UNIVERSITY OF ICELAND

Towards Clinically Useful Neonatal Seizure Detection Algorithms

Ana Borovac

2024

FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Towards Clinically Useful Neonatal Seizure Detection Algorithms

Ana Borovac

Dissertation submitted in partial fulfillment of a
Philosophiae Doctor degree in Computer Science

Supervisor
Prof. Steinn Gudmundsson

Doctoral Committee
Prof. Steinn Gudmundsson
Prof. Thomas P. Runarsson
Prof. Sampsa Vanhatalo

Opponents
Asst. Lect. Alison O'Shea
Prof. Maarten De Vos

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, January 2024

Towards Clinically Useful Neonatal Seizure Detection Algorithms
(Towards Clinically Useful NSDAs)

Dissertation submitted in partial fulfillment of a *Philosophiae Doctor* degree in Computer Science

Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
School of Engineering and Natural Sciences
University of Iceland
Hjarðarhagi 2-6
107 Reykjavik
Iceland

Telephone: 525-4000

Bibliographic information:
Ana Borovac (2024) *Towards Clinically Useful Neonatal Seizure Detection Algorithms*, PhD dissertation, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, 118 pp.

ISBN 978-9935-9742-8-0

Printing: Haskolaprent ehf.
Reykjavik, Iceland, January 2024

Abstract

Seizures are the most common neurological emergency among neonates and if left untreated they can cause permanent brain damage. The current gold standard for neonatal seizure detection is continuous electroencephalogram (EEG) visually interpreted by a human expert. Such expertise is rarely available round-the-clock and reliable automatic seizure detection could fill the gap. The detectors have been in development for about three decades and they are approaching human-level classification performance. In this work, we analyse and improve a detector based on deep learning. First, we study how to utilise the rather small data sets normally available for the development of detectors. In this case, we found that it is important to use the data for which multiple experts agreed on the seizure annotation. The size of the data sets could be increased if data from multiple institutions would be combined. To share the data information safely without breaking data-sharing policies, we propose to use an ensemble of locally developed seizure detectors. Second, we analyse the classification performance of a detector for a reduced number of input EEG signals and results indicate the performance is acceptable even when the number of input signals is small. The detector is further improved by utilizing calibration methods to be able to inform the user about uncertain predictions. Last, we gather user experience with a commercial seizure detector by conducting interviews with nurses. A common finding was that automatic seizure detections need to always be verified, but the nurses still find the detector useful despite many falsely detected seizures.

Útdráttur

Flog eru tiltölulega algeng meðal nýbura. Mikilvægt er að greina þau snemma því ómeðhöndluð flog geta leitt til varanlegs heilaskaða. Nákvæm greining á flogum byggir á lestri heilarita en það krefst sérfræðiþekkingar. Slík þekking er ekki alltaf til staðar og í þeim tilfellum geta sjálfvirkar greiningaraðferðir komið að góðum notum. Rannsóknir á sjálfvirkum aðferðum til að greina flog spanna rúma þrjá áratugi en á síðustu árum hafa komið fram aðferðir með greiningarnákvæmni sem stendur sérfræðingum ekki langt að baki. Í þessu verkefni var unnið með greiningaraðferð sem byggir á djúpum tauganetum, hún skoðuð ítarlega og síðan endurbætt. Til að hægt sé að þróa aðferðir sem byggja á tauganetum þarf talsvert magn af gögnum sem búið er að greina en í tilfelli nýburafloga eru gögn almennt af skornum skammti. Í verkefninu var skoðað hvernig mætti nýta lítil gagnasöfn sem best. Í ljós kom að mikilvægt er að styðjast við greiningar sem fleiri en einn sérfræðingur sammælast um. Vegna persónuverndarlaga er ekki einfalt að safna gögnum frá mörgum sjúkrastofnunum á einn stað til að mynda eitt stórt gagnasafn. Því var þróuð aðferð sem gerir stofnunum kleift að þjálfa tauganet á eigin gögnum og sameina svo líkönin þannig að persónuupplýsingar séu ekki í hættu. Nákvæmni greiningaraðferðinnar var skoðuð með hliðsjón af því að flest heilarit innihalda tiltölulega fáar rásir. Í ljós kom að ásættanleg nákvæmni fæst í þessum tilvikum. Æskilegt er að sjálfvirkar aðferðir gefi notendum til kynna þegar óvissa er í greiningu og því var unnið að því að bæta kvörðun aðferðarinnar. Að lokum var gerð könnun meðal hjúkrunarfólks á nýburugjörgæslu á notkun á vöktunarbúnaði fyrir flog en þessi búnaður er í almennri noktun á mörgum sjúkrahúsum. Niðurstöður sýndu að þrátt fyrir margar falskar viðvaranir, sé kerfið almennt gagnlegt, að því gefnu að allar viðvaranir séu skoðaðar sérstaklega.

Contents

List of Figures

List of Tables

List of Original Publications

Paper I: **Ana Borovac**, Steinn Gudmundsson, Gardar Thorvardsson and Thomas P. Runarsson. “Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network.” In: *NeurIPS Data-Centric workshop*. 2021.

Paper II: **Ana Borovac**, Steinn Gudmundsson, Gardar Thorvardsson, Saeed M. Moghadam, Päivi Nevalainen, Nathan Stevenson, Sampsa Vanhatalo and Thomas P. Runarsson. “Ensemble learning using individual neonatal data for seizure detection.” In: *IEEE Journal of Translational Engineering in Health and Medicine* 10 (2022). DOI: 10.1109/JIEHM.2022.3201167.

Paper III: **Ana Borovac**, Thomas P. Runarsson, Gardar Thorvardsson and Steinn Gudmundsson. “Neonatal seizure detection algorithms: The effect of channel count.” In: *Current Directions in Biomedical Engineering* 8.2 (2022), pp. 604-607. DOI: 10.1515/cdbme-2022-1154.

Paper IV: **Ana Borovac**, Thomas P. Runarsson, Gardar Thorvardsson and Steinn Gudmundsson. “Calibration of automatic seizure detection algorithms.” In: *2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. 2022. DOI: 10.1109/SPMB55497.2022.10014868.

Paper V: **Ana Borovac**, David H. Agustsson, Thomas P. Runarsson, and Steinn Gudmundsson. “Calibration methods for automatic seizure detection algorithms.” Accepted for publication.

Paper VI: Xiaowan Wang, **Ana Borovac**, Agnes van den Hoogena, Maria L. Tataranno, Manon J. N. L. Benders, and Jeroen Dudink. “Nurses’ experiences and perspectives on (a)EEG monitoring in neonatal care: A qualitative study.” In: *Journal of Neonatal Nursing* (2023). DOI: 10.1016/j.jnn.2023.08.003.

Abbreviations

ACC	accuracy
aEEG	amplitude integrated electroencephalogram
API	application programming interface
AUC	area under the curve
CNN	convolutional neural network
CONF	confidence
CFR	code of federal regulations
CUS	cranical ultrasound
DNN	deep neural network
ECE	expected calibration error
ECG	electrocardiogram
EEG	electroencephalogram
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
gRPC	Google remote procedure call
HIPAA	Health Insurance Portability and Accountability Act
INFANS	INtegrating Functional Assessment measures for Neonatal Safeguard
IP	internet protocol
HIE	hypoxic-ischemic encephalopathy
MDR	medical device regulations
ML	machine learning
MRI	magnetic resonance imaging
NICU	neonatal intensive care unit
NIRS	near-infrared spectroscopy

NSDA	neonatal seizure detection algorithm
OE	overconfidence error
ReLU	rectified linear unit
SCE	static calibration error
SE	sensitivity
SP	specificity
SDA	seizure detection algorithm
SVM	support vector machine
TCP	temporal central parasagittal
UMCU	University Medical Centre Utrecht
WKZ	Wilhelmina Children’s Hospital

Acknowledgments

I extend my sincere appreciation to Steinn Gudmundsson, Thomas P. Runarsson and Gardar Thorvardsson, for their guidance and support throughout the course of this doctoral study. Your theoretical and practical expertise in artificial intelligence applied to problems in healthcare has been a significant asset to this study. Your constructive feedback and thoughtful insights have played a crucial role in shaping the research methodology and outcomes. I would also like to thank Sampsa Vanhatalo for providing clinical expertise and experience whenever needed.

