lines concentrated in the lower frequencies below 60 Hz, indicating that most of the signal's energy is contained in the low-frequency range. The magnitudes in the spectrogram are very small, with a minimum of 1.56×10^{-21}, a maximum of 6.72×10^{-10}, and a mean of 7.07×10^{-12}. Therefore, to highlight lower frequency components, the spectrogram's power spectral density was converted to a logarithmic scale. This, changes the values to a range with a minimum of -20.81, a maximum of -9.17 and a mean of -12.26. Fig. 4.6 (b) shows the resulting log-scaled spectrogram where the lower-frequency components are now more clearly distinguishable in the spectrogram image.

The spectrograms were then cropped to a specific lower frequency, capturing the primary harmonics of the ECG signal [**?**], the purpose of which is to reshape the spectrogram to a more balanced size, optimizing it for CNN processing. The adjusted dimensions of the spectrogram are better suited for convolutional and pooling layers, which perform better with more uniform aspect ratios.

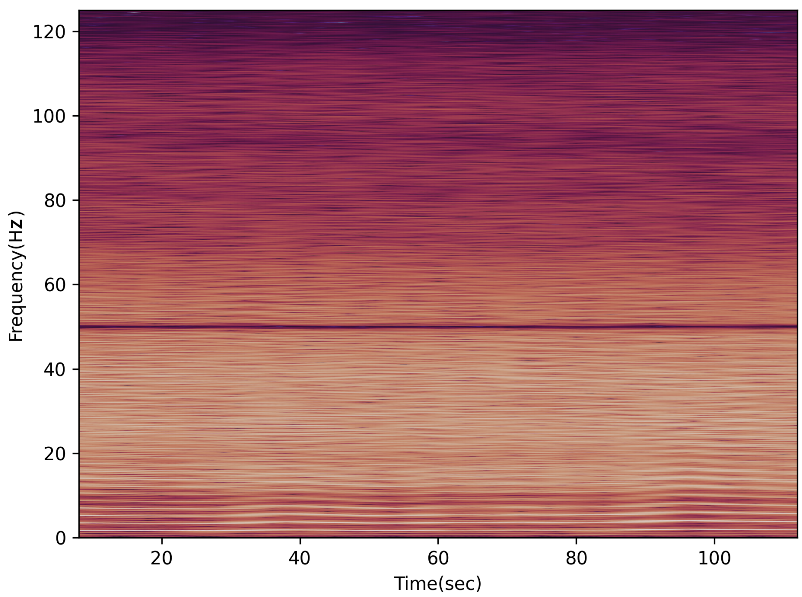
For HIE grading, the logarithm of power spectral density of the spectrogram is cropped until 9 Hz of frequency. Therefore, the spectrogram size changes from 3585×213 to a more balanced size of 253×213. The result is shown in 4.7 (a). For seizure detection, the logarithm of power spectral density of the spectrogram calculated and result is cropped until 7 Hz of frequency. Therefore, the spectrogram size changes from 2001×105 to a more balanced size of 113×105. The result is shown in 4.7 (b).

Various spectrogram configurations were evaluated with the proposed CNN model in this thesis, as summarized later in the results section in Table 4.3. Parameter selection prioritized generating square-shaped spectrograms to ensure balanced and reliable representations.

Fig. 4.8 shows a sample of cropped log-scaled spectrogram and NN interval series



(a)



(b)

Figure 4.6: Visualization spectrogram of 2-minute segments of ECG signal: (a) Spectrogram of a 2-minute ECG signal with a Hamming window of 16 Seconds and 1-second shift; (b) Logarithmically scaled spectrogram of a 2-minute ECG segment.

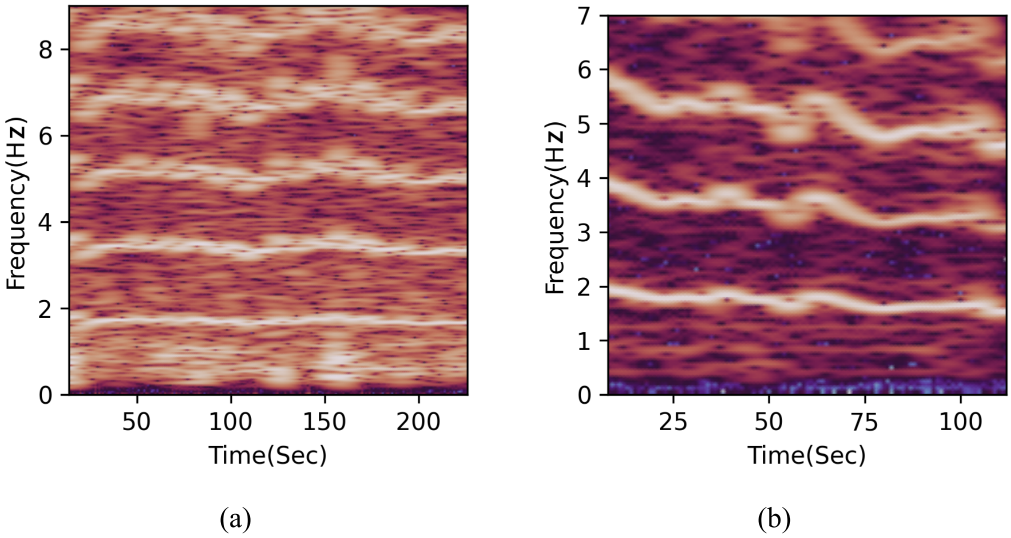


Figure 4.7: Visualization spectrogram of ECG signal: (a) log-scaled spectrogram of a 4-minute ECG segment, cropped to a maximum frequency of 9 Hz, with a size of 253×213. (b) log-scaled spectrogram of a 2-minute ECG segment, cropped to a maximum frequency of 7 Hz, with a size of 113×105.

for the same ECG segments. The low-frequency components in the spectrogram correspond to the slow variations in heart rate, also evident in the heart rate signal.

The generated spectrogram matrix is a 2D array where each element represents the log power spectral density of the ECG signal at a given frequency and time. The input to the proposed CNN in this chapter is the generated 2D spectrogram matrix, representing frequency content over time. CNNs are particularly well-suited to this type of input because they can learn spatial hierarchies of features by applying local filters across both time and frequency dimensions.

These matrices were standardized using the global mean and standard deviation of all matrices, in the training dataset to have zero mean and unit variance before being used as input for CNN classifier. The input tensor was constructed with shape of $1 \times H \times W$ where the first dimension represents the number of channels, whereas H and W denote the height (frequency axis) and width (time axis), re-

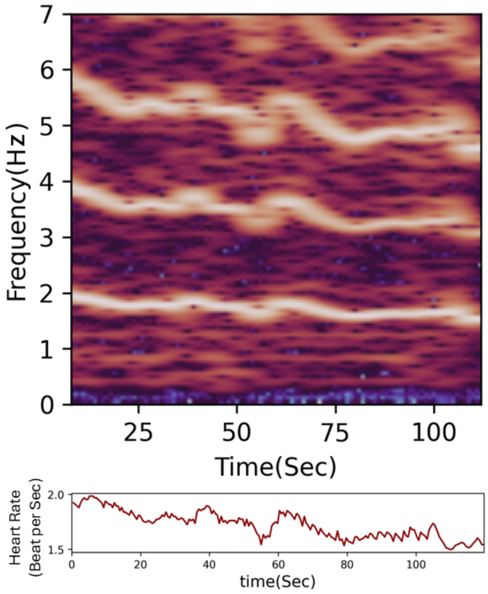


Figure 4.8: Visualization of ECG segment features: The top figure shows the log-scaled spectrogram and the bottom figure shows the heart rate trace.

spectively. Due to the simplicity of the input data, a lightweight yet effective CNN architecture was designed to balance performance and computational efficiency.

4.5.2 Proposed CNN architecture for HIE grading

The proposed CNN architecture for HIE grading consists of four sequential convolutional blocks, each followed by batch normalization, ReLU activation, and max pooling to progressively extract hierarchical features while reducing spatial dimensions. The network begins with a convolutional layer that applies 8 filters of size 7×7 to the input, followed by a second layer with 16 filters of size 5×5, a third with 32 filters of size 3×3, and a final convolutional layer with 64 filters, also of size 3×3. All convolutional layers use a stride of 1 and no padding. To maintain low computational cost, the network begins with a small number of filters (8 here), since the number of filters increases in deeper layers. As the network

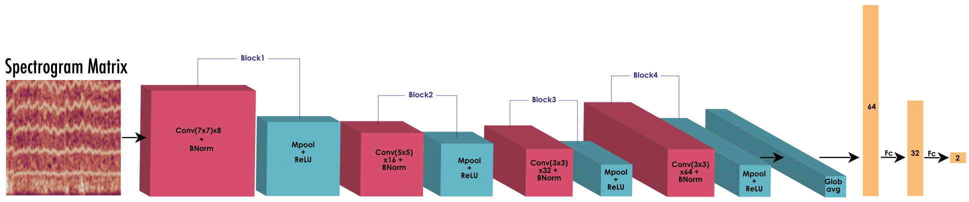


Figure 4.9: Architecture of the proposed CNN model for HIE grading. The input is a two-dimensional spectrogram matrix generated from a 4-minute ECG segment. The network comprises four convolutional blocks, each including a convolutional layer (Conv), batch normalization (BNorm), and max pooling (MPool). A global average pooling layer (Glob-Ave) is used to aggregate the spatial features, followed by a fully connected classification head consisting of two linear layers.

deepens, this design increases the receptive field. Max pooling with a 2×2 kernel size and batch normalization were applied after each block to downsample the feature maps.

After the final convolutional stage, a global average pooling layer reduces the spatial dimensions to 1×1, resulting in a 64-dimensional feature vector. This vector is flattened and passed through a fully connected layer with 32 units, followed by batch normalization, ReLU activation, and dropout with a probability of 0.5. The final classification is performed using a dense layer with 2 output units, corresponding to the target classes. An overview of the proposed CNN architecture is shown in Fig. 4.9. Table 4.1 provides a detailed summary of the changes in feature map dimensions and the number of channels at each stage of the network.

4.5.3 Proposed CNN architecture for seizure detection

Since seizure events are short, a smaller spectrogram size was selected for seizure detection than for HIE grading, requiring a tailored CNN architecture to match the input size. The proposed CNN architecture for seizure detection consists of three sequential convolutional blocks, each followed by batch normalization, ReLU activation, and max pooling to progressively extract hierarchical features while re-

Table 4.1: Feature map dimensions at each layer of the proposed CNN for HIE grading for an input spectrogram matrix of size $1 \times 253 \times 213$

Layer Name	**Output Shape (Channels $\times$ Height $\times$ Width)**
Conv1	$8 \times 247 \times 207$
MaxPool1	$8 \times 123 \times 103$
Conv2	$16 \times 119 \times 99$
MaxPool2	$16 \times 59 \times 49$
Conv3	$32 \times 57 \times 47$
MaxPool3	$32 \times 28 \times 23$
Conv4	$64 \times 26 \times 21$
MaxPool4	$64 \times 13 \times 10$
Global Avg Pool	$64 \times 1 \times 1$

ducing spatial dimensions. The network begins with a convolutional layer that applies 16 filters of size 5×5 to the input, followed by a second layer with 32 filters of size 3×3 and a final convolutional layer with 64 filters, also of size 3×3. All convolutional layers use a stride of 0 and no padding. To maintain low computational cost, the network begins with a small number of filters (16 here), since the number of filters increases in deeper layers. As the network deepens, this design increases the receptive field. Max pooling with a 2×2 kernel size and batch normalization were applied after each block to downsample the feature maps.

After the final convolutional stage, a global average pooling layer reduces the spatial dimensions to 1×1, resulting in a 64-dimensional feature vector. This vector is flattened and passed through a fully connected layer with 32 units, followed by batch normalization, ReLU activation, and dropout with a probability of 0.5. The final classification is performed using a dense layer with 2 output units, corresponding to the target classes. An overview of the proposed CNN architecture is shown in Fig. 4.10. Table 4.2 provides a detailed summary of the changes in feature map dimensions and the number of channels at each stage of the network.

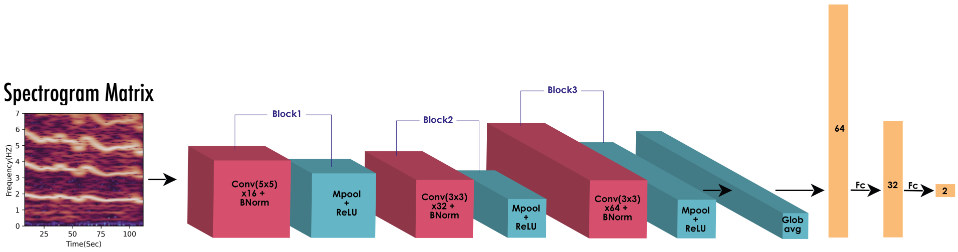


Figure 4.10: Architecture of the proposed CNN model for seizure detection. The input is a two-dimensional spectrogram matrix generated from a 2-minute ECG segment. The network comprises three convolutional blocks, each including a convolutional layer (Conv), batch normalization (BNorm), and max pooling (MPool). A global average pooling layer (Glob-Ave) is used to aggregate the spatial features, followed by a fully connected classification head consisting of two linear layers.

Table 4.2: Feature map dimensions at each layer of the proposed CNN for seizure detection for an input spectrogram matrix of size $1 \times 105 \times 113$

Layer Name	**Output Shape (Channels $\times$ Height $\times$ Width)**
Conv1	$16 \times 109 \times 101$
MaxPool1	$16 \times 54 \times 50$
Conv2	$32 \times 52 \times 48$
MaxPool2	$32 \times 26 \times 24$
Conv3	$64 \times 24 \times 22$
MaxPool3	$64 \times 12 \times 11$
Global Avg Pool	$64 \times 1 \times 1$

4.5.4 Performance evaluation

The classification performance of the model on the test dataset was evaluated using several standard metrics such as AUC, sensitivity, specificity and confusion matrix as used in chapter 3.

4.6 Application of the proposed methods on the ANSeR dataset

For the experiments in this chapter, the ANSeR datasets for HIE and seizure explained in chapter 2, section 2.9 were used. These datasets are large and diverse, including many infant recordings, making them diverse across individuals. All experiments were patient-independent, meaning the model was tested on infants it had never seen during training, helping it become more generalizable and robust.

In all models, the train dataset was randomly split into training (80%) and validation (20%) subsets for the purpose of loss optimization. Training was conducted with a batch size of 1024 for over 100 iterations, and was terminated early once the validation loss began to increase, indicating potential overfitting. During training, validation AUC was monitored at each iteration, and the model checkpoint corresponding to the highest observed validation AUC was selected as the final model to ensure optimal generalization performance and was then evaluated on the unseen test dataset. The models were initialized using the Kaiming method [**?**] to ensure efficient weight distribution at the start of training. Training was conducted using a cosine annealing learning rate scheduler with a warm up of 20 epochs. Optimization was performed using the AdamW optimizer, with cross-entropy loss employed as the objective function. Additionally, Weight decay was added to the optimizer to prevent overfitting by penalizing large weights and encouraging model generalization.

4.6.1 HIE grading

For the HIE grading task, the HIE ANSeR dataset explained in chapter 2, section 2.9 was used. CNN was trained on the ANSeR2 dataset (73 patients) and tested on the completely unseen ANSeR1 dataset (58 patients). A binary classification is implemented to determine whether the neonate needs to be cooled or not. Finally, the predicted probabilities of all segments in each 1-hour epoch in the test dataset were averaged to indicate the grade of that epoch. The initial learning rate for the scheduler was set to 1×10^{-5}, with a minimum learning rate of 1×10^{-7} and a weight decay of 1×10^{-7} for generalization.

4.6.2 Seizure detection

For seizure detection, the seizure ANSeR dataset explained in chapter 2, section 2.9 was used. CNN was trained on the ANSeR1 dataset (42 patients) and tested on the totally unseen ANSeR2 dataset (23 patients). The initial learning rate for the scheduler was set to 1×10^{-3}, with a minimum learning rate of 1×10^{-6} and a weight decay of 1×10^{-5}.

The training set contains a total of 267,266 samples, with 258,109 non-seizure samples and 9,157 seizure samples. This distribution exhibits a strong class imbalance, where seizure samples representing only about 3.4% of the data. To address this imbalance during training, class weights were initially calculated as follows:

$$W_{nonseizure} = \frac{267266}{2 \times 258109} = 0.5 \tag{4.5}$$

$$W_{seizure} = \frac{267266}{2 \times 9157} = 14.6 \tag{4.6}$$

Therefore, class weights of $[0.5, 14.6]$ were applied to the loss function during training.

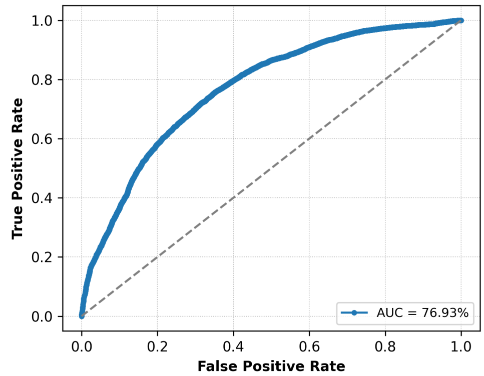
In addition, seizure samples were augmented by adding Gaussian noise with three different standard deviations: $[0.05, 0.1, 0.2]$. Gaussian noise was randomly applied, so applying it twice with the same standard deviation produced different augmented versions. In total, this resulted in six augmented samples for each original seizure sample. As a result, the total number of seizure samples increased to 64,099. Despite this augmentation, the dataset remained imbalanced; therefore, class weighting was still applied during training.

In the end, a 5-fold patient-independent cross-validation with an 80/20 split the same as chapter 3 was performed, with each fold containing 32 patients for training and 13 patients for testing. The goal was to utilizes the entire dataset and assess the method's performance. All folds were evaluated and compared to ensure a comprehensive and fair analysis. The patients in each fold follow the same grouping used in Chapter 3 for the seizure-detection cross-validation on the ANSeR dataset.

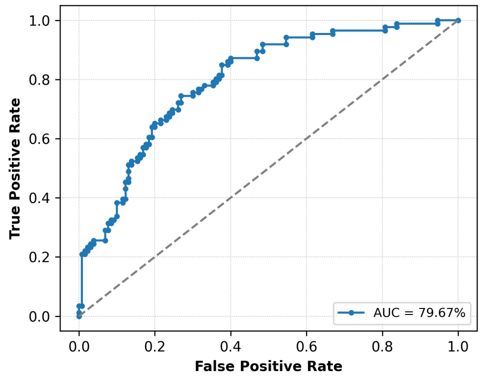
4.7 Results

4.7.1 HIE grading

An AUC of 79.67% is achieved for the HIE grading task. ROC curves and the corresponding AUC values for HIE grading using the CNN model were calculated based on all spectrogram segments, followed by averaging across each 1-hour ECG epoch. The results are shown in Fig. 4.11.



(a)



(b)

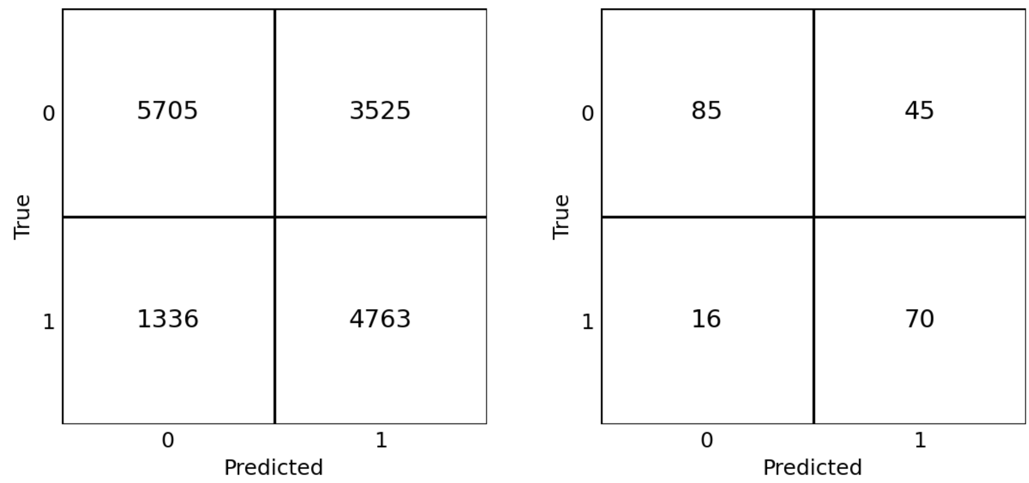
Figure 4.11: CNN-based HIE grading performance: (a) ROC curves and AUC values from all spectrogram segments; (b) AUC values averaged per patient over 1-hour ECG epochs.

Table 4.3: Comparison of CNN HIE grading AUC results for spectrograms with various parameters. Segment length and window size are given in seconds(s); image size is in pixels (P); maximum frequency (Freq) is in Hertz (Hz); and AUC values are reported as percentages.

Segment len(s)	Image size(P)	Window size(S)	Max Freq(Hz)	AUC(%)
240	253×213	28	9	79.67
240	55×59	6	9	74.7
300	251×251	50	5	78.81

Confusion matrices for CNN-based HIE grading were generated from all spectrogram segments and then aggregated across 1-hour ECG epochs to provide a representative evaluation. The final results are displayed in Fig. 4.12.

The performance of the proposed method was found to depend on spectrogram length, shape, and window size. To optimize results, various spectrogram configurations were evaluated with the CNN, as summarized in Table 4.3. Parameter selection prioritized generating square-shaped spectrograms to ensure balanced and reliable representations. The best outcome was obtained with a 4-minute spectrogram using a 28-second window size.



(a) (b)

Figure 4.12: CNN-based HIE grading performance: (a) confusion matrix computed from all spectrogram segments; (b) patient-level confusion matrix based on 1-hour ECG epoch averages.

4.7.2 Seizure detection

Seizure detection results on the ANSeR2 dataset using CNN models with different strategies for handling class imbalance are presented in Table 4.4. Without augmentation, and using class weights of $[0.5, 14.6]$, the model achieved an AUC of 61.36%. When Gaussian noise augmentation was applied, the model detected more seizure events, showing an improvement in sensitivity. The AUC remained nearly unchanged (61.17%), suggesting that augmentation altered the trade-off between sensitivity and specificity rather than improving the overall discriminative capability of the model. Nevertheless, the overall performance for seizure detection remains relatively weak, highlighting the challenge posed by the extreme class imbalance in this dataset.

The 5-fold cross validation results for each fold along with the number and length of seizure events in that fold are presented in Table 4.5. The patients in each

Table 4.4: Seizure detection performance on the ANSeR2 dataset using the proposed CNN with augmentation and class weight.

Augmentation	Class weight	TP	FP	TN	FN	AUC (%)
No	[0.5,14.6]	635	7540	93975	3697	61.36
Gaussian noise	[0.5,14.6]	827	11541	89974	3505	61.17

Table 4.5: Five-fold cross-validation results of CNN with ECG spectrograms for seizure detection, along with their average performance, reported in terms of Sensitivity, Specificity and AUC. The seizure event distribution of each fold is also presented.

Fold	Performance Metrics				Seizure Event Analysis	
	Sens (%)	Spec (%)	Acc (%)	AUC (%)	average of max seizure event length per patient (min)	number of events longer than 10 mins
0	30.59	81.23	80.58	61.96	11.26	7
1	30.12	85.97	83.82	**63.63**	**25.6**	23
2	17.54	86.52	84.34	55.46	10.12	13
3	26.06	86.06	84.16	62.82	18.32	24
4	12	91.94	88.49	53.66	14.86	52
Avg	23.26	86.34	84.28	59.51	16.032	23.8

fold follow the same grouping used in Chapter 3 for the seizure-detection cross-validation on the ANSeR dataset. As shown, the model achieved consistently high specificity across most folds, indicating strong performance in identifying non-seizure cases. However, sensitivity remained relatively low, reflecting a limitation in detecting actual seizures. Additionally, similar patterns as those observed in chapter 3 can be seen. Fold 1 achieved the highest AUC, which had the longest seizure events on average and folds 3 and 5 have the lowest AUC, indicating the model weak performance with more short seizure events in the dataset.

4.8 Conclusion

The CNN model trained on spectrogram representations achieved an AUC of 79.67% for the HIE grading task. These results indicate that ECG spectrograms combined with deep learning models have potential for HIE grading, showing slightly better performance than traditional HRV-based methods (AUC = 79.34%). Importantly, this approach eliminates the need for NN interval extraction from ECG signals, which is typically performed using the Pan-Tompkins method, a process that is time-consuming, prone to errors, and requires extensive data cleaning. Moreover, representing ECG signals as a spectrogram enables the use of state-of-the-art, robust deep learning architectures, further enhancing model flexibility and performance.

Seizure detection results indicate generally weak performance; however, the achieved AUC of 61.36% is slightly higher than traditional HRV-based methods discussed in chapter 3 (AUC = 60.04%), suggesting a modest improvement. Data augmentation also contributed by helping to balance the dataset and producing a small performance gain. Also, the 5-fold cross-validation results shows considerable variability across datasets, underscoring uncertainty in the model's performance. The relatively weak overall performance can likely be attributed to several factors discussed in Chapter 3, including severe class imbalance that could not be fully mitigated by the applied methods, the short duration of seizure events, potential misalignment of labels, high variability between seizure and non-seizure spectrogram representations, and the presence of multiple seizure types.

5 | HIE Grading and Seizure Detection Using Raw ECG

5.1 Introduction and literature review

In this chapter, a method is proposed that directly utilizes raw ECG signals. This approach is motivated by the inherently noisy nature of ECG recordings. Additionally, using raw EEG signal had highly promising results for neonatal seizure detection (AUC of 98.5%) in a recent study by O'Shea [**?**]. HRV feature-based methods depend on accurate R-peak detection; however, in the presence of noise, R-peaks detection is less accurate [**?**] [**?**], leading to incomplete or invalid extracted features. Similarly, time–frequency representations such as spectrograms require a minimum level of signal consistency; severely corrupted segments often result in non-informative representations.

Using the the raw ECG signal preserves the temporal information of the data while eliminating preprocessing steps such as heart-rate calculation and spectrogram generation. As a result, the proposed pipeline avoids the potential failure modes associated with feature extraction and maintains robustness across noisy segments. Moreover, operating directly on the raw signal reduces preprocessing complexity and prevents issues with missing features, making the approach more suitable for real-time applications. Additionally, the reduced computational load enables more timely output generation, which enhances the practicality of the model for clinical deployment.

Based on the existing literature, no prior study has utilized raw ECG signals for grading HIE and detecting seizures. Existing approaches all rely on HRV-based methods or the EEG signal as summarized in Chapter 2, Section 2.7. To effectively model these raw sequences, transformer based architectures were leveraged, which are well-suited for capturing long-range dependencies in time-series data. This strength is especially useful for classifying the severity of HIE from ECG signals, where recognizing subtle changes requires understanding both wider temporal context and the timing between heartbeats. Transformer-based models have shown promising results in the ECG domain such as arrhythmia classification [**?**] [**?**], detection of obstructive sleep apnea [**?**] and stress detection [**?**].

In this chapter, the Transformer architecture, introduced by Vaswani et al. [**?**], is first described. The proposed transformer-based architectures for HIE grading and seizure detection are then presented. Each task is addressed separately; although the same Transformer-based architecture is employed for both HIE grading and seizure detection, the input dimensions differ for each task. The models are subsequently trained and experimentally evaluated on two distinct ANSeR datasets—the HIE ANSeR dataset and the seizure ANSeR dataset—introduced in Chapter 2, Section 2.9. Additionally, a Vision Transformer (ViT)-based model is applied exclusively to the seizure detection task for comparison purposes.

Finally, all methods presented in Chapters 3, 4, and this chapter are systematically compared for each task, with HIE grading and seizure detection evaluated separately in terms of performance and practical applicability.

5.2 Transformer architecture

The Transformer architecture, represents a paradigm shift in sequence modeling. Unlike RNNs or CNNs, which either process input sequentially or within local receptive fields, Transformers enable global information flow across the entire sequence at every layer. This is achieved via a self-attention mechanism that allows the model to dynamically attend to all parts of the input, regardless of their relative position. Such capacity is particularly beneficial in time series analysis, where capturing long-range dependencies and nonlinear temporal interactions is essential.

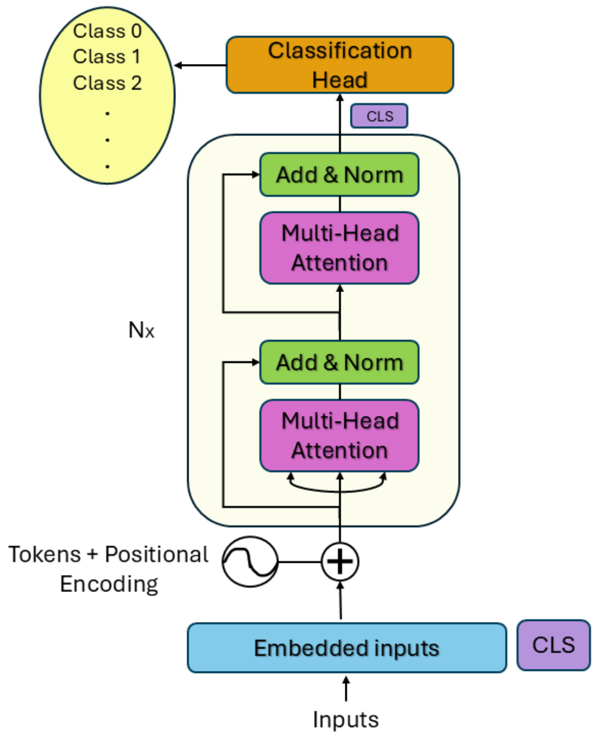


Figure 5.1: Architecture of the transformer encoder. Adapted from [?].

The transformer architecture introduced in [?] consists of two main components: an encoder and a decoder. The decoder component is typically used in sequence generation tasks such as machine translation. In classification tasks, encoder-only transformer architectures are commonly employed; this thesis also utilizes such a model. This setup discards the decoder component, reducing architectural complexity and computational overhead while retaining the model's ability to learn expressive temporal representations. A linear layer or MLP is then applied on top of the encoder output to perform the final classification. The Encoder architecture for classification is presented in Fig. 5.1

5.2.1 Input embedding

Before feeding raw data into a Transformer encoder, it must first be divided into smaller units, called tokens in natural language and time-series contexts, and patches in image processing. These tokens or patches are then projected into a high-dimensional vector space to serve as the input sequence for the model. The input sequence with dimension d_{in} and length L, denoted as $X = (x_1, ..., x_n)$, is first passed through a linear projection layer that maps the raw signal into the model's embedding dimension d_{model}:

$$X_{\text{embed}} = XW_{\text{proj}} + B_{\text{proj}}, \quad X \in \mathbb{R}^{L \times d_{\text{in}}}, \quad X_{\text{embed}} \in \mathbb{R}^{L \times d_{\text{model}}} \tag{5.1}$$

Where W_{proj} and B_{proj} are the projection weights and bias, respectively. The resulting matrix X_{embed} forms the embedded representation of the raw ECG sequence.

5.2.2 Positional encoding

Embedding transforms raw ECG sequential data into fixed-length numerical vectors (tokens) that capture temporal patterns. To preserve temporal order, since the Transformer has no inherent notion of sequence, a positional encoding vector (PE) is added to each input embedding vector:

$$Y = X_{\text{embed}} + \text{PE}, \quad \text{PE} \in R^{L \times d_{\text{model}}} \tag{5.2}$$

The resulting representation Y is then passed to the first encoder layer.

In the original transformer paper [**?**], positional information is injected into the input embeddings using sinusoidal positional encodings:

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{5.3}$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{5.4}$$

This encoding allow the model to reason about the relative and absolute positions of tokens in the input sequence.

5.2.3 Encoder layer

The encoder is the core component of the Transformer model responsible for learning contextual representations from the input sequence. As depicted in the Fig. 5.1, the transformer encoder consists of a stack of N identical layers, each designed to model dependencies between all positions in the sequence, regardless of their distance. Each encoder layer is composed of two main sublayers:

- **Multi-head attention mechanism and add and norm**

  The Scaled Dot-Product Attention mechanism forms the computational heart of the Transformer. The input consists of queries and keys of dimension d_k, and values of dimension d_v. For a given input token sequence Y, three learned linear projections are applied to obtain the query (Q), key (K), and value (V) matrices:

$$Q = YW^Q, \quad K = YW^K, \quad V = YW^V \tag{5.5}$$

  The output of scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.6}$$

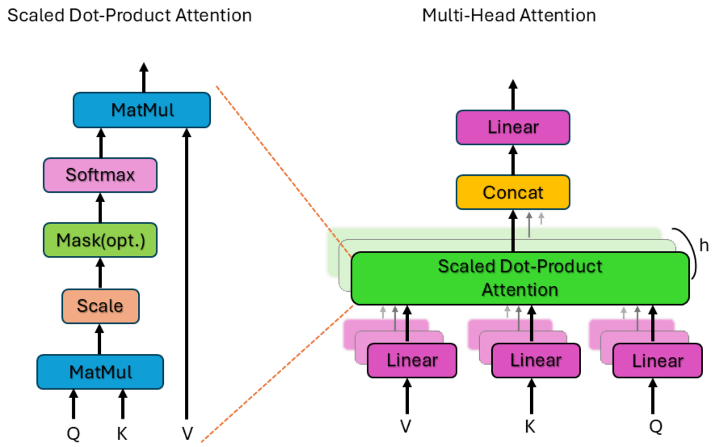


Figure 5.2: (left) Scaled Dot-Product Attention and Multi-Head Attention (right) consists of several attention layers running in parallel. Adapted from [**?**].

Instead of performing a single attention operation, the Transformer uses h parallel attention heads to allow the model to jointly attend to information from different representation subspaces:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad i = 1, \ldots, h \tag{5.7}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5.8}$$

This mechanism enriches the model's ability to capture diverse types of temporal relationships. The architecture of the Scaled Dot-Product attention mechanism and Multi-head attention mechanism are shown in Fig. 5.2.

The multi-head attention layer is wrapped with a residual (skip) connection which adds the input of each the layer to its output, preserving the original information and mitigating vanishing gradients. Then layer normalization is applied to help stabilize the hidden states across training iterations.

- **Position-wise feed-forward network and add and norm**
  Each position in the sequence is then passed through a fully connected feed-forward network. This component consists of two linear transformations with a non-linear activation function (usually ReLU) applied to each position separately and identically. It allows the model to transform the attended features into a richer representation space.

$$\text{FFN}(y) = \max(0, yW_1 + b_1)W_2 + b_2 \tag{5.9}$$

  To stabilize training and improve gradient flow, the feed-forward layer is also wrapped in a residual connection and followed by layer normalization, similar to the multi-head attention layer.

5.2.4 Classification head

For classification, a single learnable vector ([CLS]) is prepended to the sequence of embedded tokens, as shown in Fig. 5.1. This vector aggregates information from all tokens throughout the encoder, and its final representation is passed to a linear or MLP head to produce the classification output.

5.3 Proposed method using raw ECG signal and transformer-based models

Transformers are particularly effective at capturing long-range dependencies in data through their self-attention mechanism. This strength is especially useful for classifying the severity of HIE from ECG signals, where recognizing subtle changes requires understanding both the temporal context and the timing between heartbeats. In contrast to CNNs or traditional approaches that depend on predefined representations, such as spectrograms or statistical features, transformers can operate directly on the raw ECG signal. This enables the model to automatically learn relevant patterns without relying on hand-crafted features. For both HIE grading and seizure detection, TMFormer (Trend-Augmented Multiscale Tokenization Transformer) was proposed. TMFormer is a transformer-based model that introduces a tokenization approach derived from the ECG signal attributes, capturing spatial patterns at multiple scales and tracking the signal's long-term progression.

5.3.1 ECG preprocessing

First the ECG recordings were filtered to remove common noise artifacts for higher signal quality and computational efficiency. A high-pass Butterworth filter with a 0.5 Hz cutoff was applied to eliminate baseline wander. To reduce computational load and improve training efficiency, the ECG signal was downsampled from 256 Hz to 80 Hz. This significantly lowers the size of the input sequence, allowing for faster processing, larger batch sizes, and deeper models. Since most important ECG information falls below 40 Hz as shown in the ECG specrograms in Chapter 4, Fig. 4.6, this reduction preserves the essential information while eliminating unnecessary high-frequency noise. The downsampled ECG signal was divided into four-minutes segments for HIE grading and two-minutes segments for seizure detection with 80% overlap, consistent with the approach used in Chapter 3 and 4. Similar segment lengths were used because they capture sufficient abnormalities and changes in the signal, and they provide an adequate amount of data for analysis. For seizure detection, this segment length also offers a suitable window to include enough seizure events. In addition, using comparable segment lengths allows the results in this chapter to be directly compared with those presented in

Chapters 3 and 4. Each segment was standardized using the global mean and standard deviation to achieve zero mean and unit variance before being input to the model.

5.3.2 Overall architecture of the proposed transformer-based model

The overall design of the proposed model consists of three main components, a tokenizer for signal embedding, two transformer encoder blocks and a classification head, as shown in Fig. 5.3. The tokenizer handles the ECG signal directly, without manual feature extraction, enabling the model to automatically learn and detect relevant patterns. The transformer encoder blocks then refine these learned representations, capturing complex temporal and frequency-based relationships in the ECG signal to improve contextual understanding. Finally, the classification head examines the enhanced representations to determine the grade of HIE.

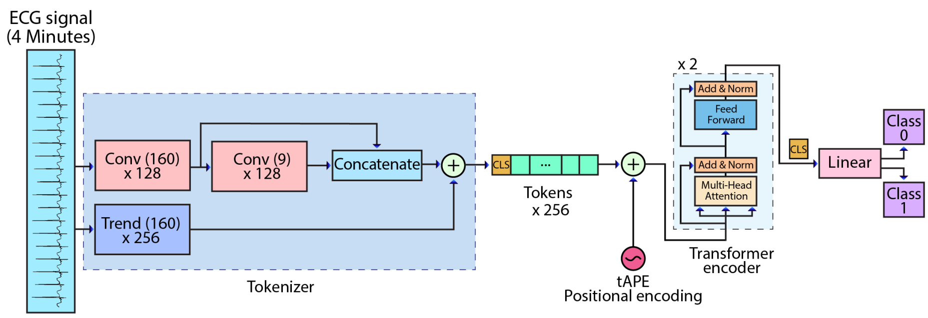


Figure 5.3: The proposed TMFormer model for HIE grading and seizure detection. HIE grading input is shown in this architecture (a 4-minute raw ECG segment); however, for seizure detection the input will be a 2-minute segment. The framework consist of a Trend-Augmented Multi-scale Tokenization, two transformer encoder blocks and a classification head.

5.3.3 Trend-Augmented Multi-scale tokenization

In several studies, CNNs or other feature-extraction methods are applied to the input before feeding it into the Transformer. These approaches provide more informative representations of the raw data and reduce the input sequence length, thereby lowering the computational cost of the Transformer while improving its performance. In this thesis, a novel Tokenization or input embedding technique specialized for ECG is proposed. The first stage of the designed tokenization has a multi-scale architecture consisting of 2 layers. The first layer tokenizes the ECG signal into large segments with minimal redundancy. A large kernel is applied to capture global patterns, such as overall waveform morphology and long-term dependencies. Therefore, a convolutional layer with a large kernel of size 160 is used to extract tokens. Size 160 is chosen since it is equal to 2 seconds of ECG which normally contains four to five consecutive P-QRS-T complexes in neonatal ECG. This is consistent with the neonatal NN interval of approximately 0.32–0.86 seconds (heart rate 70–190 bpm). A small overlap of 11 is used to ensure minimal redundancy while preserving continuity between segments. The one dimension ECG data is tokenized into 128 tokens and each is then embedded into a 128-dimensional vector space. This transformation enhances the representation, allowing the model to learn more intricate and meaningful features. The second layer uses a smaller kernel of size 9 on each token, extracting finer details from them. This enables the model to capture local features, such as peaks, slopes, and subtle variations in the signal. In addition, the output of the second layer preserves both the number of tokens and the embedding dimension at 128, consistent with the first convolutional layer's output. Ultimately, the tokens extracted by both kernels are concatenated channel-wise, maintaining the number of tokens at 128 while increasing the embedding dimension to 256. This allows the model to learn both contextual and detailed features simultaneously. Additionally, the computational load is lower compared to processing the entire ECG with small kernels.