The first three years of the research were possible and funded by the European Union's Horizon 2020 research and innovation programme. The University of Iceland is also thanked for providing funds for the last, fourth, year of the studies in the form of a Doctoral teaching assistant grant. Moreover, the university offered support through the School of Engineering and Natural Sciences student service and the entire community at the Faculty of Industrial Engineering, Mechanical Engineering and Computer Science. Kvikna Medical ehf. is thanked for providing the computer and the required hardware for the project. I would like to express my gratitude to Samuel Thomas for technical support and Heidar Einarsson for the development of the software.

As the project was a part of the Marie-Curie training network, the project provided additional knowledge transfer through secondments. I would like to thank all the hosting institutions; the University of Helsinki and BABA Center are thanked for the clinical requirements related to neonatal continuous brain monitoring, eemagine Medical Imaging Solutions GmbH is thanked for broadening my knowledge with hardware development for neonatal brain monitoring, and University Medical Center Utrecht (UMCU) for giving me the in-clinical experience. Through secondments and other activities organised by the project, I was able to meet like-minded PhD students, and I am grateful that some of them became my friends.

I wish to acknowledge all collaborators; Nathan Stevenson for sharing the code, Päivi Nevalainen for providing some additional data, Saeed M. Moghadam for running the experiments and the help with a schematic representation of the methodology, David H. Agustsson for running the experiments and analysis of the results, Xiaowan Wang for collaborating on all stages of a study, Agnes van den Hoogena for assistance with qualitative study approach, Maria L. Tataranno, Manon J. N. L. Benders and Jeroen Dudink for clinical input and supervision of a study. All collaborators are also thanked for their assistance with writing and revising manuscripts.

Lastly, I extend heartfelt gratitude to my friends in Slovenia, Iceland, and other parts of the world, as well as to Matic and my family, especially mom, Barbara and dad for their support throughout my academic journey. It truly would not be possible without you.

1 Introduction

This research is a part of INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS) project funded by the Marie Skłodowska-Curie Actions programme.[1] The overreaching aim of the INFANS project is to improve neonatal brain monitoring with an emphasis on translating research progress into clinical practice.

Neonatal mortality rate has significantly decreased in past decades [51] and there is now a growing interest in improving the long-term outcome of every neonate with brain-focused care in the first month after birth [18]. Magnetic resonance imaging (MRI) and less invasive cranical ultrasound (CUS) are used in neonatal intensive care units (NICUs) for assessing long-term outcomes of neonates, however, these techniques do not allow continuous brain monitoring that is crucial for preventing brain injury and reducing neurological impairment later in life [112]. Currently, electroencephalogram (EEG) and near-infrared spectroscopy (NIRS) are the only non-invasive techniques allowing continuous brain monitoring and giving insights into the present brain condition and its development (*maturation*) [88, 102]. EEG is an established technique and is also recommended by international organisations in certain cases [125]. On the other hand, more research is required to confirm the added value of the NIRS monitoring [28, 126].

Maintaining hours and even days long EEG recordings in clinical practice can be difficult due to handling of the baby (e.g. diaper change and feeding) and movements of the baby, which can dislocate EEG electrodes. The signal quality is therefore often reduced and contaminated with artefacts [119]. For these reasons, there is an urgent need for the development of novel sensors comfortable for the patient and sensors that can provide good quality signals for an extensive period of time.

Since EEG recordings are long in duration, reviewing takes time [11] and it needs special expertise, neither of which are sufficiently available for the bedside clinicians in most places of the world [94]. Having automatic tools in clinical practice can be a valuable asset in making accurate patient-specific diagnoses [112] and can also promote more widespread use of continuous brain monitoring where is currently not in use due to lack of expertise.

This project focuses on the development of reliable machine learning (ML) tools for automatic analysis of EEG, specifically neonatal seizure detection algorithm (NSDA) based on deep neural networks (DNNs). Such tools should provide performance comparable to human experts and should alert the user if the tools are not entirely certain about their analysis, in order to prevent potential errors. Additionally, in collaboration with Kvikna Medical ehf. we designed and implemented a system allowing fast and easy integration of analytical tools into the Stratus EEG monitoring system (Kvikna Medical ehf., Iceland).

[1] Grant agreement ID: 813483

1.1 Motivation

Seizures, with an estimated prevalence of 1 – 5 per 1000 live births, are the most common neurological emergency amongst neonates. The prevalence rises to approximately 10 times higher values for neonates born *preterm* (before full 37 weeks of gestation) [127, 136]. Neonatal seizures are usually *acute provoked*, i.e. have an underlying cause. The most common causes are hypoxic-ischemic encephalopathy (HIE) [90], a brain injury caused by a decreased level of oxygen in the brain [6], and intraventricular haemorrhage, bleeding into the areas of the brain with cerebral spinal fluid, for preterm neonates [136]. The cause of seizures is also a dominant factor for short- and long-term outcomes of the neonate experiencing seizures. About half of the neonates affected by seizures reportedly have permanent brain damage associated with neurological, cognitive and motor impairments later in their lives. The mortality rate of affected neonates remains high and is estimated to be around 30 % [121].

Most neonatal seizures are only seen in the neuronal activity monitored with EEG [9], hence called *subclinical*, or *only electrographic*, whereas the sick newborn infants may have a large repertoire of motor behaviours that are easily mistaken for seizures. Relying only on visual observation of the neonate can therefore lead to an over- or under-diagnosis of seizures [70]. Currently, a continuous video EEG observed by a human expert is the gold standard of neonatal seizure detection [89]. EEG is recorded by using 2 – 19 scalp or needle electrodes positioned according to a standardized 10 – 20 system modified for neonates and their small heads [94, 99, 113]. Signals are later visualized as bipolar derivations (*channels*) that together form a *montage*. A schematic representation of EEG acquisition is in figure 1.1.

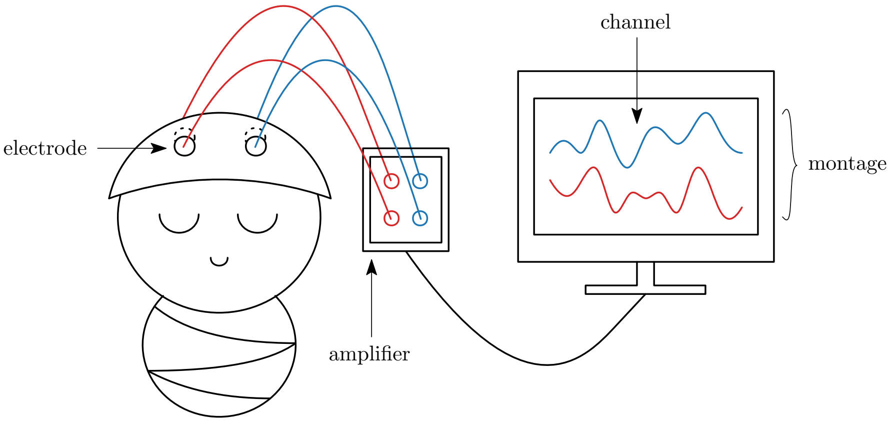


Figure 1.1. Schematic representation of EEG acquisition. Four electrodes (two in front and two at the back of the head) are attached to the scalp through a specialized cap. Electrodes are connected to an amplifier that amplifies the input signals and converts them from analogous to digital format. The amplifier is connected to a computer where the signals are visualized and processed. For example, blue (red) denotes a channel from the signals on the left (right) side of the head. Both channels together form a montage.