Time series decomposition involves separating a time series into several components, with each capturing a unique pattern or fundamental structure [**?**]. Among these components, the trend is particularly important as it reflects the series' long-term direction or movement. Since the trend component reflects changes in the signal's pattern over extended periods, and ECG pattern alterations are of interest in the context of HIE, analyzing the trend component of the ECG signal can be effective for detecting these changes. In this chapter, a refined approach utilizing the ECG trend component in the second stage of the tokenization is proposed. This was achieved using an average pooling layer to extract the trend, which is then incorporated into the model's convolution output in a residual manner. This, enhances the model's ability to capture long-term variations in ECG signals, providing a more comprehensive representation of HIE severity. Here, the trend can be calculated as (5.10):

$$T_j = \frac{1}{k} \sum_{i=0}^{k-1} x_{s \times j + i} \tag{5.10}$$

The estimate of the trend T at position j, is obtained by averaging values of the ECG signal x, within k samples of i. x starts at index $s \times j$, where s is the stride. Here, k is set to 160, to match the kernel size of the first convolutional layer. A stride of 149 corresponding to an overlap of 11 samples is used, consistent with the overlap in the first layer. This ensures that the output contains exactly 128 tokens, allowing for proper alignment during residual addition. Finally, the residual trend calculation is performed across all 256 tokens dimensions.

Time Absolute Position Encoding (tAPE) [**?**] improves upon traditional sinusoidal absolute position encoding, which was originally designed for language models that typically use high-dimensional embeddings. Unlike the conventional method, tAPE incorporates the sequence length into the frequency components of the sine and cosine functions which was explained in equation (5.4), resulting in a smoother and more consistent dot product between positional embeddings regardless of the series length. By scaling the frequency according to the embedding dimension, tAPE preserves distance awareness even in lower-dimensional embeddings, which are common in time-series applications. The original transformer frequency term $\omega_k^{original}$ and the modified tAPE version ω_k^{tAPE} are shown below.

$$\omega_k^{original} = 10000^{-2k/d_{model}} \tag{5.11}$$

$$\omega_k^{tAPE} = \frac{\omega_k^{original} \times d_{model}}{L} \tag{5.12}$$

Where L is the sequence length and d_{model} is the embedding dimension.

Integrating the proposed tokenizer as the first component of this thesis architecture not only enables the capture of both local and global patterns within the raw ECG signal but also, reduces the sequence length, optimizing the data for tAPE position encoding while lowering computational complexity. Hence, tAPE positional encoding is added to the extracted tokens by the proposed tokenizer, to pass to the encoder layer. Further, a dropout rate of 0.3 is used for regularization and avoiding the model's dependence on particular feature embeddings and encouraging a broader and better understanding of the whole ECG segment.

5.3.4 Transformer encoder layers

The model consists of two Transformer encoder layers, each including layer normalization, multi-head self-attention, and a feedforward neural network (MLP). A classification token (CLS token) is added at the beginning of the input sequence to act as a summary representation for classification purposes. Eight heads were chosen for the attention mechanism to keep a balanced structure which are large enough to capture important features, yet small enough to keep computation reasonable and avoid overfitting. A small dropout rate of 0.2 is applied to support regularization. The MLP expands the embedding from 256 to 512 to enrich feature representation while keeping the model shallow enough to avoid unnecessary complexity or overfitting given the data size and complexity of ECG segments. A GELU activation function [**?**] is applied between these layers to introduce non-linearity.

5.3.5 Classification head

Once the Transformer encoders have processed the input sequence, including the CLS token, the output associated with the CLS token is used for classification. This output is then fed into a feedforward network consisting of layer normalization followed by a linear layer to generate the final class prediction.

5.3.6 Explainability with attention rollout

Attention rollout [**?**] combines attention weights from all transformer layers by multiplication to trace the flow of information from the input tokens to the CLS token. This method reveals how information from each token propagates through the network and identifies which parts of the input the model focuses on at each layer. The result is a cumulative attention vector that represents the overall impact of each token on the final prediction.

Here, the goal is to identify the portions of the ECG signal that contain the most relevant physiological features, which the model relies on to make its final decision. To execute this, both the attention flow at each layer and the cumulative attention flow across all layers (attention rollout) are calculated. Since heart rate, which is derived from NN intervals, fluctuates during HIE, attention flow is compared with the corresponding NN intervals. This comparison explores how variations in NN intervals correspond to the regions of the ECG signal that the model attends to across layers.

5.4 Seizure detection using CNN-ViT and ECG spectrogram (comparative approach)

For the purpose of comparison, a vision transformer [**?**] was used in this section as an alternative to proposed TMFormer for seizure detection task. The vision transformer model was applied on the spectrogram images to utilize the potentially more informative features for seizure detection.

Therefore, a Vision Transformer with Convolutional Patch Embedding (CNN–ViT) was designed and applied to ECG spectrograms to explore whether the strengths of transformer-based architectures could be effectively utilized for seizure detection.

In the CNN–ViT architecture, the spectrogram matrices generated in Chapter 4 were utilized as inputs. In the designed CNN-ViT, the spectrogram matrix is first divided into patches using a convolutional layer with a kernel size of 7, producing a total of 240 patches. Each patch is projected into a 64-dimensional embedding space. This configuration balances capturing local spectral patterns, providing

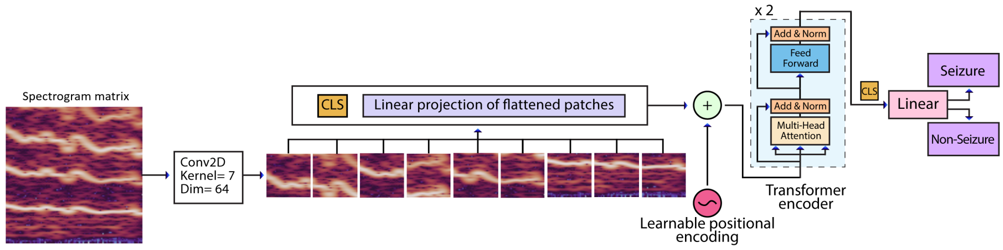


Figure 5.4: Architecture of the proposed CNN-ViT model used for seziure detection. The input is a 2-dimensional spectrogram matrix generated from a 2-minute ECG segment.

enough tokens for the transformer to learn relationships, and keeping computational cost manageable. To retain positional information, learnable position encodings are added to the patch embeddings. The Transformer backbone consists of two encoder layers, each composed of layer normalization, a multi-head self-attention mechanism, and a feedforward network (multi-layer perceptron, MLP). A learnable classification token (CLS) is prepended to the sequence and serves as the representative feature for the final prediction. The self-attention module operates with 8 heads and employs a dropout rate of 0.2 for regularization. The MLP expands the embedding dimension from 64 to 128 through two linear transformations, with a GELU activation applied between them to introduce non-linearity. After processing by the Transformer encoders, the output corresponding to the CLS token is extracted and passed through a lightweight classifier. This classifier consists of layer normalization followed by a linear layer, which maps the representation to the final class scores. The flowchart of this model is presented in Fig. 5.4.

5.5 Application of the proposed methods on the ANSeR dataset

For all models, the training data was randomly split into two subsets: 80% for training and 20% for validation. Training was performed using a batch size of 1024 over more than 150 iterations, with early stopping applied when the validation loss began to increase, suggesting potential overfitting. The model that achieved the highest AUC on the validation set was selected for evaluation on the independent test dataset. All models were initialized using the Kaiming method to ensure effective weight distribution at the start of training. A cosine annealing learning rate scheduler with a 50-epoch warm-up phase was used to manage the learning rate throughout training. Optimization was carried out using the AdamW optimizer, with cross-entropy loss serving as the objective function. Further, Weight decay was added to the optimizer to prevent overfitting by penalizing large weights and encouraging model generalization.

5.5.1 HIE grading

For the HIE grading task, HIE ANSeR dataset explain in chapter 2, section 2.9 was used. The TMFormer model was trained on the ANSeR2 dataset and evaluated on the completely unseen ANSeR1 dataset. As before, a binary classification approach was used to determine whether a neonate required cooling (category 1) or did not required cooling (category 0). To assign a coolling class to each 1-hour epoch in the test set, predicted probabilities from all spectrogram segments within the epoch were averaged. The learning rate scheduler was initialized with a starting learning rate of 1×10^{-5}, a minimum learning rate of 1×10^{-6}, and a weight decay of 1×10^{-5} to encourage generalization.

Additionally, TMFormer was compared to two baseline transformer models: one using a simple tokenization method (with a kernel size of 160 and an equal number of parameters), and another with the same architecture as TMFormer but without incorporating the trend component during tokenization. These comparisons were designed to assess the effectiveness of TMFormer's tokenization strategy and the specific contribution of trend information to its overall performance.

5.5.2 Seizure detection

For seizure detection, seizure ANSeR dataset explain in chapter 2, section 2.9 was used. Both TMFormer and CNN-ViT models were trained on ANSeR1 dataset and tested on totally unseen ANSeR2 dataset. The initial learning rate for the scheduler was set to 1×10^{-3}, with a minimum learning rate of 1×10^{-6} and a weight decay of 1×10^{-5}.

Since the seizure segments were highly imbalanced compared to the non-seizure segments, the same strategies used in Chapter 4 were applied to address this issue. Specifically, a class weight of $[0.5, 14.6]$ was assigned in the loss function, and Gaussian noise was added to the seizure segments for data augmentation.

In the end, a 5-fold patient-independent cross-validation with an 80/20 split, the same as chapters 3 and 4, was performed. All folds were evaluated and compared to ensure a comprehensive and fair analysis.

Table 5.1: HIE grading results of applied models based on all 4- minute ECG segments and averaging across each 1-hour ECG epoch.

Model	AUC(%) (1-hour ECG epoch)	AUC(%) (All 4-minute segments)
TMFormer	80.21	76.50
TMFormer without trend	79.02	75.73
Transformer	78.6	75.50

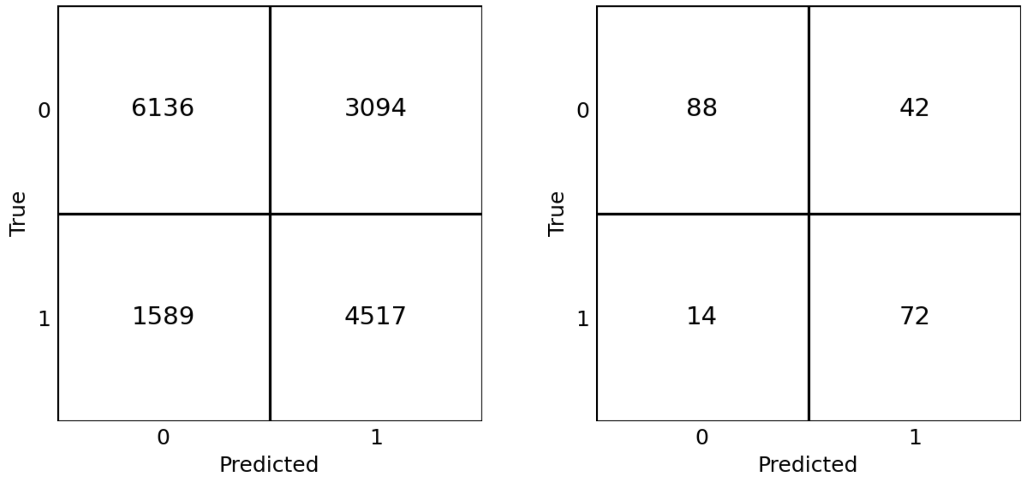
5.6 Results

5.6.1 HIE grading

An AUC of 80.21% is achieved for the HIE grading task using the proposed TMFormer. AUC values for HIE grading using various applied transformer-based models were calculated based on all ECG segments, followed by averaging across each 1-hour ECG epoch. The results of the models evaluated on all 4-minute segments, as well as on the labels for each 1-hour ECG epoch, are presented in Table 5.1.

Confusion matrices for HIE grading using TMFormer were generated from all ECG segments and then aggregated across 1-hour ECG epochs to provide a representative evaluation. The final results are displayed in Fig. 5.5.

Fig. 5.6 presents a 4 minute input ECG signal, overlaid with the attention flow from the CLS token at the first layer, the second layer, and the cumulative attention flow across both layers, along with the extracted NN intervals derived from the ECG signal. Each layer's attention flow is visualized as a heatmap, where darker pink regions indicate areas that the model focuses on more strongly during classification, based on the attention weights of that specific layer. The cumulative attention flow (attention rollout) is shown as the combined (multiplied) attention from both layers. The extracted NN intervals, calculated from the peaks of the



(a) (b)

Figure 5.5: TMFormer HIE grading performance: (a) confusion matrix computed from all 4-minute ECG segments; (b) patient-level confusion matrix based on 1-hour ECG epoch averages.

ECG signal, are included for comparison. The plots show that the model's attention consistently emphasizes regions of the ECG signal that align with fluctuations in the NN interval pattern, suggesting a connection between the model's focus and heart rate dynamics, which are often altered in cases of HIE.

5.6.2 Seizure detection

The result of the proposed TMFormer for seizure detection was not satisfying. Since the results using HRV features and ECG spectrograms were better than the result using raw ECG, the idea was to apply a vision transformer to the spectrogram images to leverage the potentially more informative features for seizure detection.

As presented in Table 5.2, TMFormer using raw ECG achieved an AUC of 59.2%, while the CNN–ViT utilizing ECG spectrograms reached a higher performance with an AUC of 61.57%.

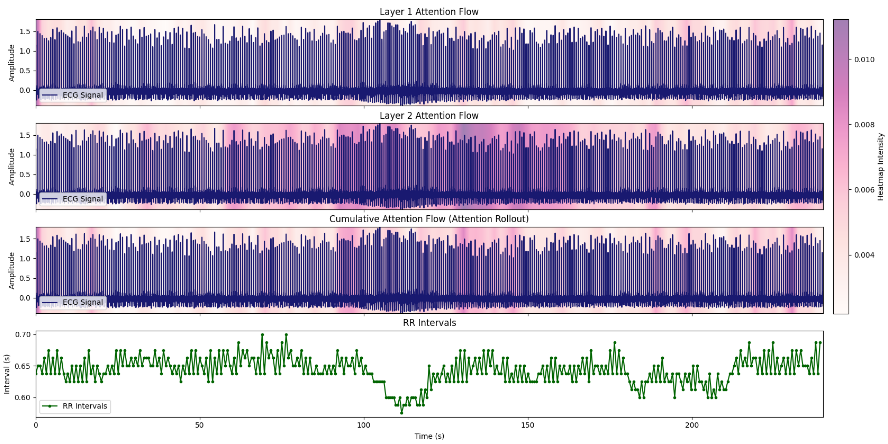


Figure 5.6: A 4-minute ECG segment with overlaid TMFormer attention flow heatmaps from the CLS token at the first and second layer and the cumulative attention flow across both layers, alongside NN intervals for comparison. The darker pink regions highlight areas of the ECG signal that the model attends to more strongly during classification. The ECG recording is from hour 6 of a neonate with with target category 1 (TH is required).

Table 5.2: Seizure detection result of applied models on test set.

Model	Input	TP	FP	TN	FN	AUC (%)
TMFormer	Raw ECG	1087	15400	86264	3249	59.2
CNN-ViT	ECG spectrogram	2690	48086	53429	1642	61.57

Table 5.3: Five-fold cross-validation results of TMFormer with raw ECG signal for seizure detection, along with their average performance, reported in terms of Sensitivity, Specificity and AUC. Seizure event distribution of each fold is also presented.

Fold	Sens (%)	Spec (%)	AUC (%)	average of max seizure event length per patient (min)	number of events longer than 10 mins
0	19.47	84.94	55.78	11.26	7
1	16.63	88.02	**55.25**	**25.6**	23
2	16.82	87.46	57.054	10.12	13
3	19.14	90.27	57.943	18.32	24
4	13.2	87.95	52.088	14.86	52
Avg	17.05	87.73	55.62	16.032	23.8

The 5-fold cross validation results of both methods for each fold along with the number and length of seizure events in that fold are presented in Table 5.3 and 5.4. As shown, the TMFormer with raw ECG input, achieved consistently high specificity across most folds, indicating strong performance in identifying non-seizure cases. However, their sensitivity remained very low, reflecting a limitation in detecting seizures. Additionally, the patterns observed here differ from those in Chapters 3 and 4. This is not surprising, as the results indicate that the model lacks an understanding of the seizure segments; performance below 60% are random. However, the average AUC across all folds for the CNN-ViT model is higher than the average AUC of the TMFormer model. This reflects the strength of the CNN-ViT architecture, where CNN layers extract local time-frequency patterns while the Transformer component captures global dependencies across the ECG spectrogram.

Table 5.4: Five-fold cross-validation results of CNN-ViT with ECG spectrograms for seizure detection, along with their average performance, reported in terms of Sensitivity, Specificity and AUC.

Fold	Sens (%)	Spec (%)	AUC (%)	average of max seizure event length per patient (min)	number of events longer than 10 mins
0	76.33	49.67	67.05	11.26	7
1	77.15	40.1	**65.45**	**25.6**	23
2	58.15	53.85	58.61	10.12	13
3	55.69	59.61	60.08	18.32	24
4	48.62	55.8	53.21	14.86	52
Avg	63.19	51.81	60.88	16.03	23.8

5.7 Methods comparison

5.7.1 HIE grading

One of the aims of this study was to develop an automatic, real-time system for clinical support in hospitals to assist clinicians in grading HIE regarding applying TH. Therefore, identifying the most effective solution was a central objective.

Here, the results of all HIE grading approaches applied to the ANSeR dataset, described in detail in Chapters 3, 4, and 5, are systematically compared. The comparison focuses on classification accuracy and the amount of data preparation required for each model. Additionally, the outputs are evaluated against expert-provided clinical labels. To illustrate practical applicability, the predictions from each model are plotted for a continuous ECG recording, providing a direct visual comparison of method performance in a real-world scenario.

Table 5.5 presents the AUC score of all models.

If comparing the results of this thesis with AUC of 81% reported by Ahmed et al. [**?**], the AUCs are lower. However, their study was conducted on a much smaller

Table 5.5: AUC values for HIE grading regarding requirement of TH using all applied models and transformer variations used in this paper.

Model	Input data	AUC score (%)
XGBoost	HRV features	79.26
MLP	HRV features	79.34
CNN	ECG Spectrogram	79.67
Transformer	Raw ECG	78.6
TMFormer	Raw ECG	**80.21**

Table 5.6: Percentage-based True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) metrics for HIE grade predictions regarding requirement of TH on all 4-minute segments across all applied models.

model	TP (%)	FP (%)	TN (%)	FN (%)
XGBoost	35.4	22.4	44.7	8.6
MLP	34.5	25.1	41.7	9.4
CNN	34.3	25.4	41.2	9.7
Transformer	35.1	26.6	38.2	7.8
TMFormer	32.6	22.3	44.2	11.5

dataset (54 patients compared to 120 in this thesis) and relied on leave-one-out cross-validation, which is known to produce optimistically biased estimates.

Since the classification results for all applied models were based on the average predicted probability across all test samples within each one-hour ECG epoch, the corresponding True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) metrics, calculated from predictions on all 4-minute segments, are provided in Table 5.6.

The execution time required for input data generation across all methods is summarized in Table 5.7. The measurement was performed for each 1-hour ECG epoch and the per-segment execution time was calculated based on the number of 4-minute segments within a 1-hour epoch. Using raw ECG has no processing overhead, making it suitable for real-time or low-resource scenarios. Since no computational time is required for raw ECG, it was not included in the table.

Table 5.7: The execution time for data generation of all applied methods.

Generated data type	1-hour ECG epoch (ms)	4-minutes ECG segment (ms)
Spectrogram	1.55	0.022
NN Interval + HR Features	7453.0	104.972

Spectrogram generation is highly efficient, requiring just 0.022 milliseconds per segment, which makes it over 4700 times faster than HR feature extraction which, takes over 100 milliseconds per segment.

Fig. 5.7 shows the classification results of all models which are compared across the five HIE-EEG grades explained in chapter 2. For each 1-hour ECG epoch the corresponding EEG grade is used. Each bar represents the correct and incorrect binary classification of 1-hour of ECG epochs by the respective classifier. For each grade, the bars show the number of incorrect classifications stacked on top of the correct ones. The chart highlights the misclassification patterns across five different EEG-based severity levels, with the total number of 1-hour of ECG epochs per grade annotated above each bar. A vertical line marks the borderline between category 0 and category 1.

An important finding was the consistent misclassification of Normal/mild HIE-EEG cases as category 1. Upon further expert review by a neurophysiologist, 17 one-hour EEG segments were flagged as ambiguous, falling between mild and moderate classifications. Given the subjective nature of EEG grading and the signal complexity, these ambiguous cases were difficult to label. Therefore, an analysis of the model performance was varied when the corresponding 1-hour ECG epochs were excluded. As illustrated in the confusion matrices in Fig. 5.8, this exclusion led to a marked drop in misclassification rates across all models. Furthermore, as detailed in Table 5.8, specificity improved for all models, however sensitivity remained unchanged.

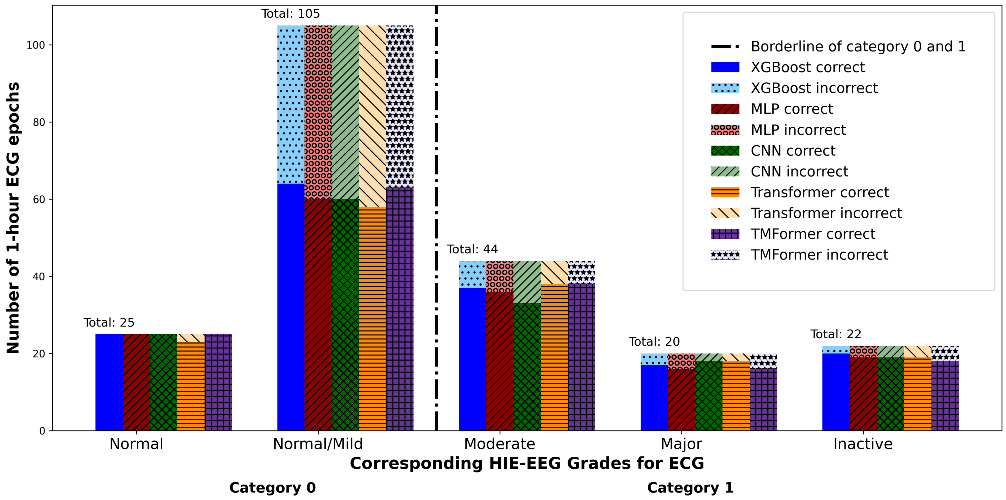


Figure 5.7: The classification results of all models for each 1-hour ECG epoch across the corresponding five HIE-EEG grades for the ECG signal. Each bar shows correct (bottom) and incorrect (top) classifications of 1-hour ECG epochs. A vertical line marks the threshold between category 0 (TH not required) and catergory 1 (TH required), used for binary classification.

Table 5.8: Evaluation measures of all models using the full dataset and after excluding borderline HIE 1-hour epochs reported in terms of Sensitivity and Specificity.

Model	Acc (%)	Sens (%)	Spec (%)
XGBoost (all data)	75.46	86.05	68.46
XGBoost (borderline HIE removed)	80	86.05	75.44
MLP (all data)	72.22	82.56	65.38
MLP (borderline HIE removed)	76.5	82.56	71.93
CNN (all data)	71.76	81.4	65.38
CNN (borderline HIE removed)	77	81.4	73.68
Transformer (all data)	72.22	87.21	62.31
Transformer (borderline HIE removed)	77	87.21	69.3
TMFormer (all data)	74.07	83.72	67.69
TMFormer (borderline HIE removed)	78.5	83.72	74.56

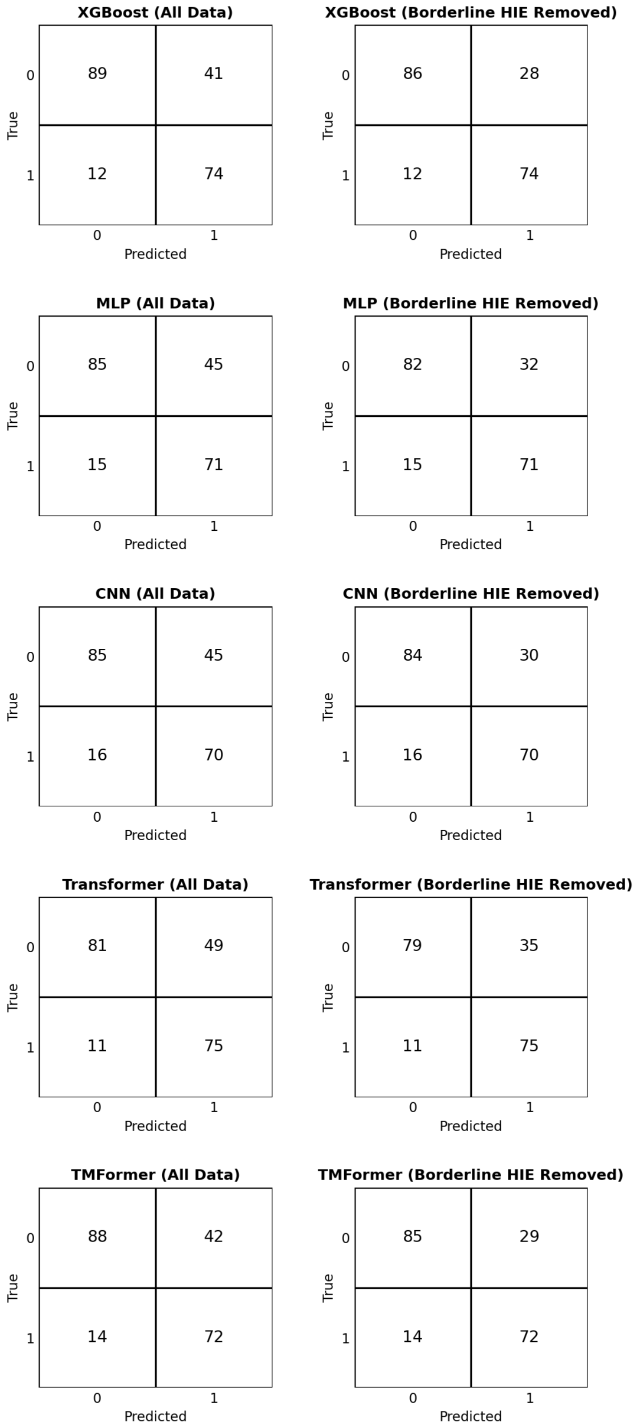


Figure 5.8: Confusion matrices of all models using the full dataset and after excluding borderline HIE 1-hour epochs.

Since the aim of this study was to develop an automatic real-time system for clinical support in hospitals to assist clinicians in decision-making regarding TH, the model's predictions are shown here for a continuous ECG recording from an example neonate between hour 6 and hour 48 after birth. Fig. 5.9 shows the prediction probability of binary classes (TH not required (category 0) and TH required (category 1) produced by XGBoost and MLP models using HR features, CNN using ECG spectrogram and TMFormer using raw ECG. Predictions were generated for each segment, and a rolling average window was applied to smooth the probability curves. Vertical lines indicate the boundaries of five 1-hour ECG epochs labelled by a neurophysiologist, with the assigned target categories displayed above each epoch. As shown the predicted probabilities increase when the category is 1 and decrease when the category is 0. The infant condition improved during hours 30 to 40, therefore the category is zero, however it may have worsened later due to anti-seizure medication. When comparing the prediction curves, the probabilities generated by TMFormer using the raw ECG method more closely align with the target categories, indicating greater confidence and accuracy in distinguishing between category 0 and category 1 compared to other methods. An important observation is that, the predictions from XGBoost and MLP using extracted features show missing segments between hours 6 and 9 and 30 and 33. These gaps correspond to noisy portions of the ECG signal, where QRS peak detection using the Pan-Tompkins algorithm failed. As a result, heart rate features could not be computed for those segments, and they were subsequently discarded. This issue is not observed in other methods, highlighting a limitation of HRV-based approaches for real-time applications.

To enhance visual understanding and interpretation, the model predictions for continuous ECG recordings from two additional neonates are presented. Fig. 5.10 illustrates the predicted HIE probabilities for a neonate with target category 1,

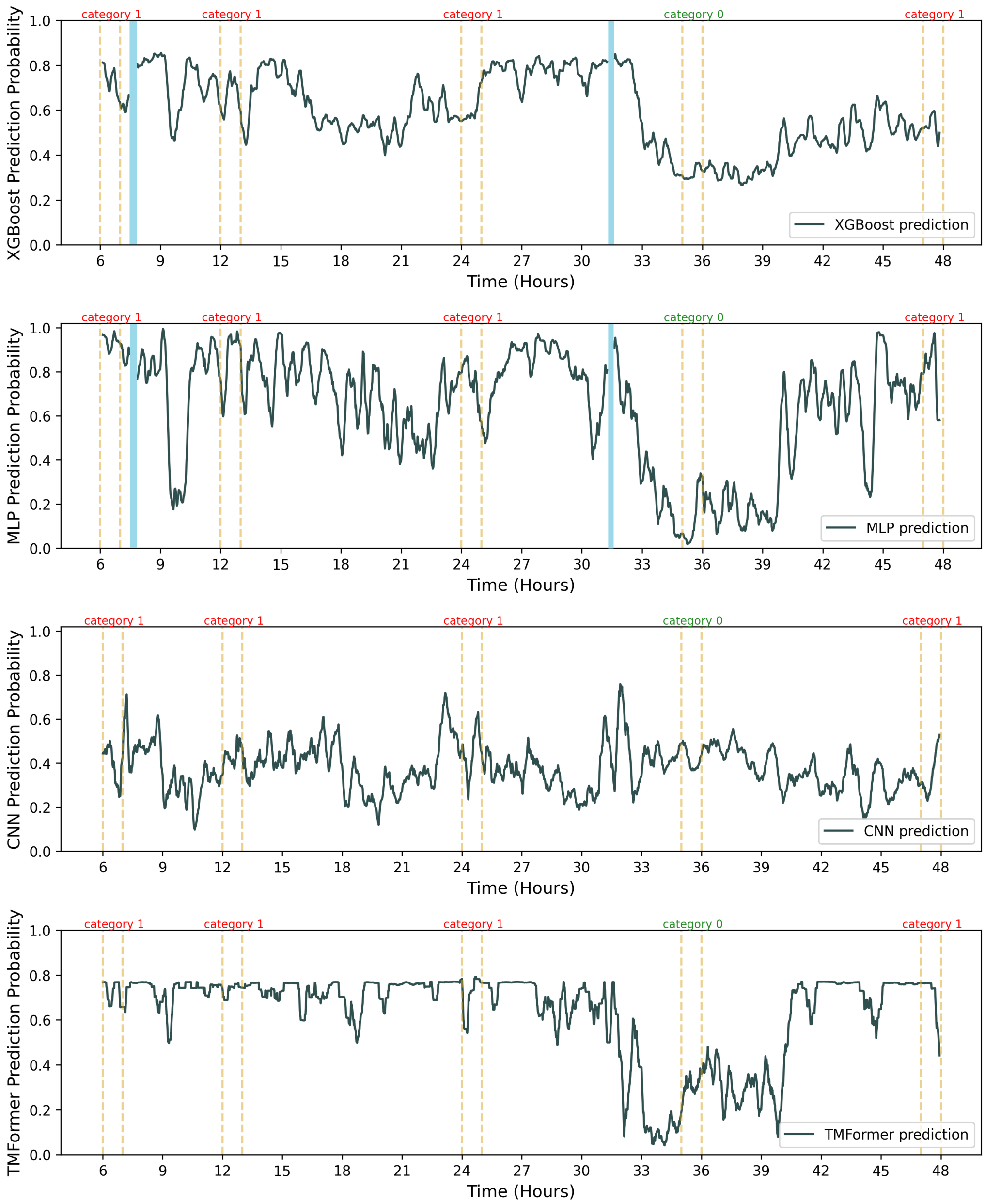


Figure 5.9: The model predictions of a continuous ECG recording from a single neonate between hour 6 and hour 48 after birth with fluctuating HIE grade (category 0 and 1). The plots, from top to bottom, show: XGBoost predictions using HR features, MLP predictions using HR features, CNN predictions using ECG spectrograms, and TMFormer model predictions using raw ECG data.

covering the period from hours 6 to 48 after birth. Notably, the TMFormer model using raw ECG, produces a consistently strong prediction, showing a high degree of certainty in predicting the correct target category. Also, the predictions from XGBoost and MLP using extracted features show missing segments between hours 9 and 12 and 18 and 21, corresponding to noisy intervals in the signal. Fig. 5.11 displays the prediction probabilities for a neonate with target category 0, recorded between hours 6 and 34. This signal had minimal noise, resulting in no intermittent gaps. While predictions appear to be less stable in this case, the TMFormer again demonstrates more reliable and consistent performance using the raw ECG input.

As the result of each classifier varied across patients, an ensemble analysis was performed using all four methods: XGBoost and MLP using HR features, CNN using spectrogram and TMFormer using raw ECG. Since the models were trained on different feature types, their predicted probabilities were combined without additional training or fine-tuning, using simple averaging. This ensemble analysis was conducted to assess the relative contribution of each model to overall performance. A leave-one-out approach was employed, where each method was removed in turn and the AUC of the remaining ensemble was calculated. The results presented in Table 5.9 show that the full ensemble achieved an AUC of 82.03%. Notably, the exclusion of TMFormer led to the largest decrease in performance, indicating its substantial contribution to the ensemble's predictive capability. Conversely, the removal of XGBoost resulted in a slight performance improvement, suggesting that it may introduce some degree of redundancy or noise in the combined predictions.

Since timely identification of infants who may benefit from therapeutic hypothermia (TH) is critical, the performance of TMFormer was evaluated using ECG recordings from progressively earlier time windows after birth. Specifically, model

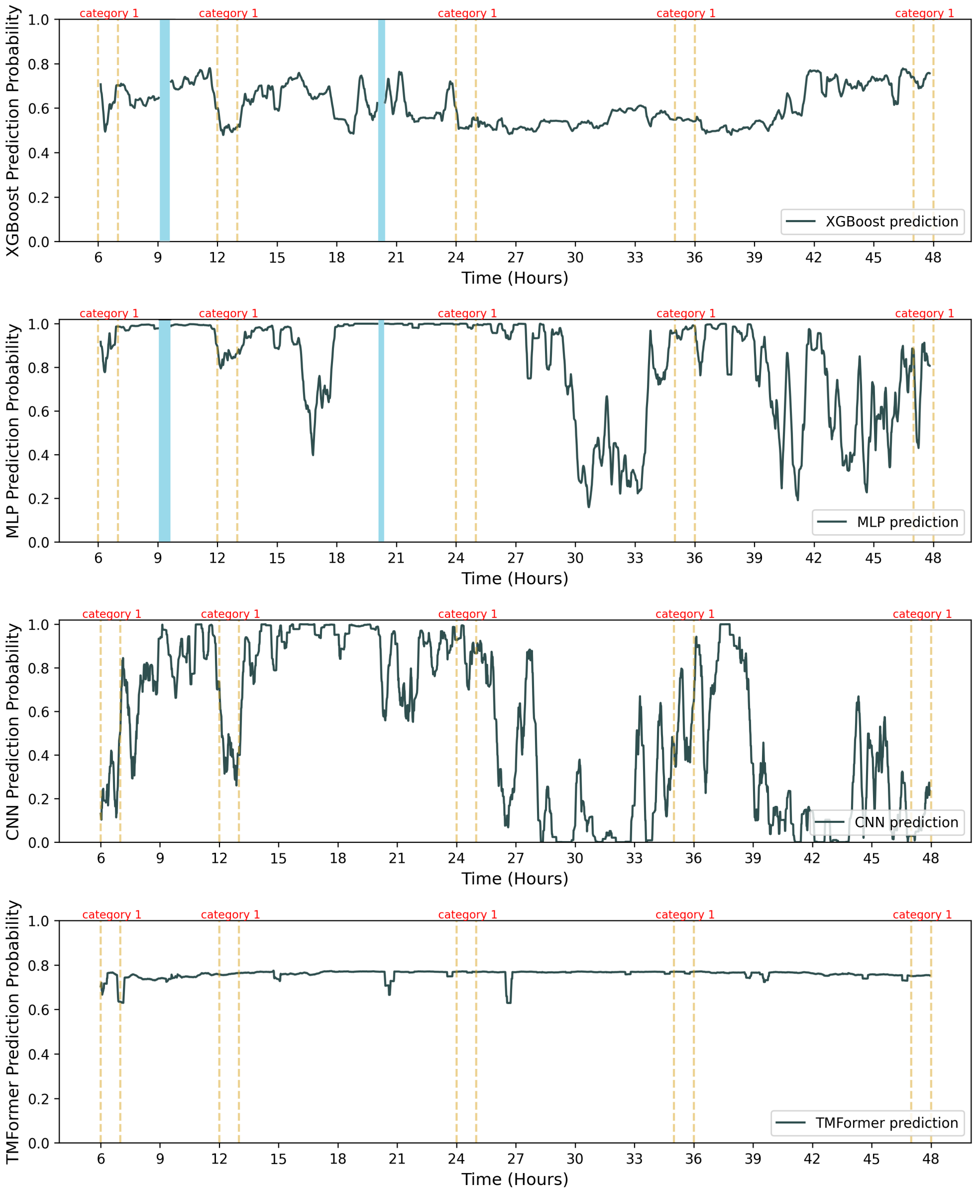


Figure 5.10: The model predictions of a continuous ECG recording from a single neonate between hour 6 and hour 48 after birth with target category 1. The plots, from top to bottom, show: XGBoost predictions using HR features, MLP predictions using HR features, CNN predictions using ECG spectrograms, and TMFormer model predictions using raw ECG data.

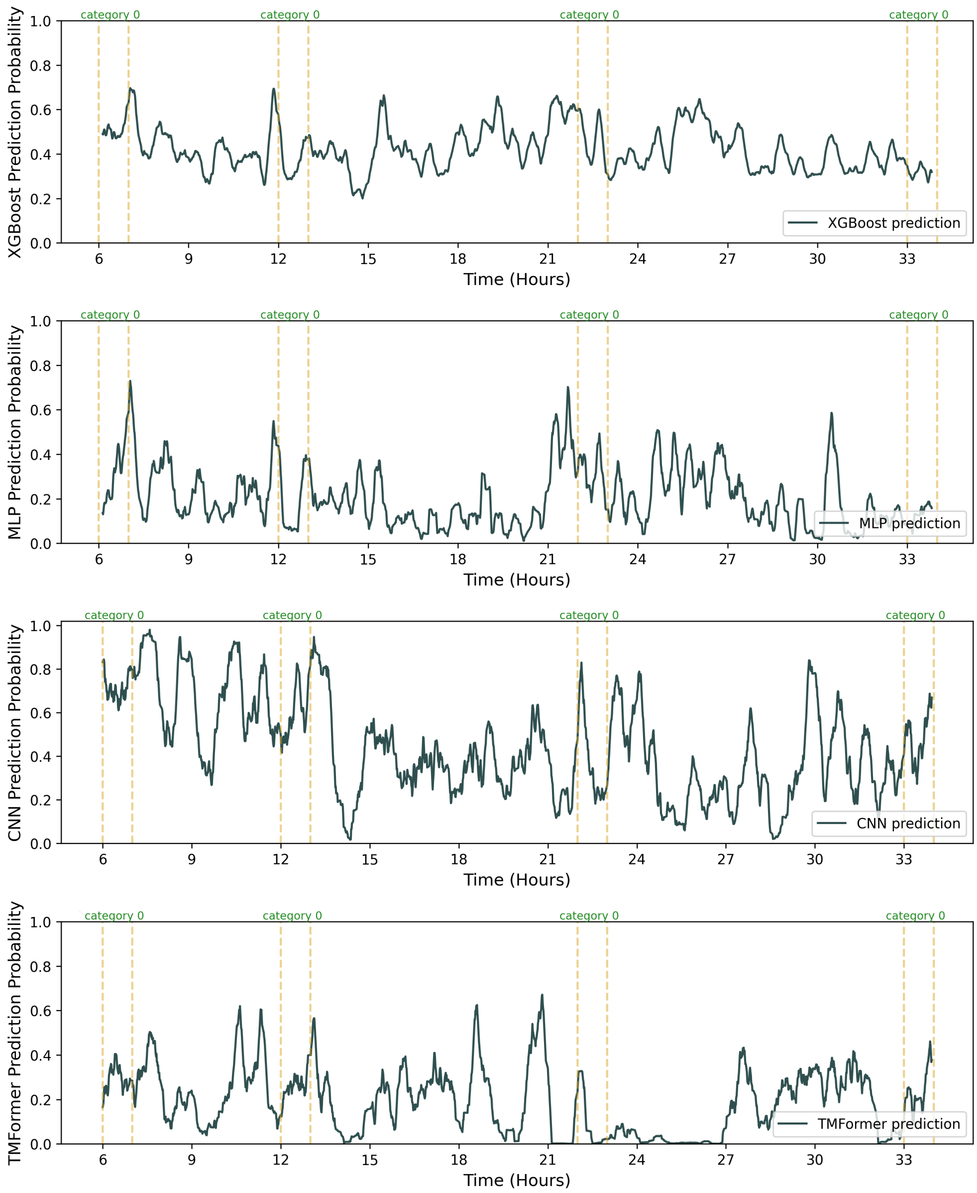


Figure 5.11: The model predictions of a continuous ECG recording from a single neonate between hour 6 and hour 34 after birth with target category 0. The plots, from top to bottom, show: XGBoost predictions using HR features, MLP predictions using HR features, CNN predictions using ECG spectrograms, and TMFormer model predictions using raw ECG data.