A part of EEG recording is identified as a seizure, if it lasts at least 10 s and has distinguishable beginning, middle, and end phases, with sudden and repetitive waveforms that evolve over time [15]. Most of the neonatal seizures are *focal* (*partial*) and can be observed just on a few EEG channels [74]. Annotations of neonatal seizures are however somewhat ambiguous even for human experts with years of experience [108] and precise temporal and spatial annotations are hardly possible to obtain.

In order to simplify and expedite the review of long EEG recordings that can span from multiple hours to multiple days, the EEG signals can be filtered, time-compressed and viewed on a semi-logarithmic scale, a method called amplitude integrated electroencephalogram (aEEG) [110]. However, analysing aEEG trends alone leads to falsely detected seizures due to artefacts or missed short seizures due to time compression [21].

Since not all NICUs have dedicated EEG experts available at the bedside all the time [12, 94] and annotating long EEG recordings is time-consuming, a lot of effort has gone into developing automatic NSDAs. Using automatic detection would help with prompt seizure detection and treatment that is crucial for preventing severe brain damage [8].

1.2 Background

Since the early 1990s, there has been a significant effort put into the development of accurate NSDAs [86, 114]. Initially, neonatal detectors were based on seizure detectors for adults and computed features from EEG. These features most often captured the periodic and high amplitude nature of seizures [14, 34, 69, 75, 77]. In [100] they extract features capturing the level of synchronisation between the channels since it is frequently increased during seizures. After extracting the features, human-defined thresholds were used to determine whether a particular EEG segment corresponded to seizure.

The next generation of NSDAs replaced human-defined thresholds with traditional ML approaches, e.g. linear discriminant analysis [37], support vector machines (SVMs) [1, 3, 111, 115] and neural networks [44, 131]. In these detectors, features were still hand-crafted, but the decision boundary between seizure and non-seizure classes was determined based on available data and not the human experience.

At the moment, DNNs are the state-of-the-art in automatic neonatal seizure detection [86]. By utilizing DNNs, features are not hand-crafted but instead *learned* from available data. During this process, DNNs learn the boundary between seizure and non-seizure activity. However, the boundary may also be determined by traditional ML methods, e.g. random forests [4] and SVMs [105].

The most frequently used type of DNNs for automatic seizure detection are convolutional neural networks (CNNs) [86]. They were first used for recognition of handwritten digits [68] and became popular for image classification about a decade ago [40]. The CNNs were inspired by the visual cortex of cats [50] and the idea is to learn small patterns (e.g. edges) that are combined into more complex objects with increased depth (number of layers) of the network. Methods for recognizing patterns in images were then adapted to pattern recognition in time series as are also neonatal EEG signals. In addition to applications for neonatal seizure detection, CNNs have been used for sleep stage classification [2, 73], EEG artefact detection [47, 82, 133] and classification of background activity [72] including estimation of HIE severity [93]. In [35, 84] fully convolutional networks were used for neonatal seizure detection and in [16] an improved version with residual connections was used. Residual connections make the development of networks with a large number of layers easier and often improve the accuracy of the classifier [46]. Since most neonatal seizures are focal and seen only on a subset of channels [74], a CNN was combined with an attention layer in [56] and with graph attention networks in [91]. By employing the attention mechanism any number of channels can be handled and annotations of specific channels containing seizure activity are not required. Moreover, this approach allows visualizing the location of seizures, giving the user additional information about the specific EEG channels that contributed to seizure detection. In [92] graph convolutional network is used to enhance the spatial resolution within the detector. This approach accounts for the varying proximity between different channels, which may be of particular value when dealing with a larger number of channels (e.g. 18 channels). In such cases, seizure activity is likely seen across several neighbouring channels.

In [13] a transfer learning approach was used to address the lack of available neonatal EEG data with seizure annotations. Features were extracted by a CNN developed on a large amount of image data and the decision boundary was determined based on available EEG data. It was assumed that even if the CNN used non-EEG data during the development, it was still able to extract relevant patterns for seizure detection.

1.3 Challenges

Evaluation of commercially available NSDAs and seizure detection algorithms (SDAs) for adults in clinical environments shows a discrepancy between reported and *real* seizure detection performances [61, 62, 98]. A decrease in performance in a clinical setting may partly be explained by the relatively small data sets used for the evaluation. High inter-patient variability present in the EEG signals [29] makes it difficult for small data sets to capture the full spectrum of EEG patterns. The same source of data (e.g. hospital) contributes to monotonicity in the data sets since different hospitals use different recording equipment and setups. The used data is also sometimes carefully selected and clean of artefacts which is far from what is encountered in the clinical environment. The reported performance values in the literature, therefore, apply for a subset of EEG recordings and the same performance is not guaranteed on all EEG signals [22, 135]. More attention needs to be given to these challenges and consequently making detectors with human-level performance. The challenges are split into three groups below.

1.3.1 Limited neonatal data with high-quality annotations

Development of NSDAs requires data with high-quality annotations, i.e. markings of precise time intervals with seizure activity. However, in reality, the availability of such data is often limited. First, annotating many long EEG recordings is excessively time-consuming [11] and requires special effort from human experts as precise temporal annotations of each seizure are not part of routine EEG analysis. Second, obtaining high-quality annotations requires review from multiple human experts due to the ambiguity of neonatal EEG patterns [108] which further compounds the expense of obtaining high-quality annotations. Third, seizure activity can represent less than 10 % of EEG recordings with seizures [106] and additionally limits the amount of seizure data. And, finally, neonatal EEG is a physiological signal and is therefore a subject of strict privacy regulations which prevents data sharing between institutions and building large and diverse data sets [20].

Hence, when designing and developing NSDAs based on a DNN one needs to keep in mind that a large amount of data [67] is often unavailable and the data should be used in a well-considered manner. Special attention also needs to be given to the imbalance present in the seizure EEG data in order to limit biases towards non-seizure predictions [57]. If multiple institutions are collaborating, it is necessary to select a training method that leverages the data available at each location but does not require sharing data in order to meet privacy regulations.

1.3.2 Different recording settings

NICUs use different hardware (e.g. electrodes and amplifiers) and software for EEG acquisition. These differences alone can affect the signal quality and consequently affect the automatic analysis. The use of equipment also differs depending on the internal protocols [64, 94]. The number of channels can vary between the NICUs and is dependent on the purpose of EEG monitoring. In case a patient needs long-term monitoring, fewer channels simplify the acquisition. For this type of monitoring, subcutaneous needle electrodes are also a more suited choice in comparison with gel electrodes. The quality of signals recorded with the gel electrodes decreases over time since the conductive gel dries out [27]. Furthermore, repetitive gel application may irritate the vulnerable neonatal skin [117]. On the other hand, if abnormalities are spotted during the continuous EEG monitoring and extra information is required from a wider area of the brain, more channels can be added. Such recordings are normally short in duration and gel electrodes may be used. The choice of electrodes together with a human factor also enhances the differences in inter-electrode distances despite the use of the standardised electrode position system. For example, in case the subcutaneous needle electrodes are applied and pointed towards each other, the distance between the electrodes is smaller for double the needle length in comparison with gel electrodes located at the same location where the skin holes were made for the needle electrodes. These differences may be small but are proportionately large.