Table 5.9: AUC scores for the full ensemble and after removing each model. XGBoost and MLP using HR features, CNN using spectrogram and TMFormer using raw ECG methods were used and the impact of each model on ensemble performance is shown.

Models	AUC (%)
XGBoost + MLP + CNN + TMFormer	82.03
XGBoost removed	82.35
MLP removed	81.88
CNN removed	81.57
TMFormer removed	81.08

Table 5.10: TMFormer performance on 1-hour ECG epochs from the first 12 and 24 hours of birth.

Hours	Number of 1-hour ECG epochs	AUC (%)	Sensitivity (%)	Specificity (%)
12	84	85.91	88.24	78
24	135	84.9	88.68	70.73

performance was assessed using the first 12 hours and the first 24 hours of postnatal ECG data and the results are shown in Table 5.10. Using 84 one-hour ECG epochs from the initial 12 hours of life, the model achieved an AUC of 85.91%. When extended to the first 24 hours, the model achieved a comparable AUC of 84.9%. Notably, both results exceed the performance obtained using the full 48-hour dataset, 216 one-hour epochs, which yielded an AUC of 80.21%. These findings suggest that TMFormer is capable of identifying TH-eligible infants using substantially less data, highlighting its potential value for early clinical decision-making.

5.7.2 Seizure detection

The performance of all models tested on the unseen ANSeR2 dataset are shown in Table 5.11. Based on the results, the CNN-ViT model using ECG spectrograms achieved the highest AUC and sensitivity of 61.57% and 62.06% respectively, while the CNN model attained the highest specificity of 88.63%, respectively.

Table 5.11: Seizure detection performance on the test dataset (ANSeR2) using all the proposed methods.

Model	Input	Sensitivity (%)	Specificity (%)	AUC (%)
TMFormer	Raw ECG signal	25.07	84.85	59.2
CNN ViT	ECG spectrogram	62.06	52.03	61.57
CNN	ECG spectrogram	19.09	88.63	61.17
RF	HRV features	50.52	64.01	60.04
MLP	HRV features	30.72	78.43	55.93

5.8 Conclusion

This chapter demonstrates the potential of utilizing raw ECG signals in combination with a state-of-the-art transformer-based model for HIE grading. The use of raw ECG signals offers key advantages, such as eliminating the need for preprocessing and enabling real-time analysis, enhancing its clinical practicality. The model produced more reliable and stable prediction probability curves over time, indicating stronger confidence and more consistent behaviour in its outputs compared to traditional HRV-based and spectrogram-based CNN approaches, while also achieving slightly higher evaluation metrics. Additionally, the lightweight TMFormer used in this study is efficient enough to run on portable devices and remains robust under noisy conditions.

The evaluatioon metrics were slightly compared to traditional HRV-based and spectrogram-based CNN approaches, however the prediction probablity curves over time produced by the model are more stable and confident.

TMFormer achieved an AUC of 80.21% for HIE grading and the comparison of cumulative attention flow with the extracted NN intervals of the ECG signal, showed that the model's attention consistently emphasizes regions of the ECG signal that align with changes in the NN interval pattern, suggesting a connection between the model's focus and heart rate dynamics, which are often altered in cases of HIE. After checking the HIE grade result of models with a neurophysiologist, excluding ambiguous cases which were difficult to label, led to a marked drop in misclassification rates across all models. After visualizing the predictions of all the models on continuous ECG recordings, the TMFormer model using raw ECG demonstrated consistently strong and stable performance. It showed a high degree of confidence in correctly predicting the HIE grade, with no missing segments due to noise, outperforming the other methods.

The seizure detection task achieved better performance with the CNN–ViT model using ECG spectrograms, reaching an AUC of 61.57%, compared to TMFormer applied to raw ECG signals, which obtained an AUC of 59.2%. CNN-ViT performance was higher than that of HRV-based methods (AUC = 60.04%) and spectrogram-based CNNs (AUC = 61.36%) as well. A likely explanation is that seizure detection relies on short ECG segments, where derived representations such as statistical features or spectrograms capture more informative patterns than raw signals. Additionally, the 5-fold cross-validation results of the CNN–ViT show higher sensitivity than specificity, whereas all other methods exhibited higher specificity. This highlights the CNN–ViT's superior ability to correctly identify seizure segments.

6 | Conclusions

6.1 Concluding summary and main contributions

Hypoxic-Ischemic Encephalopathy (HIE), the most common underlying cause of seizures in the neonatal period, can lead to a wide range of neurological and developmental disorders. Timely diagnosis and treatment of these conditions are critical, highlighting the growing need for automated detection systems. Although EEG is the current gold standard for neonatal brain injury monitoring, it is often limited in clinical practice due to the need for specialised equipment, expert interpretation, and continuous electrode placement on the scalp. As an alternative, ECG signals are more widely available, easier to acquire in neonatal intensive care units, and continuously monitored as part of routine care. By analysing and processing ECG signals, valuable diagnostic information related to autonomic and physiological changes associated with brain injury can be extracted. Applying machine learning methods to these ECG-derived features enables automated classification, which can support clinicians in making timely clinical decisions.

This work presents automated system designs to serve two main purposes:

- Detection of Seizures from Neonatal ECG signals
- Grading the severity of HIE injury regarding TH requirement using neonatal ECG signals

For this purpose, multiple representations of the ECG signal were explored to make more efficient use of the available data. The alternative approaches proposed in this thesis, such as ECG spectrograms and raw ECG signals, outperformed the traditional HRV feature-based method. To the best of our knowledge, this work is the first to investigate the use of both ECG spectrograms and raw ECG signals for neonatal seizure detection and HIE grading. These representations eliminate the need for preprocessing steps such as heart-rate calculation and manual feature extraction, and they do not suffer from missing data caused by peak-detection failures. Therefore, the proposed representations support real-time analysis even under noisy conditions, improving the automatic seizure-detection and HIE-grading capabilities for practical clinical use. Additionally, the processing time required for generating input data for the models was significantly reduced.

These representations enable the application of modern deep-learning architectures developed in this thesis, including the custom-designed CNN and Transformer models. The proposed Transformer-based model learns rich and discriminative representations directly from the data through a tailored tokenization approach, enabling robust and physiologically meaningful feature extraction.

As a result, the methods introduced in this work outperformed and show consistent improvements in metrics such as AUC compared to HRV-only baselines. Additionally, the lightweight architectures were specifically designed for efficiency, making them suitable for deployment on portable devices such as smartphones while maintaining strong performance.

Additionally, this thesis benefits from a very large neonatal ECG dataset compared to those used in previous studies. All methods proposed in this thesis, were implemented and compared on the same dataset to ensure a fair and consistent benchmarking process. This enables more generalizable evaluation and more robust performance metrics, including sensitivity, specificity, and AUC.

urthermore, a detailed analysis of the data and results was conducted to evaluate the contribution and effectiveness of each proposed method. The performance of the models was assessed in relation to neurophysiologist-provided labels, demonstrating how the models capture relevant changes in the ECG signal associated with neonatal brain activity. In particular, the results highlight that the models are sensitive to seizure-related patterns, including variations in the number and duration of seizure events. In addition, explainability techniques were applied to the Transformer-based model to provide further insight into its decision-making process and improve interpretability of the learned representations.

HIE grading in this study, achieved using raw ECG signals and the Transformer model, reached an AUC of 80.21%. While this is lower than EEG-based results reported by Yu et al. [**?**] (AUC 86.23% on 181 patients and 653 hours of recordings), the dataset used in this thesis is substantially larger, comprising 131 patients and 1,756 hours of recordings, providing a more robust and generalizable evaluation. For seizure detection, the highest AUC obtained was 61.57%, which is lower than EEG-based results reported by O'Shea [**?**] (AUC 98.5%). However, the current study utilized 65 patients with 2,442 hours of recordings, compared to only 18 patients and 834 hours in O'Shea's work, offering stronger reliability and broader generalizability. These results highlight the potential of ECG-based approaches for HIE grading and seizure detection. Furthermore, the high accessibility of ECG signals suggest that deployment on portable devices could facilitate further improvements in model performance.

The main conclusions and contributions of each chapter are now summarized.

Chapter 2 outlined relevant clinical information regarding neonatal HIE and seizures, emphasizing the importance of timely treatment through therapeutic hypothermia. ECG signals were discussed as an alternative to EEG signals, including their potential for supporting the development of automated systems for HIE grading and seizure detection to support clinicians. Previous studies on HIE grading and seizure detection methods were reviewed, highlighting the need for extensive investigation into the use of ECG signals for these tasks. Additionally, approaches employed in this thesis were briefly summarized. Finally, the datasets used to extract information from the ECG signals and to train and evaluate the machine learning architectures developed in this work were described in detail.

HIE grading and seizure detection were investigated using traditional methods in **chapter 3**. In this chapter, the extraction of heart rate from the ECG signal using the Pan–Tompkins method was explained, and its limitations were discussed. Signal preprocessing techniques were examined to identify the most effective approaches for noise removal and signal segmentation. Furthermore, the most informative HRV-based features and classical machine learning models, selected based on findings from previous state-of-the-art studies, were described and further explored. These methods were initially evaluated on a small dataset to identify the most discriminative features and determine the optimal machine learning models. The experimental results underscored the critical importance of specific features in both HIE grading and seizure detection. In particular, muNN and LF power emerged as key features across both tasks, consistent with previous findings by Goulding et al. [**?**] [**?**]. Additionally, other HRV features selected in this study, such as Trindex, CSI, and ApEn, also demonstrated effectiveness in improving classifier performance.

Classifier results on the small dataset showed that, for HIE grading, XGBoost using sixteen features proposed in this thesis achieved the highest AUC of 77.75%. For seizure detection, the RF model employing the same sixteen features attained the highest AUC of 69.21%. When these models, along with a MLP, were applied to the larger dataset (ANSeR2 as the training set and ANSeR1 as the test set), the MLP achieved a slightly higher performance of 79.34% for HIE grading. However, for seizure detection, the RF model again performed best, though with a lower AUC of 60.04%. The drop in seizure detection performance, from 69.21% on the smaller dataset to 60.04% on the larger one, suggests that the results obtained using leave-one-patient-out cross-validation on the small dataset may not generalize well to unseen data. In summary, while HIE grading performance improved slightly on the larger, patient-independent ANSeR dataset, seizure detection performance decreased notably, indicating potential overfitting and limited generalizability of the small dataset models.

To further leverage the entire ANSeR seizure dataset and obtain a more comprehensive assessment of model performance, a 5-fold cross-validation was applied. The results showed that model performance varied across different folds; however, the overall seizure detection outcomes were consistent with those observed in the ANSeR1 and ANSeR2 splits. Sensitivity was notably lower than specificity, and the AUC was lower than accuracy in every fold. This indicates that the model is more effective at correctly identifying non-seizure segments than detecting seizure segments. These patterns reflect the underlying class imbalance and suggest that the decision boundary may be biased toward the majority (non-seizure) class. Despite applying SMOTE to oversample seizure segments, the imbalance persisted and continued to affect the results. The relatively weak overall performance can likely be attributed to several factors, including the severe class imbalance that could not be fully mitigated by oversampling, the short duration of seizure events, potential misalignment of labels, high variability between seizure and non-seizure feature spaces, and the presence of multiple seizure types.

Additionally, the AUC of each model was examined in relation to the number and duration of seizure events in each fold's test set. In this context, a seizure event refers to the full seizure episode from the long recording, rather than the 2-minute segments. For each fold, the average of the maximum seizure event lengths per patient and the number of events longer than 10 minutes were calculated. The results showed that the folds containing, on average, the longest seizure events achieved the highest AUC, while folds with shorter seizure events made it more difficult for the models to detect meaningful patterns. Furthermore, in folds where the average seizure length was not particularly short but the number of seizure events was high, the reduced duration of individual events also appeared to hinder model performance.

In **Chapter 4**, entirely new methods were investigated that had not previously been applied to ECG signals for neonate HIE grading or seizure detection. In this chapter, an optimized time–frequency representation of the ECG signal was developed using spectrograms computed directly from the raw ECG data, enabling the use of deep learning models such as CNNs. This approach significantly reduced the computation time for model input and overcame several limitations associated with traditional heart rate calculation, handcrafted feature extraction, and classical classification methods.

A tailored spectrogram computation was employed, involving log-scaled spectrograms cropped to the 7 Hz or 9 Hz frequency range (depending on the task) to produce square-shaped matrices. Instead of using spectrogram images, the spectrogram matrices were used directly as model inputs, which further improved CNN processing efficiency. Lightweight CNN architectures were designed for both HIE grading and seizure detection tasks. These architectures were tailored to the specific input dimensions and spectral characteristics of the spectrograms and were designed to effectively capture hierarchical representations through progressively expanding receptive fields. To address class imbalance in the seizure detection task, data augmentation was applied to increase the number of seizure samples.

Finally, the CNN models trained on the spectrogram representations achieved an AUC of 79.67% for HIE grading and 61.36% for seizure detection, both slightly outperforming traditional HRV-based methods. These results demonstrate the potential of using optimized spectrogram representations and deep learning models for neonate ECG-based HIE grading and seizure detection.

In **Chapter 5**, all additional preprocessing steps, such as heart rate calculation and spectrogram generation, were completely eliminated. This simplification reduced computational overhead, enabled real-time processing, and improved the practicality of the proposed method for clinical applications. Instead of relying on handcrafted features or spectrogram-based representations, the approach directly utilizes raw ECG signals, preserving the rich temporal dynamics that are often lost during traditional feature extraction. This chapter presented a novel approach that, for the first time, employs raw ECG signals for both HIE grading and seizure detection, allowing the application of transformer models capable of capturing long-range temporal dependencies.

In this chapter, a transformer-based model, named TMFormer, was developed. The model introduces a novel tokenization technique specifically designed for ECG signals, incorporating a multi-scale layer to simultaneously capture both global patterns and fine-grained details of the waveform. Additionally, an ECG trend component was proposed in the second stage of the tokenization process to model long-term variations in the signal, thereby providing a more comprehensive representation of HIE severity. When applied directly to raw ECG signals, TMFormer achieved an AUC of 80.21% for HIE grading, outperforming both HRV-based methods and spectrogram-based CNN models. Although the performance for seizure detection was less promising (AUC = 59.2%), the CNN–ViT model designed for this task achieved an AUC of 61.57%, slightly surpassing HRV- and spectrogram-based approaches. This finding demonstrate the efficiency and potential of transformer architectures in comparison to traditional models for ECG-based HIE grading and seizure detection.

The HIE grading results obtained in this thesis are compared to the result reported by Ahmed et al. [**?**], both at around 80%. The seizure detection results are in the same range as those reported in other state-of-the-art studies [**?**] [**?**] [**?**], which report accuracies of about 60%. Previous studies relied solely on HRV features, used much smaller datasets, and typically employed leave-one-out evaluation. The methods in this thesis were applied to much larger datasets. Additionally, training was performed on a fully patient-independent dataset that was never seen during training, further improving the accuracy and reliability of the results.

Another contribution of this thesis is to demonstrate an alternative approach that eliminates the need for heart rate and HRV feature extraction. These features are often unreliable in noisy segments where peak detection fails, resulting in missing data, as illustrated in chapter 5, Fig. 5.9 and Fig. 5.10. Focus of this thesis is to show that using ECG spectrograms and raw ECG signals can achieve performance comparable to HRV-based methods while enabling real-time applicability in clinical settings as there will not be any missing segments even in noisy situation. Additionally, less processing steps make the model more cost efficient and applicable on a small device such as a smart phone, making it possible to be used in resource limited hospitals.

HRV-based methods extract HRV features based on the statistical features available. By contrast, the proposed CNN and Transformer architectures operating on ECG spectrograms and raw signals learn features directly from the data and expect to see features that may not exist in the baseline methods. Thus, it can be inferred that these learned features are as informative as enough manually selected HRV features.

These findings represent progress toward developing supportive tools that complement clinical judgment. By leveraging easily obtainable ECG signals, this method reduces reliance on EEG-based assessments by neurophysiologists and could improve access to accurate HIE assessment in lower-resource healthcare settings. However, seizure detection remains challenging and further analyses on alternative test sets with different patients revealed considerable variability in performance, highlighting uncertainty in the model's decisions, which likely arises from the number, characteristics, and annotations of seizure events.

6.2 Limitations and future work

The extensive availability of ECG data was a key strength of this thesis. Although the datasets used were relatively large, model performance could likely be enhanced by incorporating even larger and more diverse data that better reflect the variability of real clinical conditions. In particular, the seizure detection task was limited by data imbalance, highlighting the need for additional data to capture the full range of neonatal seizure patterns. Additionally, since the ECG signals were extracted from EEG recordings, using clinically monitored ECG, rather than ECG captured as part of the EEG setup, could improve signal quality and overall robustness.

This work also benefited from expert EEG annotations provided by neonatal neurophysiology specialists. A key limitation, however, was variability in inter-rater agreement, particularly for borderline HIE cases. Although labels were derived from EEG recordings, models were trained and evaluated on ECG signals. As a result, in the seizure detection task, ECG changes related to seizures may occur slightly before or after the annotated EEG onset, potentially leading to false positives and reduced model accuracy.

In addition to inherent noise in the ECG signals, a significant limitation of this study is the signal interference in some patients occurring during therapeutic hypothermia and the effects of medication, which may affect data quality.

Future work should focus on larger, more diverse datasets and better alignment between signal modalities to improve model reliability.

Additionally, the use of unsupervised learning methods could be explored, which may be particularly effective when larger datasets are available but lack annotations. Such approaches could help uncover underlying signal patterns and improve model generalization without the need for extensive manual labeling.

Future work could explore wavelet transforms as an alternative to STFT for ECG analysis, as they may better capture non-stationary signal features and reveal additional diagnostic information.

Another promising direction is the development of multi-modal detection models that integrate additional accessible physiological signals, such as respiratory rate, which has been shown to enhance seizure detection performance [**?**].

A | Appendix

Appendix

This appendix provides supplementary background material supporting the methods presented in this thesis. It includes additional details on EEG recording, relevant literature on HIE grading and seizure detection using EEG, the Pan-Tompkins algorithm, HRV features, and traditional machine learning models. These sections are included to provide further context without interrupting the flow of the main chapters.

A.1 EEG recording

For EEG recording, electrodes are typically placed over the central region as well as the frontal, temporal, parietal, and occipital lobes of the brain, using the standardized system of 10-20 electrode placement [**?**]. Each channel is named according to the brain region where the electrode is positioned, with numbers indicating the hemisphere (odd numbers for the left hemisphere and even numbers for the right). Fewer electrodes are used in neonatal EEG due to the smaller head size, including F4, F3, C4, C3, Cz, T4, T3, P4, P3, O1, and O2 [**?**]. The electrical activity detected by each electrode is amplified and converted from an analog to a digital signal, which is then processed and displayed on a computer monitor. The bipolar montage is the most commonly used configuration in neonatal EEG monitoring. In this montage, channels are arranged in a sequential chain, where the electrode serving as the input for one channel simultaneously functions as the reference for the next [**?**]. An example of neonatal EEG signal consisting of 8 channels is presented in Fig. A.1

EEG signals are prone to artifacts, which may be physiological (e.g., movement, respiration, pulse) or environmental (e.g., electrical noise, electrode detachment).

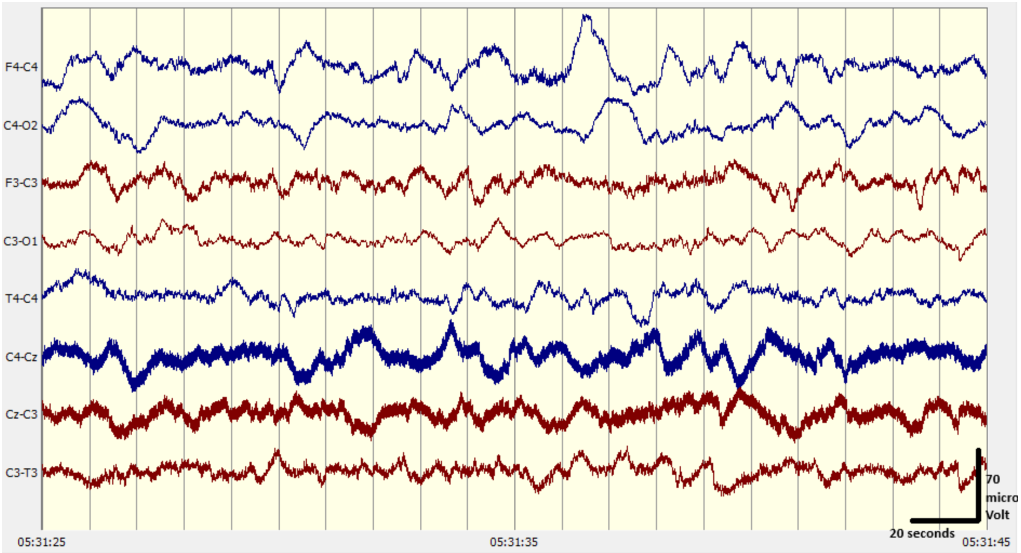


Figure A.1: Example of a healthy neonatal EEG signal from ANSeR dataset used in this thesis.

Artifacts complicate interpretation by masking or mimicking critical diagnostic features [**?**]. Samples of artifact on EEG signal are shown in Fig. A.2

Neonatal EEG signals are classified into different frequency bands, which are used in clinical practice to assess the biological significance of rhythmic brain activity. The frequency bands include alpha (8-13 Hz), beta (14-40 Hz), theta (5-7 Hz) and delta (less than 4 Hz) [**?**].

A.2 Automated HIE grading and seizure detection using EEG signal

Automated analysis of EEG signals using machine learning algorithms has emerged as a promising approach for seizure detection and HIE grading. HIE is a condition that primarily affects neonates, and therefore, no studies on HIE detection have been conducted in adult populations. However, many studies have investigated

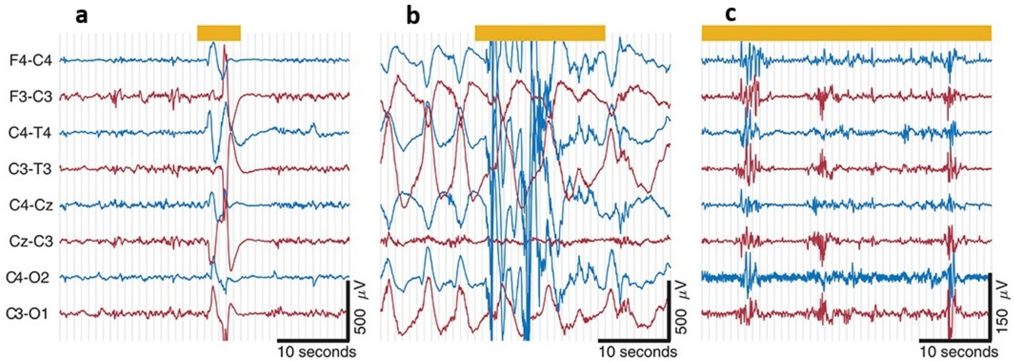


Figure A.2: Examples of infant EEG signal artifacts. Yellow lines at top represent artifact annotation. Artifact is present across all channels in (a,b); artifact in (c) is located in C4-O2, C3-O1, and F4-C4. Movement artifact in (a,b) and muscle artifact on C4-O2 in (c). Also ECG artifact present on C3-O1 and F4-C4 in (c). [**?**].

epileptic seizure detection in adults [**?**] [**?**] [**?**] [**?**] [**?**], showing promising results. Studying and collecting EEG data from neonates is considerably more challenging than in adults [**?**]. The manifestation of neonatal seizures is highly variable, and clinical signs are often extremely subtle [**?**]; therefore, dedicated studies are needed to develop methods specifically tailored to neonatal signals.

Several studies have proposed machine learning–based approaches using EEG signals for grading HIE and detecting seizures in neonates. Typically, these methods consist of a feature extraction stage followed by a machine learning model. The input features can vary widely, ranging from statistical measures to time-frequency representations of the signal, or even the raw EEG signals. Dataset size is also a critical factor influencing the performance and generalizability of these studies. Table A.1 provides a summary of studies on HIE grading and seizure detection using EEG signals, which are discussed in the following subsections.

Table A.1: Summary of studies on HIE grading and seizure detection using EEG signals. AUC: area under the receiver operating characteristic curve (measure of classification performance).

Study	Problem	Number of Patients	Total Time (hours)	Result (%)
Stevenson et al. [?]	HIE	54	62	Accuracy of 83
Ahmed et al. [?]	HIE	54	62	Accuracy of 87
Raurale et al. [?]	HIE	91	338	Accuracy of 69.5
O'Sullivan et al. [?]	HIE	181	653	Accuracy of 83.6
Yu et al. [?]	HIE	181	653	Accuracy of 86.09
Aarabi et al. [?]	Seizure	10	64	Detection rate of 79.7
Temko et al. [?]	Seizure	17	267	Detection rate of 89
Ansari et al. [?]	Seizure	22	74.3	Detection rate of 77
O'Shea et al. [?]	Seizure	18	834	AUC of 98.5

A.2.1 HIE grading using EEG

Early techniques for HIE grading focused on the use of quantitative EEG features and traditional machine learning models. Stevenson et al. [?] classified the degree of abnormalities in one-hour EEG recordings using a multi-class linear classifier. The features were derived from amplitude modulation and the instantaneous frequency of the EEG signals. This approach achieved a leave-one-subject-out cross-validation accuracy of 83% on a relatively small dataset consisting of 62 hours of eight-channel EEG recordings from 54 neonates. Ahmed et al. [?] worked on the same dataset, employing a feature set of 55 basic descriptive features extracted from the frequency, time, and information-theoretic domains. They applied a generic feature set combined with a support vector machine (SVM) supervector approach. This method resulted in a notable improvement in grading performance, achieving an accuracy of 87%.

Further, time–frequency representation of EEG signal and deep learning models were used. Sumit A. Raurale [?] developed an approach combining quadratic time–frequency distributions with a convolutional neural network (CNN), using the same dataset as previous studies. However, they evaluated the model on a large unseen dataset comprising 91 neonates with a total 338 hours of eight-channel EEG recordings. The model performance achieved an accuracy of 69.5%. The lower performance compared to earlier studies is likely due to differences in evaluation methodology as the previous works used leave-one-out or within-dataset validation, which can overestimate performance, whereas testing on a completely unseen dataset provides a more realistic measure of generalization across patients. Further, O'Sullivan et al. [**?**] applied the same architecture, augmented with artifact removal, to a larger dataset comprising 653 hours of EEG recordings from 181 neonates. Their method achieved an accuracy of 83.6% in performance evaluation. Further, Yu et al. [**?**] employed fully convolutional neural networks on raw EEG signals to grade HIE, using the same dataset as O'Sullivan et al. [**?**]. The proposed method achieved an accuracy of 86.09% and an area under the curve (AUC) of 86.23%.

Based on the research results, accuracy did not exceed 87%. Although increasing the amount of data and improving the method can lead to a more robust result, obtaining EEG signals from neonates has its own challenges, as discussed earlier.

A.2.2 Seizure detection using EEG

For seizure detection, earlier studies primarily focused on statistical feature extraction and selection using traditional machine learning models, the same as HIE grading methods. Aarabi et al. [**?**] developed a system for detecting seizure and non-seizure segments in EEG recordings, with the goal of identifying discriminative features for seizure detection, particularly in neonates admitted to intensive care units. Their approach involved bandpass filtering, amplitude normalization, and automatic artifact detection to preprocess the EEG signals. Five groups of features were extracted from the EEG signal: time-domain features, frequency-domain features, autoregressive coefficients, wavelet features, and cepstral features. Feature selection was performed based on feature relevance and redundancy, and EEG classification was conducted using a trained multilayer back-propagation neural network. The system was evaluated on EEG recordings from ten neonates, achieving a sensitivity of 74% and a selectivity of 70.1%. Temko et al. [**?**] proposed a multi-channel, patient-independent neonatal seizure detection system based on a Support Vector Machine (SVM) classifier. A total of 55 features were extracted from the EEG signals, encompassing frequency-domain, time-domain, and information-theoretic characteristics. The system was validated on a large clinical dataset comprising 267 hours of EEG recordings from 17 neonates. It achieved an average detection rate of 89% with one false detection per hour; 96% with two false detection per hour; and 100% with four false detection per hour. Subsequent studies further explored EEG-based features in combination with SVM models [**?**] [**?**].

Furthermore, other studies have applied deep learning models directly to raw EEG signals. Ansari et al. [**?**] employed a convolutional neural network (CNN) for automatic feature extraction from raw multichannel EEG data, followed by classification into seizure and non-seizure segments using a random forest model. Their approach achieved a seizure detection rate of 77% on a test set comprising EEG recordings from 22 neonates. O'Shea [**?**] designed a fully convolutional neural network to detect neonatal seizures directly from raw EEG signals. The system was evaluated on a large database comprising 834 hours of continuous EEG recordings and achieved an AUC of 98.5%.

Achieving an AUC of 98.5% is highly promising in the seizure detection task using the EEG signal, and the results were significantly improved using deep learning

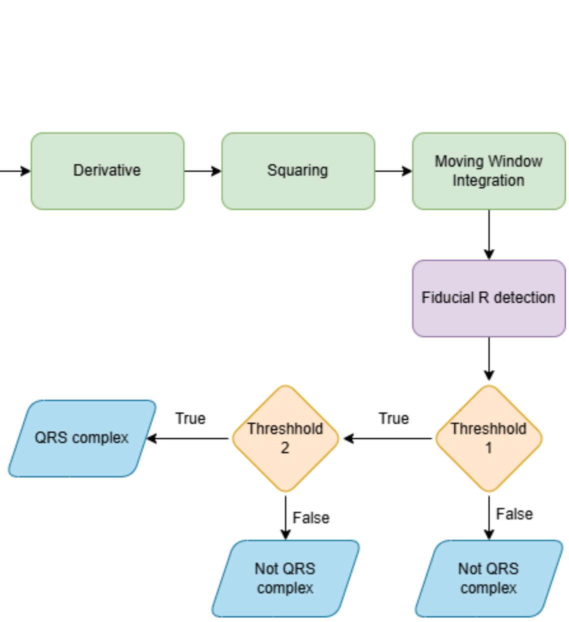


Figure A.3: Block diagram of the Pan–Tompkins algorithm based on [?].

models with a larger dataset.

A.3 Pan–Tompkins algorithm

The Pan–Tompkins algorithm [?] consists of two stages: pre-processing and decision-making. In the pre-processing stage, the signal quality is enhanced through noise removal, and QRS complexes are emphasized by smoothing the signal and enhancing the slope and amplitude. The decision stage identifies R-peaks by applying adaptive thresholds. The overall algorithm is illustrated in Fig. A.3.

The pre-processing steps are as follows:

- Band-pass filtering: Combines low-pass and high-pass filters (≈5–15 Hz) to reduce noise while preserving the QRS energy. High-frequency noise (EMG, power line interference) is removed by the LPF, and low-frequency noise (baseline wander) by the HPF.

- Derivative: Highlights the slope of the QRS complex and suppresses low-frequency P- and T-waves.
- Squaring: Converts all signal values to positive and amplifies high-amplitude regions corresponding to the QRS complex, reducing the influence of T-waves.
- Moving Window Integration (MWI): Integrates the signal over a window matching the QRS duration (here, 15 samples / 58.6 ms) to enhance waveform features for reliable peak detection.

Figure A.4 shows the complete preprocessing pipeline applied to a sample ECG signal.

The decision-making steps are as follows:

- Fiducial R detection: The QRS complex aligns with the rising edge of the integrated waveform, whose duration reflects QRS width. A fiducial point, such as the maximal slope or R-peak, can be selected from this rising edge to mark the QRS location.
- Adaptive thresholds: Pan–Tompkins uses two adaptive threshold levels, where one level is half of the other. One threshold level tracks signal peaks and one tracks noise peaks. The main threshold is set between the two levels and continuously adjusts over time. If a beat is missed, the algorithm uses the lower threshold to catch peaks that fall between the two levels, reducing false negatives. Fig. A.5 illustrates the integrated signal, the detected R-peaks, and the adaptive thresholds in Pan–Tompkins method. The algorithm tracks both signal peaks and noise peaks, shown by the dashed lines in the figure. A dynamic threshold is then computed between these two levels, allowing the method to adapt continuously to changes in signal quality.

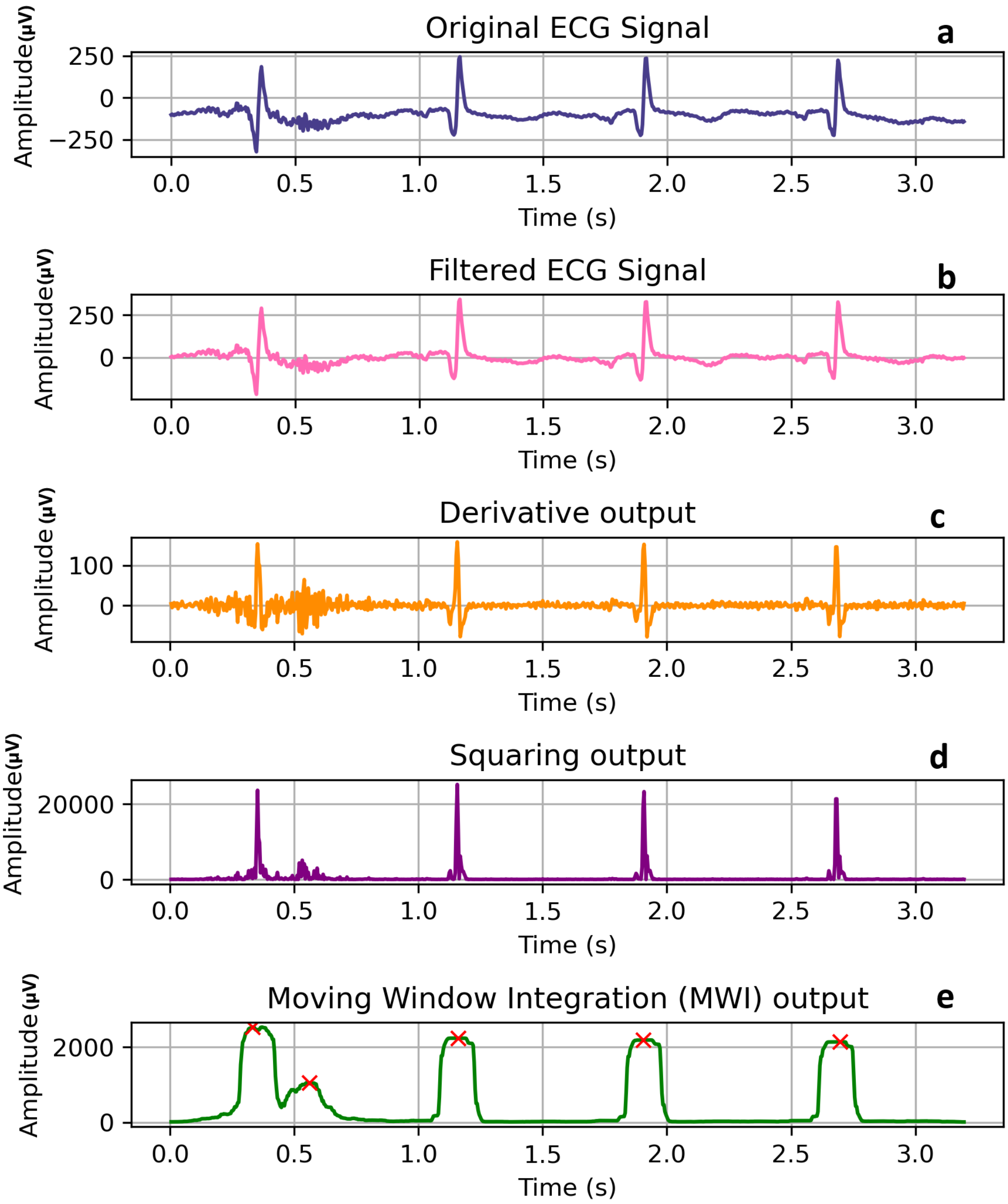


Figure A.4: Preprocessing steps of the Pan–Tompkins algorithm are illustrated. (a) shows the original ECG signal; (b) the filtered signal; (c) the derivative of the filtered ECG; (d) the squared derivative; and (e) the detected peaks after applying moving-window integration.

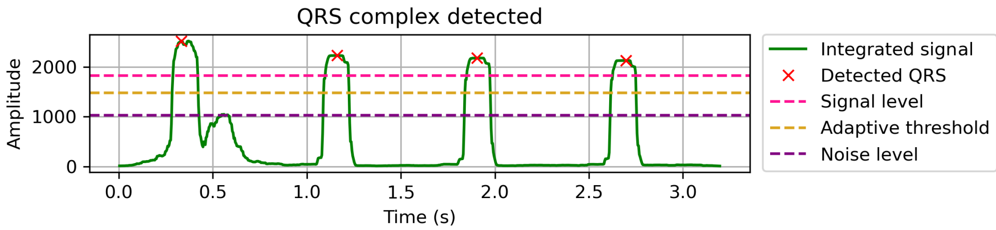


Figure A.5: Integrated signal with adaptive thresholds between noise and signal level thresholds and detected R-peaks.

- NN intervals: After detecting the R-peaks, the time differences between successive peaks are computed as R–R intervals. Peaks with a maximal slope less than half that of the preceding QRS are classified as T-waves, otherwise they are called QRS complexes. The intervals between the corrected R-peaks, called NN intervals (normal-to-normal), form the basis for heart rate variability (HRV) analysis.

A.4 Heart rate variability features

In this section, time-domain and frequency-domain features used in this thesis are explained in detail.

A.4.1 Time domain features

1. Mean of the NN interval series (muNN):

$$muNN = \overline{NNI} = \frac{\sum_{j=1}^{n} NNI_j}{n} \tag{A.1}$$

   Where n is the number of NN intervals in a segment and NNI_j is jth value.

2. The standard deviation of NN Series (SDNN):

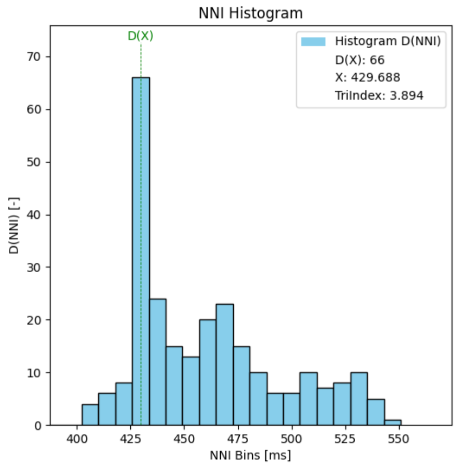


Figure A.6: Examples of Triangular Index feature extracted from an NN interval histogram with maximum histogram height D(X) of 66 corresponding to an NN value of 429.688

$$SDNN = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(NNI_j - \overline{NNI})^2} \tag{A.2}$$

3. Triangular index (Tri): This geometrical feature is derived from the NN interval histogram. The HRV Triangular Index is calculated as the total number of NN intervals divided by the maximum height of the histogram. Triangular index (Tri) is calculated as below.

$$Tri = \frac{n}{\max(D(NNI))} \tag{A.3}$$

Fig. A.3 shows the Triangular Index calculated based on the NN interval histogram with D(x) as the maximum height of the histogram bar corresponding to x in the histogram.