So, if our aim is to develop an NSDA applicable in different settings, it needs to be able to accurately process signals from different types of recording hardware and a variable number of EEG channels. Different recording protocols may also produce signals of different quality and if they differ from the ones used in the development of the NSDA, the detector should be able to inform the user that the detections are unreliable.

1.3.3 Deployment into clinical practice

In recent years, there has been substantial progress in development towards reliable ML tools in healthcare applications. However, only a minority of the proposed solutions end up in clinical practice even though it has been shown that such automatic tools can perform well in collaboration with a human user [58, 71]. Deploying and using automatic tools in healthcare institutions is challenging [96, 137] since mistakes are very costly. ML tools used during the diagnosis of patients are defined as *medical devices* and need to be designed and evaluated as such. Medical devices are subject to strict regulations that vary by country. In the European Union, these are the new medical device regulations (MDR) and in the United States the code of federal regulations (CFR) 21 subchapter H set by the Food and Drug Administration (FDA). After the initial approval, medical device manufacturers are audited on a regular basis. During the audit, the design history file for medical devices is inspected to confirm that product improvement is properly managed and documented. This includes verification and validation data as well as other technical documentation. Additionally, the ML tools must meet the requirements of the medical staff since they are the ones making diagnostic and treatment decisions. In order for users to find the automatic tools useful, they need to trust them [120].

Reported accuracies of the NSDAs in recent literature suggest clinical usefulness [86], but evaluation processes can vary substantially. Not only the validation data but also the human expert annotating the data have an influence on the results since the labels are subjective to each scorer [108]. It is also important to choose clinically relevant performance metrics [116, 122]. To obtain the evaluation in the clinical setting the detectors should be implemented in the existing EEG monitors. Besides raw performance evaluation, this would give insights into how the medical staff (e.g. nurses and neonatologists) can make the best use of the automatic detectors and what are the gains in terms of time and accuracy.

1.4 Objectives

In this project, we focused on an NSDA based on a DNN since this type of detector is currently state-of-the-art in automatic neonatal seizure detection [86]. The overreaching goal was to develop a *reliable* detector that is useful in clinical practice. To reach this goal, three research questions were defined to address the challenges described in section 1.3:

1. How to train a clinically useful NSDA based on a DNN with limited neonatal EEG data?
2. How to design, train and evaluate an NSDA based on a DNN that is applicable in wide variety of NICU recording settings?
3. To what extent does the medical staff find an existing NSDA useful in clinical practice?

These questions are addressed in three conference papers, two journal papers and one book chapter (figure 1.2).

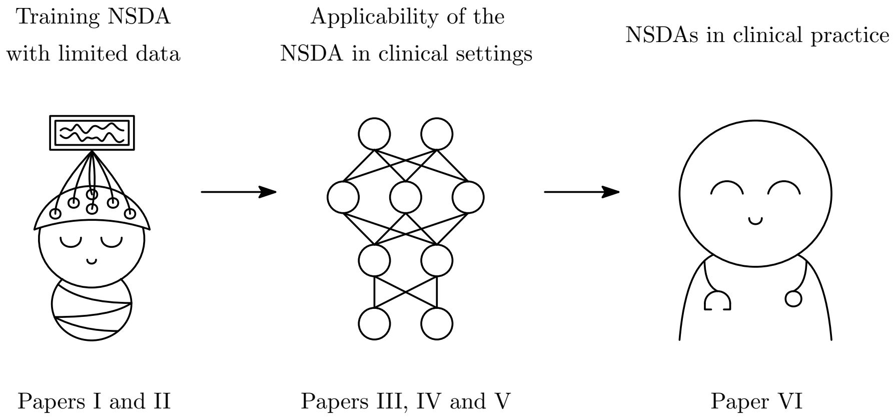


Figure 1.2. An overview of the research project with associated challenges and papers addressing them.

In papers I and II limited availability of neonatal EEG data with high-quality seizure annotations is addressed. We show the importance of high-quality annotations for the development of an accurate NSDA in paper I. The most accurate detector is obtained with parts of EEG recordings on which several human experts agreed on the annotation. This approach is often prohibitively time-consuming and we experimented with an approach that automatically detects incorrect annotations. However, the NSDA trained only on EEG the selected segments with presumably correct annotation does not result in a better-performing seizure detector. In paper II an ensemble of NSDAs developed using only a small amount of data is investigated. We show that an ensemble can perform equally to the NSDA developed utilizing all small data sets together.

Applicability of the NSDA in any recording setting is addressed in papers III, IV and V. In paper III we show that the NSDA performance is only slightly affected by the reduced number of channels and there are just certain types of artefacts causing problems. Careful analysis of the NSDA showed that there are EEG recordings for which the automatic detection fails. In papers IV and V we show that by applying calibration techniques, we can spot these recordings and inform the user that the detector is not completely reliable. By doing this, the detector could be integrated safely and effectively across different clinical settings. We additionally pursued the idea of adjusting the NSDA to each patient, making it patient-specific and accurate in any setting. We show that the current architecture of the detector is inappropriate for transfer learning and patient-specific NSDA requires a different detector and/or different adjusting approach.

The final research question is addressed in paper VI. In this work, we investigate how nurses use a continuous EEG monitoring system in a NICU and what are their concerns regarding it. As part of the study, we also interviewed them about their experience with the commercially available NSDA used in their NICU. Despite many falsely detected seizures, nurses use the detector on a daily basis and find it useful to find points of interest in long EEG recordings.

1.5 Thesis structure

The thesis consists of four parts. In section 2 the neonatal EEG data set and its preprocessing are described. This data set is used for most part of this study. The section is followed by the architecture of the seizure detector that is used throughout the project and its development (*training*) and evaluation processes. Section 3 is split into the three defined objectives and is combined from the summary and discussion of the main results obtained in the papers. Section 4 presents the summary of the project and describes directions for future work with the main focus on the clinical usefulness and applicability of detectors. This is followed by all the papers.

2 Methods

2.1 Data

The primary source of data was a publicly available neonatal EEG data set with 79 recordings [106]. Each recording is associated with one neonate/patient with a post-menstrual age between 35 and 45 weeks. All patients had suspected seizures and recordings were done on request. The data set contains 111.9 h of neonatal EEG data, individual recordings are approximately 1 h long in duration.

The acquisition of all 79 recordings was made at Helsinki University Hospital with 19 Ag/AgCl electrodes placed on the scalp according to the standardised 10-20 system. An additional electrode was positioned at the midline and served as a reference. Before reviewing the recordings for seizures, the signals were organised in the so-called *double banana* montage with 18 channels (figure 2.1): Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz and Cz-Pz. This is a rather unusual montage for neonatal EEG monitoring, typically a lower number of channels is used [94], especially if the monitoring lasts for several days.

The recordings were reviewed and annotated for neonatal seizures by three experienced human experts, denoted Expert A, Expert B and Expert C. The scorers worked independently of each other and were allowed to choose the desired filtering of the raw EEG signals, the amount of EEG visible on the screen and amplitude scaling. They had access to an electrocardiogram (ECG) but not to the clinical details of the neonates. In the final annotations, every second of the recordings is annotated as a seizure (one) or as a non-seizure (zero), but precise locations of seizures are not given.