The histogram bin size is set to 7.8125 ms, derived from the minimum recommended ECG sampling frequency of 128 Hz specified in the HRV Guidelines [?].

4. Triangular Interpolation (TINN): The Triangular Interpolation of NN intervals is a geometrical HRV measure derived from the NN interval histogram. It represents the baseline width of a triangle obtained by approximating the histogram distribution of NN intervals. The triangle is fitted such that it best approximates the overall shape of the histogram, and TINN is defined as the width of the base of this triangle.

$$TINN = N - M \tag{A.4}$$

where M and N denote the left and right endpoints of the triangular interpolation that minimize the fitting error to the NN interval histogram. TINN reflects the overall variability of NN intervals, with larger values indicating greater heart rate variability.

A.4.2 Frequency domain features

To enable frequency-domain analysis of the HR time series, the NN interval sequence needed to be resampled at uniform time intervals. This was accomplished using spline interpolation technique, which fits low-degree polynomials to localized segments of the data, ensuring a smooth and continuous representation. The sampling rate of the interpolated signal is effectively up sampled to 4 Hz , ensuring reliable spectral estimates up to 2 Hz (half the sampling rate, according to the Nyquist criterion). Frequency domain parameters are calculated by estimating the power spectral density (PSD) of the resampled NN series computed using Welch's method [?] using a hamming window. The choice of these frequency-domain features and their corresponding frequency band ranges is based on the studies by Goulding [? ?] and Doyle et al. [?]. The following four features are extracted from the periodogram:

4. Power in the Very Low Frequency band $0.01 - 0.04$ Hz (VLF)
5. Power in the Low Frequency band $0.04 - 0.2$ Hz (LF)
6. Power in the High Frequency band $0.2 - 2$ Hz (HF)
7. LF/HF power ratio

Since the PSD amplitudes were very small across all frequency bands, the logarithmic scale of power values were used. An example of the power spectral density of an NN-interval segment estimated using Welch's method is shown in Fig. A.7, where the spectrum is partitioned into VLF, LF, and HF frequency bands, and the corresponding log power in each band, along with total power and the LF/HF ratio, are reported.

A.4.3 Further features

To investigate heart rate features more comprehensively, additional features from other categories based on prior work were included. Specifically, features were selected from Doyle et al. [?], who studied 62 HRV parameters, and from Jeppesen et al. [?], who examined and refined parameters derived from the Poincare plot. This broader set of features enables a more thorough characterization of heart rate dynamics and may improve the performance of subsequent analyses.

First, some useful time domain and Nonlinear metrics were added:

8. Root mean squared of NN intervals (RMSSD):

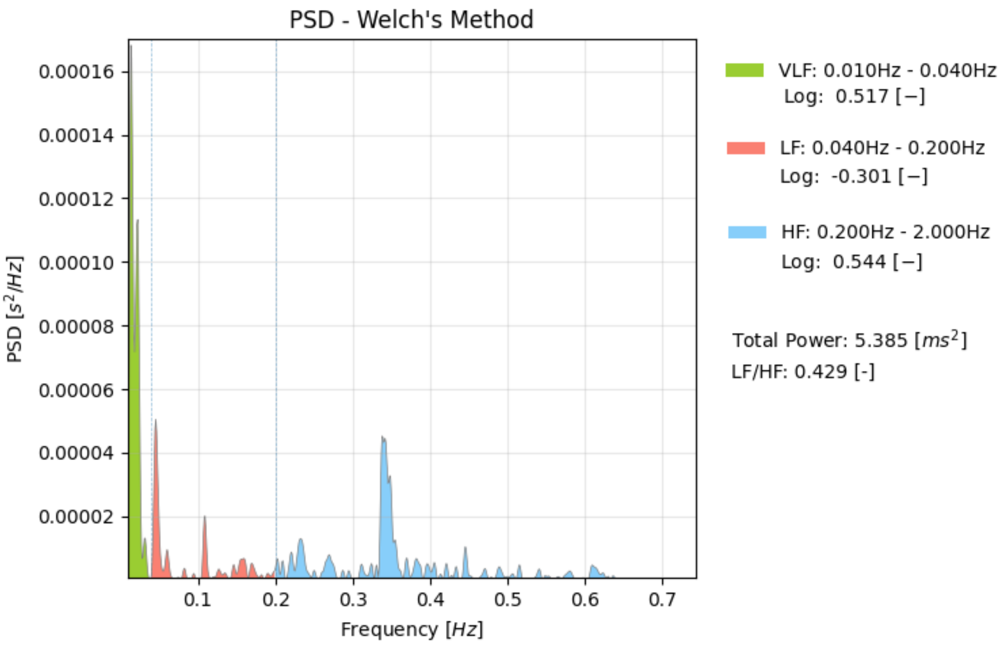


Figure A.7: Power spectral density plot generated for a two-minute ECG segment.

$$RMSSD = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n} \Delta NNI_j^2} \tag{A.5}$$

Where $\Delta NNI_j = NNI_j - NNI_{j-1}$.

9. Skewness is a measure of asymmetry which is computed as the Fisher-Pearson coefficient of skewness:

$$g_1 = \frac{m_3}{m_2^{\frac{3}{2}}} \tag{A.6}$$

where

$$m_i = \frac{1}{n}\sum_{j=1}^{n}(NNI_j - \overline{NNI})^i \tag{A.7}$$

is the biased sample ith central moment.

10. Kurtosis shows if the data is peaked or flat relative to the normal distribution.

$$kurtosis = \frac{\frac{1}{n}\sum_{j=1}^{n}(NNI_j - \overline{NNI})^4}{SDNN^4} \tag{A.8}$$

11. Approximate entropy (ApEn) quantifies regularity and complexity of stationary signals which shows the randomness of the NN intervals.

$$ApEn(NNI, r, n) = \Phi(NNI + 1, r, n) - \Phi(NNI, r, n) \tag{A.9}$$

    Where r is the threshold used to determine whether two NNIs are similar and Φ is the conditional probability that two NNIs that match within a tolerance of r will also match at the next point. r is set to 0.2 times the standard deviation of the HR signal segment.

Poincare plot or Lorenz plot analysis is a non-linear method to determine the correlation between two consecutive data points in a time series signal. Here, it is created from a series of NN intervals or peaks to calculate their beat-to-beat and overall changes.

12. SD1 parameter is the standard deviation of the NNI series along the minor axis of Poincare scatter plot, calculated using the Standard deviation of NNI differences (SDSD).

$$SDSD = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(\Delta NNI_j - \overline{\Delta NNI})^2} \tag{A.10}$$

$$SD1 = \sqrt{\frac{1}{2}SDSD^2} \tag{A.11}$$

13. SD2 parameter is the standard deviation of the NNI series along the major axis of Poincare scatterplot, calculated using the Standard Deviation of a NNI series (SDNN) which was explained in equation (A.2).

$$SD2 = \sqrt{2SDNN^2 - \frac{1}{2}SDSD^2} \tag{A.12}$$

14. The SD ratio which is the ratio between SD1 and SD2 :

$$SD_{ratio} = \frac{SD2}{SD1} \tag{A.13}$$

An example of a Poincare plot generated for an HR segment is illustrated in Fig. A.8.

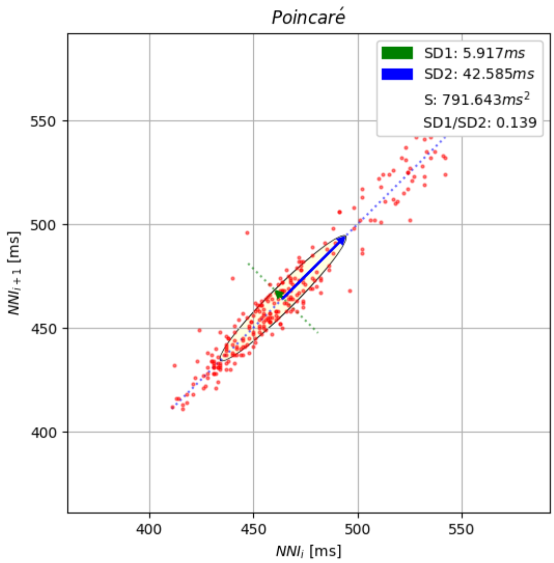


Figure A.8: Poincare plot generated for a two-minute ECG segment. Poincare plot of the NN-interval series, with NNI_i(ms) on the x-axis and NNI_{i+1}(ms) on the y-axis. The fitted ellipse illustrates short-term variability (SD1) and long-term variability (SD2), with the ellipse area S representing overall variability. The SD1/SD2 ratio characterizes the geometric shape of the plot.

Toichi et al. [?] described the transverse length (T) and the longitudinal length (L) as four times $SD1$ and $SD2$ respectively. Using these measures, two additional features can be derived, which have been shown to be effective for seizure detection by Jeppesen et al. [?].

15. Cardiac vagal index (CVI)

$$CVI = log_{10}(L \times T) \tag{A.14}$$

16. Modified cardiac sympathetic index (CSI)

$$CSI = \frac{L^2}{T} \tag{A.15}$$

A.5 Machine learning models

This section provides a detailed explanation of the classical machine learning models used in this thesis.

A.5.1 Random Forest

Decision tree learning is a supervised learning approach valued for its minimal data preparation requirements and its ability to inherently handle nonlinear relationships. It does not require feature transformation or scaling, and it makes predictions by repeatedly splitting the data on the feature and threshold that best separate the target classes, following those splits from the root to a leaf where the final decision is made.

Random Forest (RF) [?] is a tree-based ensemble learning technique that builds multiple decision trees during training and merges their predictions to improve

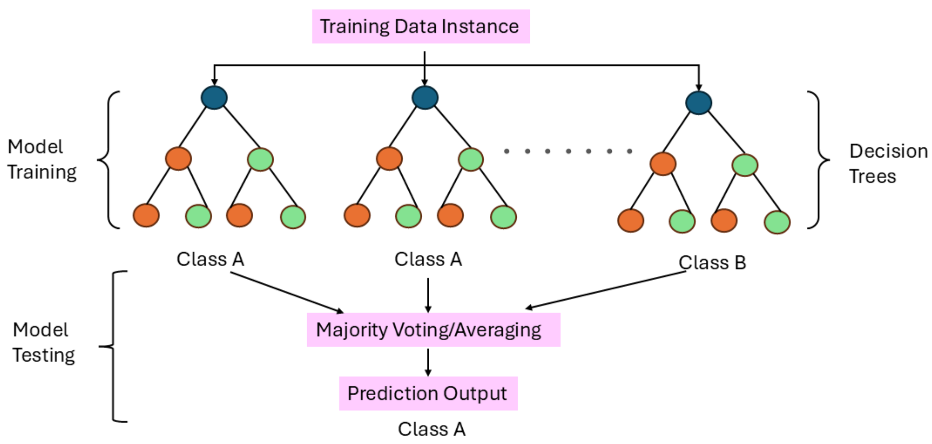


Figure A.9: Architecture of the Random Forest model.

accuracy and robustness. Each individual tree provides a vote for the predicted class of an input, and the model aggregates these votes to determine the final classification. In practical implementations, the algorithm constructs many trees with different random subsets of data and features, ensuring diversity among the trees. This approach reduces overfitting common in single decision trees and allows the model to generalize better on unseen data. Fig. A.9 shows the RF architecture.

A.5.2 Extreme Gradient Boosting

While in Random Forests the decision trees are trained in parallel, boosting algorithms take a sequential approach. Extreme Gradient Boosting (XGBoost) [**?**] is a powerful ensemble machine learning algorithm that builds decision trees sequentially using the gradient boosting framework. Each tree is trained to reduce the errors made by the previous ones by focusing on the differences between the predicted and actual values. This iterative process results in an accurate final model. In addition, the XGBoost python library is designed to be memory efficient comparing with random forest and SVM thus, making it more suitable for a large dateset training tasks. Fig. A.10 shows the XGBoost architecture.

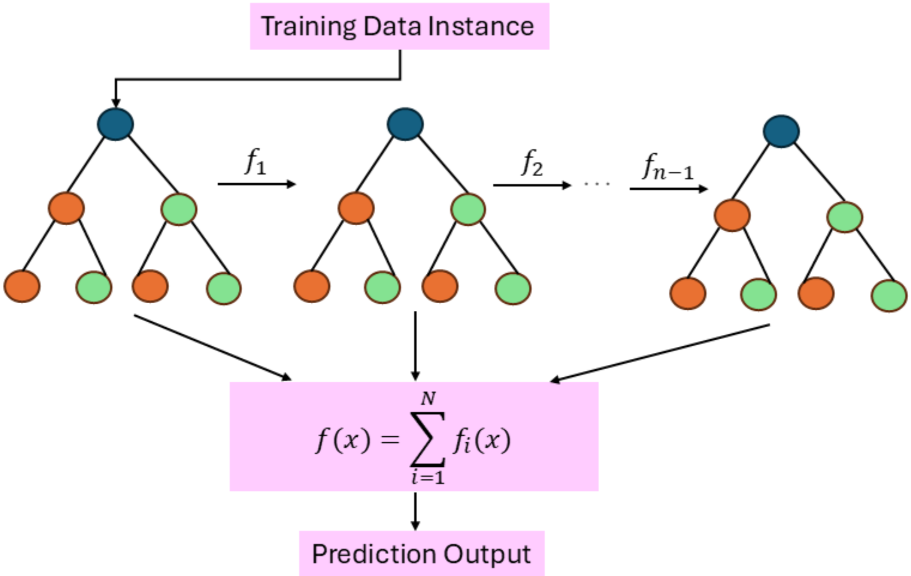


Figure A.10: Architecture of the XGBoost model.

A.5.3 Isolation Forest

Isolation Forest (IF) [?] is a tree-based unsupervised method specifically crafted for anomaly detection. It works by randomly selecting a feature and a split value between that feature's minimum and maximum in the data, recursively partitioning the dataset to isolate observations. Because anomalies are rare and differ significantly from normal instances, they require fewer splits to be isolated, resulting in shorter path lengths in the trees. This difference in path lengths allows the model to effectively distinguish anomalous points from normal ones. When implemented, the model builds a forest of such random trees and uses the average path length to assign an anomaly score to each observation.

Isolation Forest was used only for the seizure detection task in Section 3.9.1.2. In this case, the entire dataset was treated as unlabeled, and the goal was to evaluate whether Isolation Forest could identify seizure events as anomalies (outliers) within the data.

A.5.4 Support Vector Machine

A margin-based model, Support Vector Machine (SVM) [?], was also used to compare results with tree-based models, as SVMs offer a fundamentally different approach by finding the optimal separating hyperplane rather than relying on hierarchical feature splits.

SVM is a supervised learning algorithm designed to find the optimal boundary or hyperplane that best separates different classes in the feature space. The method focuses on maximizing the margin between the closest points of different classes (called support vectors) and the decision boundary. During training, the algorithm identifies these critical points and constructs a hyperplane that maximizes this margin, leading to strong generalization. In practice, SVM can handle both linear and nonlinear classification problems using different kernel functions, such as linear, polynomial, or radial basis function (RBF) kernels, which project data into higher-dimensional spaces when necessary. Fig. A.11 shows the SVM with a linear kernel function.

In this work, an RBF kernel is used, allowing the model to handle nonlinear relationships by projecting data into a higher-dimensional space.

A.5.5 Multilayer Perceptron

A deep learning approach, a Multi-Layer Perceptron (MLP) [?] was also used in this thesis. Deep learning models are capable of learning complex nonlinear relationships and automatically capturing feature interactions, providing a complementary perspective to the tree-based models and SVM.

MLP is a feed forward neural network consisting of fully connected neurons with nonlinear activation functions. The MLP algorithm trains using gradient descent

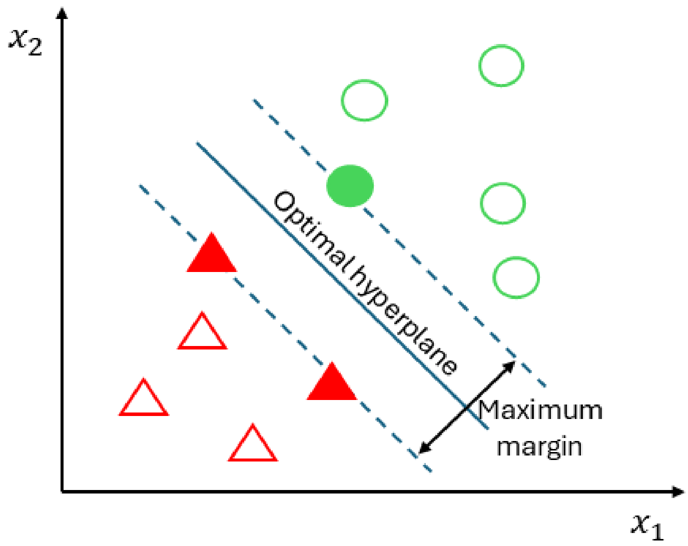


Figure A.11: Support Vector Machine classification showing the data points, the decision boundary, and the margins that define the optimal separating hyperplane.

and the gradients are calculated using Backpropagation to minimizes the cross-entropy loss function for classification task.

The MLP architecture used in this study comprises three fully connected layers. To introduce nonlinearity, the ReLU activation function is applied after each layer. The model accepts an input vector of 16 features, with the first layer containing 16 neurons. The second layer expands to 64 neurons to capture more complex patterns, followed by a third layer that reduces the dimensionality to 32 neurons. The final output layer consists of 2 neurons, representing the two target classes. This architecture is designed to strike a balance between model complexity and computational efficiency, making it well-suited for classification tasks involving 16 input features. The proposed MLP flowchart is presented in Fig. A.12.

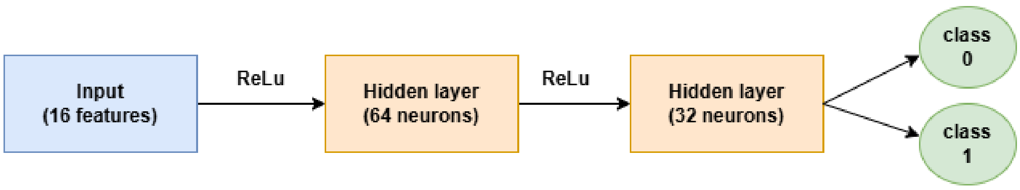


Figure A.12: Proposed MLP model flowchart.

Bibliography

[1] World Health Organization. Newborn health. `https://www.who.int/health-topics/newborn-health#tab=tab_2`, 2025. Accessed: 2025-09-01.

[2] P Greco, G Nencini, I Piva, M Scioscia, CA Volta, S Spadaro, M Neri, G Bonaccorsi, F Greco, I Cocco, and F Sorrentino. Pathophysiology of hypoxic–ischemic encephalopathy: a review of the past and a view on the future. *Acta Neurologica Belgica*, 120(2):277–288, 2020.

[3] Harvey B Sarnat and Margaret S Sarnat. Neonatal encephalopathy following fetal distress: a clinical and electroencephalographic study. *Archives of neurology*, 33(10):696–705, 1976.

[4] JM Conway, BH Walsh, GB Boylan, and DM Murray. Mild hypoxic ischaemic encephalopathy and long term neurodevelopmental outcome-a systematic review. *Early human development*, 120:80–87, 2018.

[5] Guido Wassink, Joanne O Davidson, Simerdeep K Dhillon, Kelly Zhou, Laura Bennet, Marianne Thoresen, and Alistair J Gunn. Therapeutic hypothermia in neonatal hypoxic-ischemic encephalopathy. *Current Neurology and Neuroscience Reports*, 19:1–10, 2019.

[6] Deirdre M Murray, Geraldine B Boylan, Imhad Ali, C Anthony Ryan, Brendan P Murphy, and Sean Connolly. Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 93(3):F187–F191, 2008.

[7] Enrico Facco. The role of eeg in brain injury. *Intensive care medicine*, 25(8):872, 1999.

[8] D Neubauer, D Osredkar, D Paro-Panjan, A Skofljanec, and M Derganc. Recording conventional and amplitude-integrated eeg in neonatal intensive care unit. *European Journal of Paediatric Neurology*, 15(5):405–416, 2011.

[9] Robert Michael David Goulding. *Heart rate variability and electroencephalography in Infants with hypoxic ischaemic encephalopathy*. PhD thesis, University College Cork, 2017.