By consensus, 39 recordings contain at least one seizure, i.e. the recordings include at least one seizure with agreement between all EEG experts. None of the scorers annotated a single seizure in 22 recordings. In figure 2.2 there is an example of consensus seizure for which the scorers agreed for the entire duration of the seizure. However, even if the experts did not completely agree on the start and/or end timestamp of a seizure and annotations intersect just for some time, this period was considered a *consensus seizure*. An example with seizure annotation exclusive to Expert C is in figure 2.3. Table 2.1 shows statistics on seizure annotations for each EEG expert and consensus annotations. Based on the statistics, Expert C visually identified the largest number of seizures, but on average their duration is the shortest in comparison with the other experts. The longest seizures were annotated by Expert B and also the sum of their durations results in the largest seizure burden, specifically, the seizure activity represents 15.71 % of all data. On the other hand, the consensus seizure activity represents only 9.75 % of the entire EEG data.

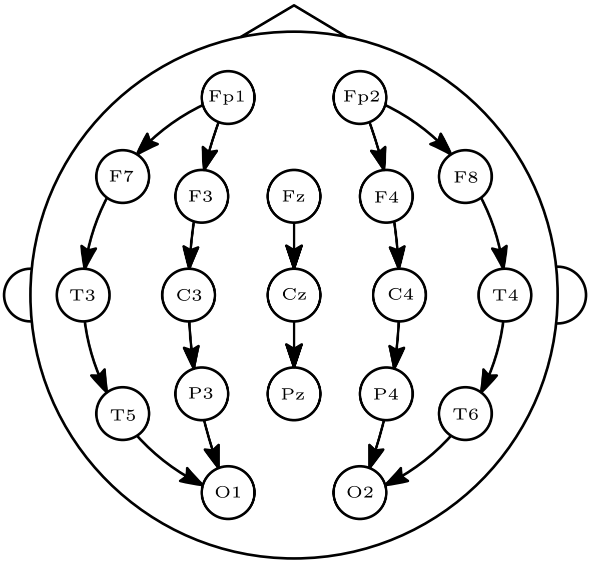


Figure 2.1. Schematic representation of a double banana montage. Electrodes (circles) are positioned according to the standardised 10-20 system and each arrow represents a channel.

Table 2.1. Statistics for seizure annotations of the neonatal EEG data set. Consensus seizures are seizure segments with complete agreement between the three EEG experts who annotated the recordings. Standard deviations are given in parentheses.

	Expert A	Expert B	Expert C	Consensus
Number of recordings with seizures	46	45	53	39
Number of seizures	402	429	548	343
Total duration of seizures [h]	13.3	17.6	14.6	10.9
Average duration of seizures per recording with seizures [min]	17.4 (22.2)	23.4 (27.7)	16.5 (22.3)	16.8 (22.9)
Average duration of a seizure [s]	119.3 (175.1)	147.5 (246.2)	95.8 (148.6)	114.5 (164.5)

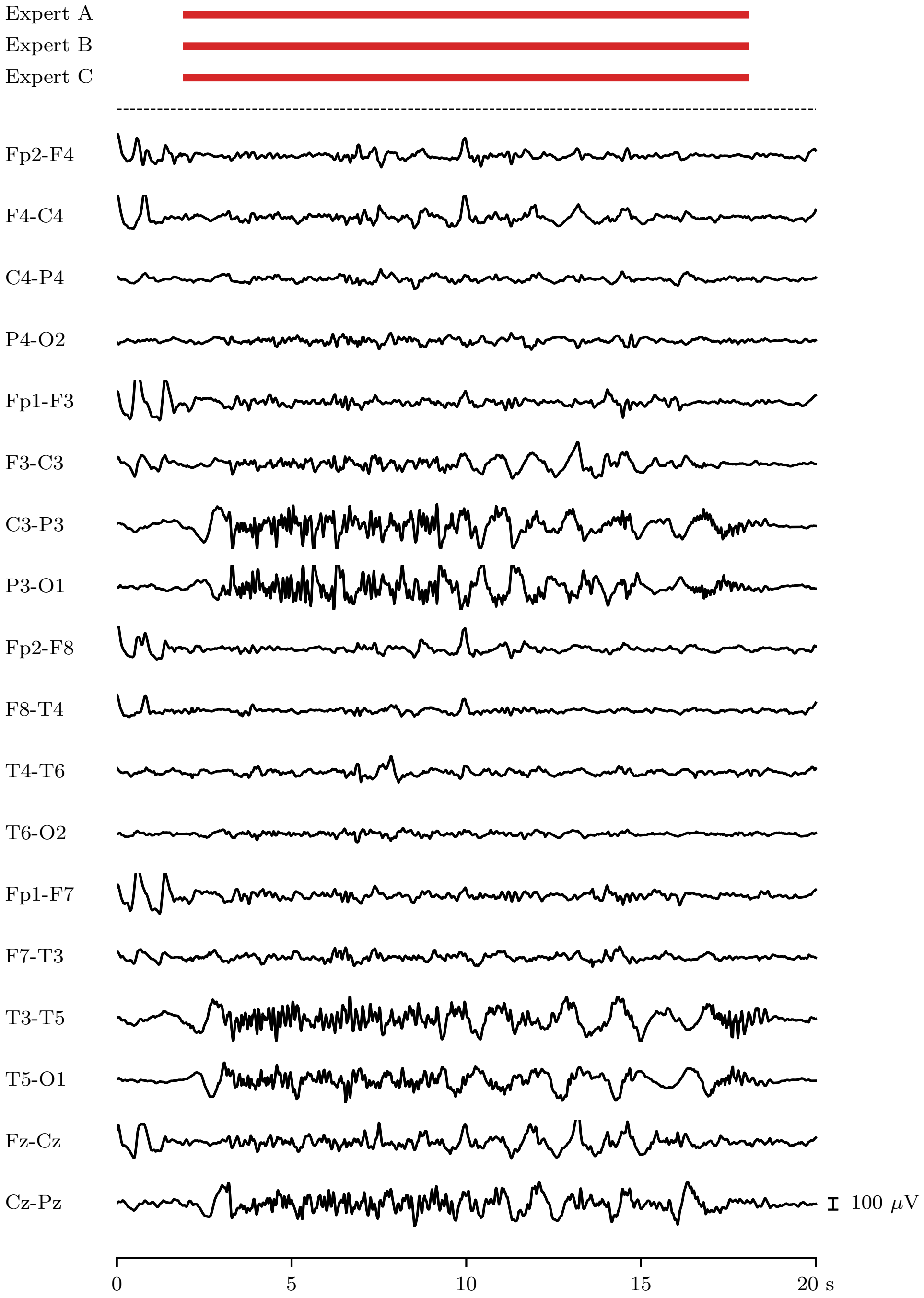


Figure 2.2. An example of a 20 s *long EEG segment containing a 16* s *long seizure with complete agreement between the experts (red lines).*

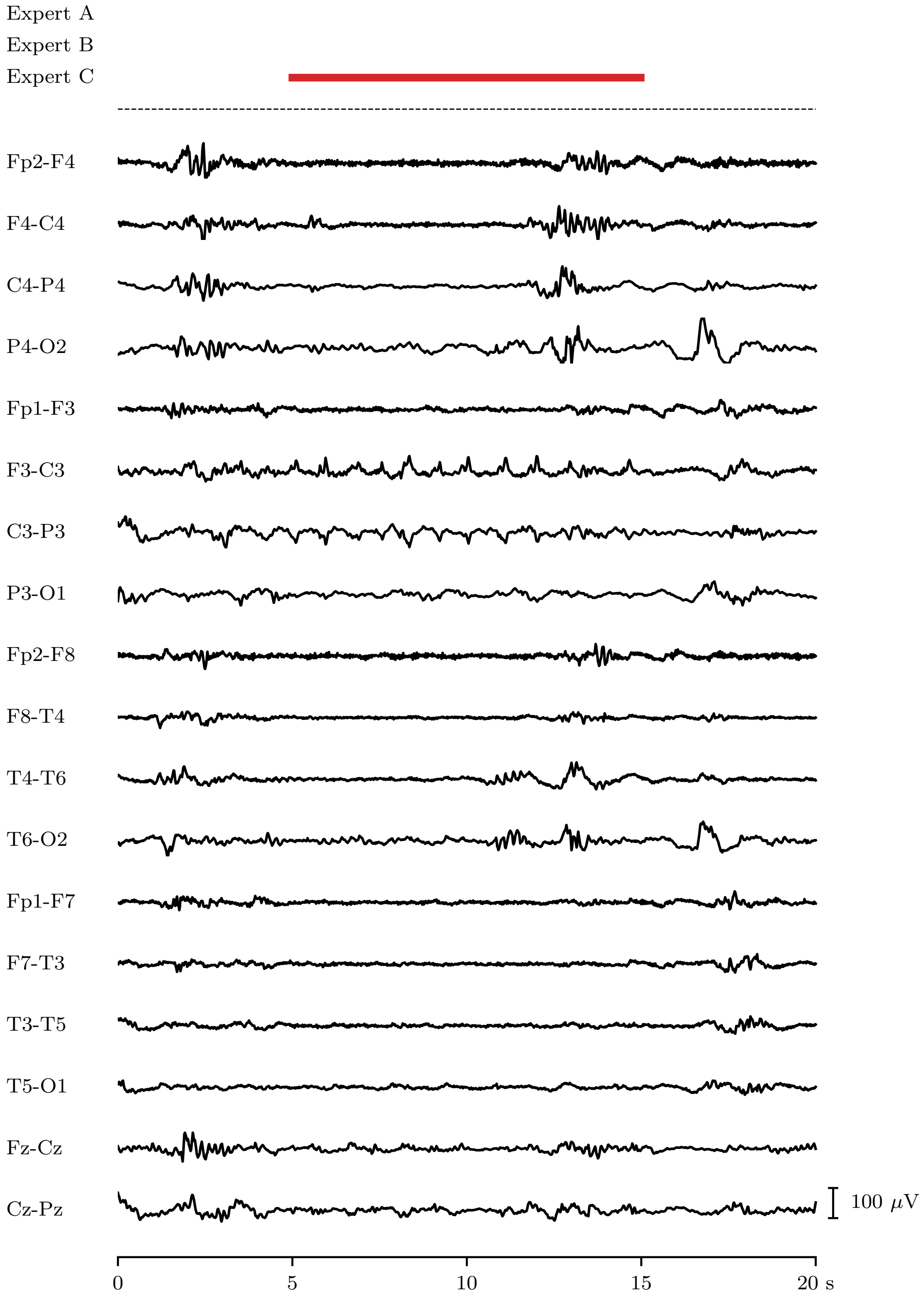


Figure 2.3. An example of a 20 s *long EEG segment containing a 10* s *long exclusive seizure annotation to only one expert (red line).*

2.1.1 Preprocessing

In all the papers, the EEG channels were organised in the double banana montage since the annotations are based on it [106]. After deriving the montage each signal was filtered, downsampled and cut into 16 s long EEG segments with 12 s overlap to increase the number of segments available for training the NSDA. In papers I – IV we applied a band-pass filter to the signals, setting the cut-off frequencies to 0.5 Hz and 16 Hz. This frequency band contains frequencies typical for neonatal seizure activity [33, 60]. The signals were further downsampled from an initial sampling rate of 256 Hz to 32 Hz to reduce the size of the input to the seizure detector and consequently the number of learnable parameters of the NSDA. The frequency band was expanded to 0.5 – 30 Hz in paper V to include frequencies typical for adult seizures [39]. In this case, the EEG signals were downsampled to 62 Hz.

In the initial papers I – IV, individual signals were normalised such that the mean amplitude was zero and the standard deviation was equal to one, a common step in the data preprocessing [45]. In later work (paper V), amplitude normalization was omitted since we found that this led to a small boost in classification performance.

Segments were generally labelled based on consensus annotations and parts of EEG recordings with disagreement between the human experts were left out from the training and test sets. Segments containing seizure activity for only a short period and not the entire duration of the segment were also excluded from the data sets. This implies that seizures of duration less than 16 s were excluded, constituting approximately 1 % of the consensus seizure data. However, it was observed that for instances where only a portion of the segment contains seizure activity, the prediction of a seizure can be made.

2.1.2 Adult data

To further validate some of the methods, we developed an SDA for adults. A dedicated detector was trained to address distinctions between adult and neonatal EEG (seizure) patterns [10]. A detector trained on one set of recordings is therefore not expected to perform well on the other due to these inherent differences. Notably, adult EEG seizure patterns are characterized by higher frequencies, typically ranging from 3 to 30 Hz, whereas neonatal seizure patterns can be as slow as 0.5 Hz [33, 39]. For the development of SDA for adults we used a publicly available data set; TUH EEG seizure corpus (version 2.0.0) [66]. The EEG recordings were done with different recording set-ups, on people of various ages, and diverse seizure types. We used a subset of EEGs recorded with an average reference. The data set contains 297 patients in the training set, 41 in the validation set and 41 in the test set. In total, there is 1095.3 h of EEG data of which 44.9 h contains seizure activity.

Preprocessing of the adult EEG signals was similar to the neonatal, but temporal central parasagittal (TCP) montage with 22 channels was used since the reviewers used it to annotate the seizures. The montage includes Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, T3-C3, C3-Cz, Cz-C4, C4-T4, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, A1-T3 and T4-A2 channels. For further details regarding the data set and preprocessing see papers IV and V.

2.2 Neonatal seizure detection algorithm (NSDA)

The NSDA used throughout this project takes 16 s long multichannel EEG segment as input and outputs probabilities of the input being a non-seizure and seizure segment. The detector first extracts features on each channel separately then combines these feature vectors into one vector by an attention layer, and, finally, maps features into two outputs (figure 2.4).