[10] Andreea M Pavel, John M O'Toole, Jacopo Proietti, Vicki Livingstone, Subhabrata Mitra, William P Marnane, Mikael Finder, Eugene M Dempsey, Deirdre M Murray, and Geraldine B Boylan. Machine learning for the early prediction of infants with electrographic seizures in neonatal hypoxic-ischemic encephalopathy. *Epilepsia*, 64(2):456–468, 2023.

[11] Kelly Q Zhou, Alice McDouall, Paul P Drury, Christopher A Lear, Kenta HT Cho, Laura Bennet, Alistair J Gunn, and Joanne O Davidson. Treating seizures after hypoxic-ischemic encephalopathy—current controversies and future directions. *International Journal of Molecular Sciences*, 22(13):7121, 2021.

[12] James JP Alix, Athi Ponnusamy, Elizabeth Pilling, and Anthony R Hart. An introduction to neonatal eeg. *Paediatrics and Child Health*, 27(3):135–142, 2017.

[13] MD Lamblin and A de Villepin-Touzery. Eeg in the neonatal unit. *Neurophysiologie Clinique/Clinical Neurophysiology*, 45(1):87–95, 2015.

[14] I Korotchikova, NJ Stevenson, BH Walsh, DM Murray, and GB Boylan.

Quantitative eeg analysis in neonatal hypoxic ischaemic encephalopathy. *Clinical Neurophysiology*, 122(8):1671–1678, 2011.

[15] Deirdre M Murray, Geraldine B Boylan, Cornelius A Ryan, and Sean Connolly. Early eeg findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. *Pediatrics*, 124(3):e459–e467, 2009.

[16] Rehan Ahmed. *Dynamic classifiers for neonatal brain monitoring*. PhD thesis, University College Cork, 2016.

[17] Georgia Ramantani, Bernhard Schmitt, Barbara Plecko, Ronit M Pressler, Gabriele Wohlrab, Katrin Klebermass-Schrehof, Cornelia Hagmann, Francesco Pisani, and Geraldine B Boylan. Neonatal seizures—are we there yet? *Neuropediatrics*, 50(05):280–293, 2019.

[18] Alison O'Shea. *Neonatal seizure detection: a deep learning approach*. PhD thesis, University College Cork, 2020.

[19] Steven A Israel, John M Irvine, Andrew Cheng, Mark D Wiederhold, and Brenda K Wiederhold. Ecg to identify individuals. *Pattern recognition*, 38(1):133–142, 2005.

[20] Gari D Clifford, Francisco Azuaje, and Patrick McSharry. Ecg statistics, noise, artifacts, and missing data. *Advanced methods and tools for ECG data analysis*, 6(1):18, 2006.

[21] AC Vinzio Maggio, María Paula Bonomini, Eric Laciar Leber, and Pedro David Arini. Quantification of ventricular repolarization dispersion using digital processing of the surface ecg. *Advances in Electrocardiograms-Methods and Analysis*, 1:183–185, 2012.

[22] Denis V Azzopardi, Brenda Strohm, A David Edwards, Leigh Dyet, Henry L Halliday, Edmund Juszczak, Olga Kapellou, Malcolm Levene,

Neil Marlow, Emma Porter, and Marianne Thoresen. Moderate hypothermia to treat perinatal asphyxial encephalopathy. *New England Journal of Medicine*, 361(14):1349–1358, 2009.

[23] Jesper Jeppesen, Sándor Beniczky, Peter Johansen, Per Sidenius, and Anders Fuglsang-Frederiksen. Detection of epileptic seizures with a modified heart rate variability algorithm based on lorenz plot. *Seizure*, 24:1–7, 2015.

[24] Kaat Vandecasteele, Thomas De Cooman, Ying Gu, Evy Cleeren, Kasper Claes, Wim Van Paesschen, Sabine Van Huffel, and Borbála Hunyadi. Automated epileptic seizure detection based on wearable ecg and ppg in a hospital environment. *Sensors*, 17(10):2338, 2017.

[25] Jesper Jeppesen, Anders Fuglsang-Frederiksen, Peter Johansen, Jakob Christensen, Stephan Wüstenhagen, Hatice Tankisi, Erisela Qerama, Alexander Hess, and Sándor Beniczky. Seizure detection based on heart rate variability using a wearable electrocardiography device. *Epilepsia*, 60(10):2105–2113, 2019.

[26] Amirhossein Jahanbekam, Jan Baumann, Robert D Nass, Christian Bauckhage, Holger Hill, Christian E Elger, and Rainer Surges. Performance of ecg-based seizure detection algorithms strongly depends on training and test conditions. *Epilepsia Open*, 6(3):597–606, 2021.

[27] Saleh Baghersalimi, Alireza Amirshahi, Farnaz Forooghifar, Tomas Teijeiro, Amir Aminifar, and David Atienza. Many-to-one knowledge distillation of real-time epileptic seizure detection for low-power wearable internet of things systems. *arXiv preprint arXiv:2208.00885*, 2022.

[28] Iosif Mporas, Vasiliki Tsirka, Evangelia I Zacharaki, Michalis Koutroumanidis, Mark Richardson, and Vasileios Megalooikonomou.

Seizure detection using eeg and ecg signals for computer-based monitoring, analysis and management of epileptic patients. *Expert systems with applications*, 42(6):3227–3233, 2015.

[29] Miguel Bhagubai, Kaat Vandecasteele, Lauren Swinnen, Jaiver Macea, Christos Chatzichristos, Maarten De Vos, and Wim Van Paesschen. The power of ecg in semi-automated seizure detection in addition to two-channel behind-the-ear eeg. *Bioengineering*, 10(4):491, 2023.

[30] Marwa Qaraqe, Muhammad Ismail, Erchin Serpedin, and Haneef Zulfi. Epileptic seizure onset detection based on eeg and ecg data fusion. *Epilepsy & Behavior*, 58:48–60, 2016.

[31] Rehan Ahmed, Andrey Temko, William P Marnane, Geraldine Boylan, and Gordon Lightbody. Classification of hypoxic-ischemic encephalopathy using long term heart rate variability based features. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2355–2358. IEEE, 2015.

[32] Barry R Greene, Philip de Chazal, Geraldine B Boylan, Sen Connolly, and Richard B Reilly. Electrocardiogram based neonatal seizure detection. *IEEE Transactions on Biomedical Engineering*, 54(4):673–682, 2007.

[33] OM Doyle, A Temko, W Marnane, G Lightbody, and GB Boylan. Heart rate based automatic seizure detection in the newborn. *Medical Engineering & Physics*, 32(8):829–839, 2010.

[34] Benedetta Olmi, Claudia Manfredi, Lorenzo Frassineti, Carlo Dani, Silvia Lori, Giovanna Bertini, Cesarina Cossu, Maria Bastianelli, Simonetta Gabbanini, and Antonio Lanatà. Heart rate variability analysis for seizure detection in neonatal intensive care units. *Bioengineering*, 9(4):165, 2022.

[35] Hui Lu, Stefan Kusnik, Dilbar Mammadova, Regina Trollmann, and Alexander Koelpin. Hrv-based monitoring of neonatal seizures with machine learning. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE, 2024.

[36] Robert M Goulding, Nathan J Stevenson, Deirdre M Murray, Vicki Livingstone, Peter M Filan, and Geraldine B Boylan. Heart rate variability in hypoxic ischemic encephalopathy: correlation with eeg grade and 2-y neurodevelopmental outcome. *Pediatric Research*, 77(5):681–687, 2015.

[37] Vladimir Matić, Perumpillichira J Cherian, Devy Widjaja, Katrien Jansen, Gunnar Naulaers, Sabine Van Huffel, and Maarten De Vos. Heart rate variability in newborns with hypoxic brain injury. In *Oxygen Transport to Tissue XXXV*, pages 43–48. Springer, 2013.

[38] Laura Cabrera-Quiros, Gabriele Varisco, Zhuozhao Zhan, Xi Long, Peter Andriessen, Eduardus JE Cottaar, and Carola van Pul. Estimation of heart rate directly from ecg spectrogram in neonate intensive care units. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 320–323. IEEE, 2020.

[39] S Daphin Lilda and R Jayaparvathy. Enhancing cardiovascular disease classification in ecg spectrograms by using multi-branch cnn. *Computers in Biology and Medicine*, 186:109737, 2025.

[40] Jingshan Huang, Binqiang Chen, Bin Yao, and Wangpeng He. Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network. *IEEE Access*, 7:92871–92880, 2019.

[41] Yared Daniel Daydulo, Bheema Lingaiah Thamineni, and Ahmed Ali

Dawud. Cardiac arrhythmia detection using deep learning approach and time frequency representation of ecg signals. *BMC Medical Informatics and Decision Making*, 23(1):232, 2023.

[42] Ping Lu, Chenyang Wang, Jannis Hagenah, Shadi Ghiasi, Tingting Zhu, Louise Thwaites, and David A Clifton. Improving classification of tetanus severity for patients in low-middle income countries wearing ecg sensors by using a cnn-transformer network. *IEEE Transactions on Biomedical Engineering*, 70(4):1340–1350, 2022.

[43] Benyamin Haghi, Lin Ma, Sahin Lale, Anima Anandkumar, and Azita Emami. Ekgnet: A 10.96 μw fully analog neural network for intra-patient arrhythmia classification. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5. IEEE, 2023.

[44] Hany El-Ghaish and Emadeldeen Eldele. Ecgtransform: Empowering adaptive ecg arrhythmia classification framework with bidirectional transformer. *Biomedical Signal Processing and Control*, 89:105714, 2024.

[45] Emmanuel C Chukwu and Pedro A Moreno-Sanchez. Enhancing arrhythmia diagnosis with data-driven methods: a 12-lead ecg-based explainable ai model. In *Nordic Conference on Digital Health and Wireless Solutions*, pages 242–259. Springer, 2024.

[46] Hang Liu, Shaowei Cui, Xiaohui Zhao, and Fengyu Cong. Detection of obstructive sleep apnea from single-channel ecg signals using a cnn-transformer architecture. *Biomedical Signal Processing and Control*, 82:104581, 2023.

[47] Junheng Wang and Milos Hauskrecht. Augmentation-free contrastive learn-

ing for ekg classification. In *International Conference on Artificial Intelligence in Medicine*, pages 468–479. Springer, 2025.

[48] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 132–134, 2021.

[49] Andreea M Pavel, Janet M Rennie, Linda S de Vries, Mats Blennow, Adrienne Foran, Divyen K Shah, Ronit M Pressler, Olga Kapellou, Eugene M Dempsey, Sean R Mathieson, Elena Pavlidis, Alexander C van Huffelen, Vicki Livingstone, Mona C Toet, Lauren C Weeke, Mikael Finder, Subhabrata Mitra, Deirdre M Murray, William P Marnane, and Geraldine B Boylan. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. *The Lancet Child & Adolescent Health*, 4(10):740–749, 2020.

[50] Janet M Rennie, Linda S de Vries, Mats Blennow, Adrienne Foran, Divyen K Shah, Vicki Livingstone, Alexander C van Huffelen, Sean R Mathieson, Elena Pavlidis, Lauren C Weeke, and Mona C Toet. Characterisation of neonatal seizures and their treatment using continuous eeg monitoring: a multicentre experience. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 104(5):F493–F501, 2019.

[51] Rehan Ahmed, Andriy Temko, William Marnane, Gordon Lightbody, and Geraldine Boylan. Grading hypoxic-ischemic encephalopathy severity in neonatal eeg using gmm supervectors and the support vector machine. *Clinical Neurophysiology*, 127(1):297–309, 2016.

[52] Andreea M Pavel, Sean R Mathieson, Vicki Livingstone, John M O'Toole, Ronit M Pressler, Linda S de Vries, Janet M Rennie, Subhabrata Mitra, Eu-

gene M Dempsey, Deirdre M Murray, William P Marnane, and Gerladine B Boylan. Heart rate variability analysis for the prediction of eeg grade in infants with hypoxic ischaemic encephalopathy within the first 12 h of birth. *Frontiers in pediatrics*, 10:1016211, 2023.

[53] Andriy Temko, Eoin Thomas, William Marnane, Gordon Lightbody, and G Boylan. Eeg-based neonatal seizure detection with support vector machines. *Clinical Neurophysiology*, 122(3):464–473, 2011.

[54] Patrick S Hamilton and Willis J Tompkins. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE transactions on biomedical engineering*, (12):1157–1165, 2007.

[55] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 2007.

[56] Ivaylo I Christov. Real time electrocardiogram qrs detection using combined adaptive threshold. *Biomedical engineering online*, 3(1):28, 2004.

[57] Zahia Zidelmal, Ahmed Amirou, D Ould-Abdeslam, Ali Moukadem, and Alain Dieterlen. Qrs detection using s-transform and shannon energy. *Computer methods and programs in biomedicine*, 116(1):1–9, 2014.

[58] Jin Bo, Xuewen Cao, Yuqing Wan, Yuanyu Yu, Pun Sio Hang, Peng Un Mak, and Mang I Vai. Investigation performance on electrocardiogram signal processing based on an advanced algorithm combining wavelet packet transform (wpt) and hilbert-huang transform (hht). In *Frontier and Future Development of Information Technology in Medicine and Education: ITME 2013*, pages 959–968. Springer, 2013.

[59] Negar Farzaneh, Hamid Ghanbari, Mingzhu Liu, Loc Cao, Kevin R Ward, and Sardar Ansari. A comprehensive comparison of six publicly available

algorithms for localization of qrs complex on electrocardiograph. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[60] Feifei Liu, Chengyu Liu, Xinge Jiang, Zhimin Zhang, Yatao Zhang, Jianqing Li, and Shoushui Wei. Performance analysis of ten common qrs detectors on different ecg application cases. *Journal of healthcare engineering*, 2018(1):9050812, 2018.

[61] Mohamed Hammad, Asmaa Maher, Kuanquan Wang, Feng Jiang, and Moussa Amrani. Detection of abnormal heart conditions based on characteristics of ecg signals. *Measurement*, 125:634–644, 2018.

[62] Hongzu Li and Pierre Boulanger. Structural anomalies detection from electrocardiogram (ecg) with spectrogram and handcrafted features. *Sensors*, 22(7):2467, 2022.

[63] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[64] Gabriele Volpes, Chiara Barà, Alessandro Busacca, Salvatore Stivala, Michal Javorka, Luca Faes, and Riccardo Pernice. Feasibility of ultra-short-term analysis of heart rate and systolic arterial pressure variability at rest and during stress via time-domain and entropy-based measures. *Sensors*, 22(23):9149, 2022.

[65] Motomi Toichi, Takeshi Sugiura, Toshiya Murai, and Akira Sengoku. A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of r–r interval. *Journal of the Autonomic Nervous System*, 62(1-2):79–84, 1997.

[66] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip

Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[67] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[68] SEQUOIA HALL, ANDREW R BARRON, and XIANGYU XIAO. Parzen, e.(1962). on estimation of a probability density function and mode. ann. math. statist.

[69] Richard A Davis, Keh-Shin Lii, and Dimitris N Politis. Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt*, pages 95–100. Springer, 2011.

[70] Jiu Feng, Mehmet Hamza Erol, Joon Son Chung, and Arda Senocak. From coarse to fine: Efficient training for audio spectrogram transformers. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1416–1420, 2024.

[71] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[72] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77, 2006.

[73] Matthias Seibold, Armando Hoch, Mazda Farshad, Nassir Navab, and Philipp Fürnstahl. Conditional generative data augmentation for clinical audio datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 345–354. Springer, 2022.

[74] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101, 2021.

[75] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[76] Md Moklesur Rahman, Massimo Walter Rivolta, Fabio Badilini, and Roberto Sassi. A systematic survey of data augmentation of ecg signals for ai applications. *Sensors*, 23(11):5237, 2023.

[77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[78] Feifei Liu, Shoushui Wei, Yibin Li, Xinge Jiang, Zhimin Zhang, Ling Zhang, and Chengyu Liu. The accuracy on the common pan-tompkins based qrs detection methods through low-quality electrocardiogram database. *Journal of Medical Imaging and Health Informatics*, 7(5):1039–1043, 2017.

[79] Ziti Fariha Mohd Apandi, Ryojun Ikeura, Soichiro Hayakawa, and Shigeyoshi Tsutsumi. An analysis of the effects of noisy electrocardiogram signal on heartbeat detection performance. *Bioengineering*, 7(2):53, 2020.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[81] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.

[82] Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I Webb, and Mahsa

Salehi. Improving position encoding of transformers for multivariate time series classification. *Data Mining and Knowledge Discovery*, 38(1):22–48, 2024.

[83] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[84] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *ArXiv Preprint ArXiv:2005.00928*, 2020.

[85] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[86] Shuwen Yu, William P Marnane, Geraldine B Boylan, and Gordon Lightbody. Neonatal hypoxic-ischemic encephalopathy grading from multichannel eeg time-series data using a fully convolutional neural network. *Technologies*, 11(6):151, 2023.

[87] Carolina Varon, Katrien Jansen, Lieven Lagae, and Sabine Van Huffel. Can ecg monitoring identify seizures? *Journal of electrocardiology*, 48(6):1069–1074, 2015.

[88] John M O'Toole and Geraldine B Boylan. Quantitative preterm eeg analysis: the need for caution in using modern data science techniques. *Frontiers in pediatrics*, 7:174, 2019.

[89] Alexandros T Tzallas, Markos G Tsipouras, and Dimitrios I Fotiadis. Epileptic seizure detection in eegs using time–frequency analysis. *IEEE transactions on information technology in biomedicine*, 13(5):703–710, 2009.

[90] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and

diagnosis of seizure using eeg signals. *Computers in biology and medicine*, 100:270–278, 2018.

[91] Anubha Gupta, Pushpendra Singh, and Mandar Karlekar. A novel signal modeling approach for classification of seizure and seizure-free eeg signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(5):925–935, 2018.

[92] Hepseeba Kode, Khaled Elleithy, and Laiali Almazaydeh. Epileptic seizure detection in eeg signals using machine learning and deep learning techniques. *IEEE Access*, 12:80657–80668, 2024.

[93] Franklyn Rocha Cabrero and Orlando De Jesus. Eeg neonatal visual analysis. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

[94] NJ Stevenson, I Korotchikova, A Temko, G Lightbody, WP Marnane, and GB Boylan. An automated system for grading eeg abnormality in term neonates with hypoxic-ischaemic encephalopathy. *Annals of biomedical engineering*, 41(4):775–785, 2013.

[95] Sumit A Raurale, Geraldine B Boylan, Sean R Mathieson, William P Marnane, Gordon Lightbody, and John M O'Toole. Grading hypoxic-ischemic encephalopathy in neonatal eeg with convolutional neural networks and quadratic time–frequency distributions. *Journal of neural engineering*, 18(4):046007, 2021.

[96] Mark E O'Sullivan, Gordon Lightbody, Sean R Mathieson, William P Marnane, Geraldine B Boylan, and John M O'Toole. Development of an eeg artefact detection algorithm and its application in grading neonatal hypoxic-ischemic encephalopathy. *Expert Systems with Applications*, 213:118917, 2023.

[97] Ardalan Aarabi, Reinhard Grebe, and Fabrice Wallois. A multistage knowledge-based system for eeg seizure detection in newborn infants. *Clinical Neurophysiology*, 118(12):2781–2797, 2007.

[98] Amir H Ansari, Perumpillichira J Cherian, Alexander Caicedo, Gunnar Naulaers, Maarten De Vos, and Sabine Van Huffel. Neonatal seizure detection using deep convolutional neural networks. *International journal of neural systems*, 29(04):1850011, 2019.

[99] Alison O'Shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture. *Neural Networks*, 123:12–25, 2020.

[100] Rehan Ahmed, Andriy Temko, William P Marnane, Geraldine Boylan, and Gordon Lightbody. Exploring temporal information in neonatal seizures using a dynamic time warping based svm kernel. *Computers in biology and medicine*, 82:100–110, 2017.

[101] Karoliina T Tapani, Sampsa Vanhatalo, and Nathan J Stevenson. Time-varying eeg correlations improve automated neonatal seizure detection. *International journal of neural systems*, 29(04):1850030, 2019.

[102] MAZ Fariha, Ryojun Ikeura, Soichiro Hayakawa, and Shigeyoshi Tsutsumi. Analysis of pan-tompkins algorithm performance with noisy ecg signals. In *Journal of Physics: Conference Series*, volume 1532, page 012022. IOP Publishing, 2020.

[103] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

[104] Peter Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[105] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[106] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[107] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[108] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[109] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.