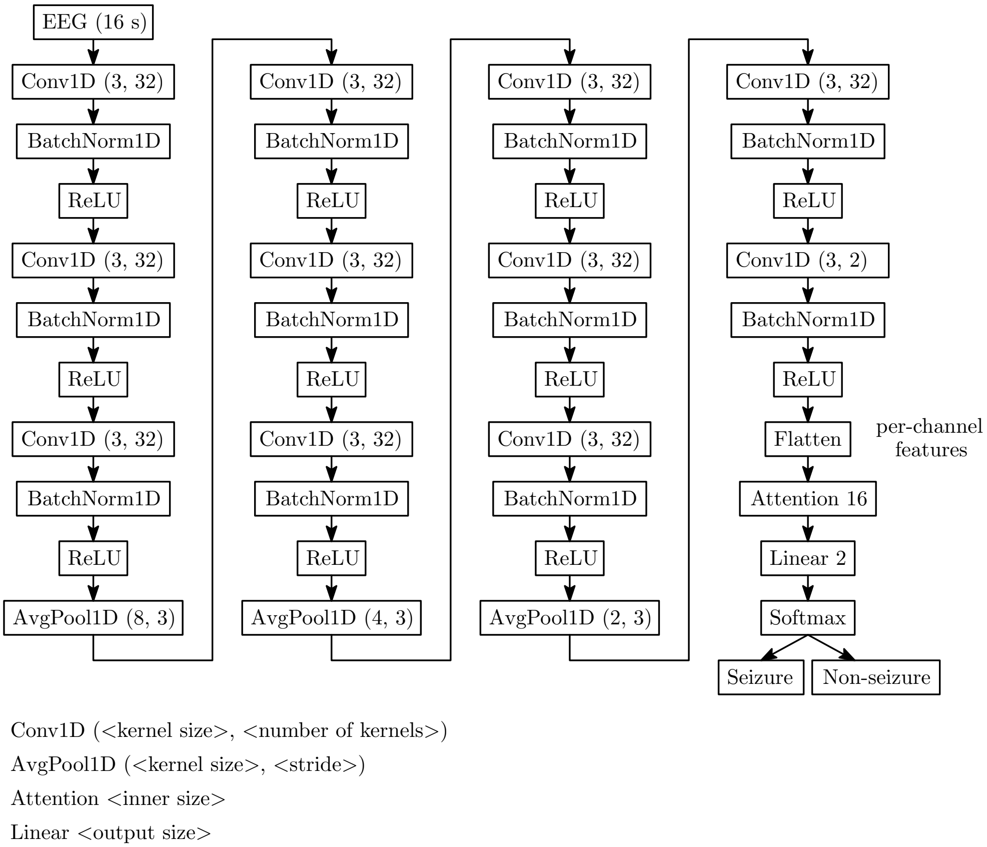


Figure 2.4. Schematic representation of the NSDA architecture. Parameters not shown in the figure were set to default PyTorch values.

A CNN was used to extract features from individual channels and the same CNN weights were shared across all the channels. The CNN had 11 convolutional layers, each followed by a batch normalisation layer and rectified linear unit (ReLU) activation [83]. With the convolutional layers, the aim was to capture patterns typical for seizure activity and since neonatal seizures are usually visible only on a few channels [74], features were extracted from each channel separately. Batch normalisation layers rescale the data which makes the training of DNNs easier and less sensitive to initial parameter values. These layers also perform regularisation and reduce overfitting of the networks to the training set [55]. ReLU activation,

$$\mathrm{ReLU}(x) = \max(0, x); \quad x \in \mathbb{R},$$

is a commonly used activation function due to its simplicity and effectiveness [30, 76]. Before the 4th, 7th and 10th convolutional layers, average pooling layers were applied to further reduce temporal resolution and to build a global representation of the input EEG signals [138].

Once the per-channel features were extracted, they were combined into a single feature vector with an attention layer [53, 56]. This layer computes a weighted sum of the feature vectors so that the channels with the largest weights contribute the most to the final seizure/non-seizure prediction. The weights are dependent on the feature values for individual channels, not on the channel positions themselves, and the idea is that more weight is given to channels with seizure-like patterns. The weights may then also be used as an indicator of seizure location. Formally, if there are C_{in} channels and L per-channel features, then the output of the attention layer is a vector of size L and can be described with PyTorch notation as

$$\text{out} = \sum_{k=0}^{C_{in}-1} a_k \cdot \text{input}(k); \quad a_k = \frac{\exp\left(w^T \tanh\left(V \text{input}(k)^T\right)\right)}{\sum_{j=0}^{C_{in}-1} \exp\left(w^T \tanh\left(V \text{input}(j)^T\right)\right)},$$

where $V \in \mathbb{R}^{L \times <\text{inner size}>}$ and $w \in \mathbb{R}^{L \times 1}$ are learnable parameters.

In the last step, the features were mapped into two (seizure/non-seizure) outputs with a fully connected (linear) layer. To interpret the outputs as probabilities they were further rescaled with the softmax function, such that the outputs summed up to one and both had values in the $[0, 1]$ range [32]. The final prediction was the class with a larger probability, or, in other words, the class with a probability larger than 0.5. The non-seizure class was denoted by zero and the seizure class by one.

The NSDA extracted $L = 24$ (58) per-channel features and had a total of $29{,}352$ $(29{,}964)$ learnable parameters if the 16 s long EEG segment was sampled at 32 Hz (62 Hz).

2.3 Training

In section 2.1 we calculated that consensus seizure activity represents less than 10 % of the entire neonatal EEG data set. To prevent bias towards the non-seizure (majority) class the same number of seizure and non-seizure segments were used in the training. We either subsampled the non-seizure segments prior to training (papers I – III) or we performed the subsampling before each pass through the data set, i.e. once during each epoch (papers IV and V). The latter strategy results in a larger number of non-seizure segments being present in the training set and is therefore preferred.

The learnable parameters of the detector (network weights) were optimised using the cross-entropy objective function and the Adam optimizer [59]. The initial learning rate was set to 0.001 and was halved every 10 or 20 epochs. The algorithm was run for 30 – 50 epochs with mini-batch sizes equal to 32, 128 or 256. A mini-batch size of 32 was used in paper II due to the small training sets (the data was split into multiple small sets). The learning rate, the number of epochs and mini-batch sizes were tuned based on training loss, i.e. cross-entropy values calculated on the training data. A dedicated validation set for hyperparameter tuning was not obtained due to the trade-off between using valuable data from the training set or having an insufficiently sized validation set for meaningful metric evaluation. We observed that the detector was insensitive to small changes in hyperparameters.

2.4 Performance evaluation

The neonatal seizure detectors were evaluated using leave-one-subject-out cross-validation. Each recording with seizures was systematically left out from the data set while the remaining data served as the training set. This evaluation technique was used due to a small number of neonates with seizures, i.e. there are 38 neonates with at least one consensus seizure longer than 16 s. In order to increase the test set in paper III, we trained an NSDA for each recording, also the ones without consensus seizure segments.

Depending on the task, a subset of metrics was used to evaluate the classification performance and the *calibration* of the NSDA. Metrics were first computed for individual patients and then averaged, we refer to these metrics as *patient*-based metrics. However, this approach is sensitive to a small number of patients and low values from a few recordings that are difficult to classify (e.g. due to seizure-like artefacts) can have a significant influence on the final result. To obtain a more stable performance estimation, the *segment*-based metrics were calculated from concatenated left-out recordings.

Classification performance was measured based on the number of correctly predicted seizure and/or non-seizure segments. Let TP denote the number of correctly predicted seizure segments (true positives), TN the number of correctly predicted non-seizure segments (true negatives), FP the number of incorrectly predicted non-seizure as seizure segments (false positives) and FN the number of incorrectly predicted seizure as non-seizure segments (false negatives). The following performance metrics were used.

The accuracy (ACC),

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},$$

measures the ratio of correctly predicted EEG segments. The proportion of correctly predicted seizure segments is measured with sensitivity (SE),

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

and the fraction of correctly predicted non-seizure segments is measured with specificity (SP),

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

Calculating the ACC, SE and SP metrics requires fixing the threshold (here 0.5) between the seizure and non-seizure classes. To measure the classification ability of a detector without choosing a specific threshold, area under the curve (AUC) can be used. The metric measures the area under the curve representing the relationship between SE and $1-\mathrm{SP}$ for all possible thresholds. If AUC equals one there exists a threshold with perfect classification.

In papers IV and V we studied the calibration of the NSDA. In other words, we investigated if the probabilities of the predicted seizure or non-seizure segments referred to as *confidences*, reflected empirical frequencies. If this is the case, the predictions for seizure and non-seizure segments with confidence estimates, say 0.9, are correct 90 % of the time. The calibration of a detector can be visualised with *reliability diagram* as shown in figure 2.5 [78]. The confidences range between 0.5 and 1 due to the fixed threshold at 0.5. The interval $[0.5, 1]$ is split into K (here $K = 5$) equally sized intervals, so-called *bins*. Each of the N testing EEG segments is then placed into one of these bins based on the confidence estimate returned by the NSDA. From the segments in an individual bin B_k ($k = 1, 2, \ldots, K$), we calculate the fraction of correct predictions (ACC) and the average confidence (CONF) in a bin. If these two values are close to each other for all the bins, the confidences reflect the empirical frequencies and the detector is considered to be *well*-calibrated. If CONF values are lower than ACC, the detector tends to be underconfident in its predictions and, conversely, is overconfident if CONF values are greater than ACC values.

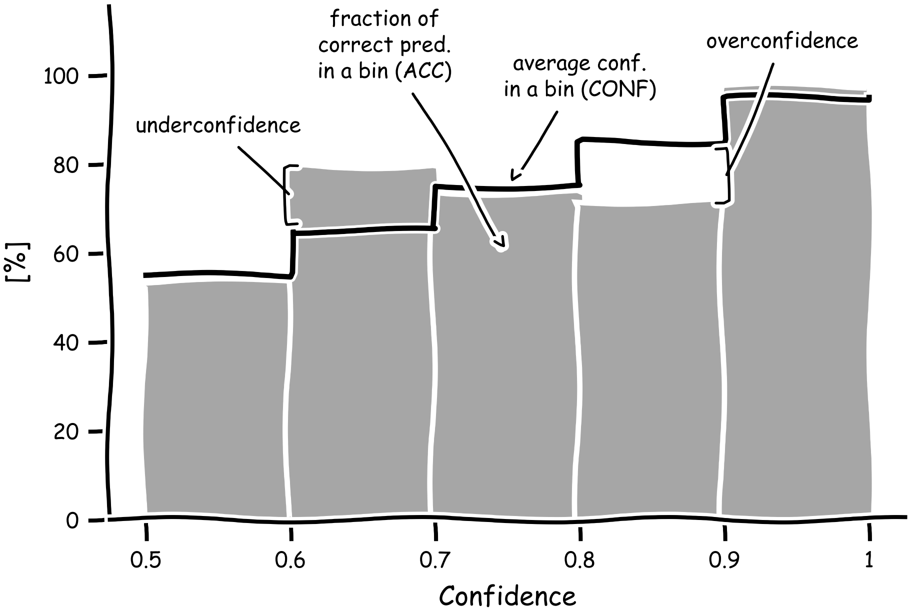


Figure 2.5. Schematic representation of the reliability diagram. The range of possible confidence levels ($[0.5, 1]$) is split into five equally sized intervals, known as bins. Each testing EEG segment is assigned to a specific bin based on the confidence of its prediction. The grey bars on the diagram indicate the ratio of correctly predicted segments within each bin (ACC), while the black curve represents the average confidence for each bin (CONF). Discrepancies between the bars and the curve suggest miscalibration, indicating that the NSDA's predictions in a bin are either under- or overconfident.

We quantified the calibration mainly using three metrics: expected calibration error (ECE) [41], overconfidence error (OE) [118] and modified static calibration error (SCE) [79]. With the ECE, the discrepancy between ACC and CONF is measured as follows

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left|\mathrm{ACC}(B_k) - \mathrm{CONF}(B_k)\right|.$$

In ECE more weight is given to the bins with more EEG segments, the metric does not provide information regarding over- and underconfidence, and possible calibration differences between the seizure/non-seizure classes are not taken into account. Since classification mistakes in medical applications are costly and a conservative detector is preferred over the overconfident one [118], we used OE,

$$\mathrm{OE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \mathrm{CONF}(B_k) \cdot \max(\mathrm{CONF}(B_k) - \mathrm{ACC}(B_k), 0),$$

and aimed for values close to zero. To address the imbalance in the seizure data and possibly different calibration performances for seizure and non-seizure EEG segments, we modified SCE,

$$\mathrm{SCE} = \frac{1}{KC} \sum_{k=1}^{K} \sum_{c=1}^{C} \frac{|B_{c_k}|}{N_c} |\mathrm{ACC}(B_{c_k}) - \mathrm{CONF}(B_{c_k})|,$$

where C is the number of classes (here $C = 2$), N_c is number of segments labelled as class c and B_{c_k} is the k-th bin with segments of class c. In words, SCE is the average of ECEs calculated on segments for each class separately.

2.5 Implementation

Preprocessing, training and evaluation of the NSDAs were done in Python 3 programming language [123]. For preprocessing and evaluation we used MNE [36], SciPy [130] and NumPy [43] libraries. The detectors were developed with PyTorch [87] and NVIDIA GeForce GTX 1080 Ti graphics card.

Code for reproducing the results in paper II is available at `github.com/anaborovac/distributed-nsda` and for the results in paper V at `github.com/anaborovac/calibrated-sda`.

3 Results and discussion

Development of reliable NSDAs faces several challenges that we divided into three research questions in section 1.4. The same framework is adapted to summarize and discuss our achieved results. The first objective addresses the scarcity of neonatal EEG data typically available for developing NSDAs and exploring utilization strategies. The second objective targets the architecture of the detector suitable for diverse recording setups, with a focus on ensuring comparable classification performance across recordings. Additionally, in order to avoid mistakes, it is preferred the detector is uncertain in the seizure/non-seizure predictions rather than certain in incorrect predictions. The third and final objective aims to investigate the user in-practice experience with NSDAs.

3.1 Training the NSDA with limited data

DNNs typically require a large amount of labelled data to achieve human-level performance [67]. However, for the development of NSDAs only a small amount of EEG data with seizure annotations is normally available and, therefore, needs to be used thoughtfully. Two main reasons preventing building larger data sets are privacy concerns related to EEG data and time-consuming scoring of long recordings. Additionally, obtaining precise time intervals with seizure activity is challenging due to ambiguous interpretation of EEG signals and the annotations are subjective to each scorer [108]. The lack of high-quality data issue is addressed in papers I and II. In paper I we show that using only parts of EEG recordings where several human experts agreed on a seizure/non-seizure annotation leads to improved classification performance of an NSDA in comparison with NSDA trained on annotations from one scorer. In paper II we address strict regulations and associated data-sharing issues by simulating ensembling multiple NSDAs. Each detector is trained on individual data sets that can not be shared due to the sensitivity of EEG data. The study suggests that the ensemble can perform as well as a detector trained on all the data united together in case there is a sufficient amount of data available in each individual data set.

3.1.1 Training the NSDA with subjective seizure annotations

In paper I we studied the effect of different annotations during training on the classification performance of an NSDA. We compared five detectors: one for each of the three scorers, one using the most frequent annotation among the scorers and one using just the EEG segments for which all scorers agreed on a seizure or non-seizure annotation (consensus labels). The latter performed best and the detector trained with labels from Expert B performed worst. In comparison with the other two scorers, annotations from this EEG expert resulted in $27 - 35$ % more seizure segments and the NSDA made $6 - 8$ % more incorrect predictions of non-seizure segments. Such differences were not observed for seizure segments, particularly, all detectors correctly predicted from 78.30 % to 80.74 % of seizure segments. This can be explained by relatively few exclusive non-seizure segments for each scorer (less than 1 %) which means that there are relatively few potentially missed seizures that could improve seizure detection if they would be detected.

The results in paper I suggest that segments with disagreements and possible incorrect annotations can have a noticeable effect on the performance of an NSDA, especially if seizures are falsely detected and these represent a big portion of the seizure class. The annotations should, therefore, be verified by multiple human experts or the so-called *noisy labels* should be taken into account in some other way [101]. It is well known that DNNs trained with noisy labels, do not perform well. Complex DNNs can even fit completely random labels [139] and from a network like that high classification performance is unexpected.

In paper I detector performing best was trained on consensus labels, in other words, on labels that are presumably correct. However, this approach requires multiple experts carefully reviewing long EEG recordings which can be prohibitively resource expensive. Automatic removal of training EEG segments with possibly incorrect seizure/non-seizure annotations would be therefore of significant use. For example, if Expert B is the only one annotating the recordings, it would be helpful to automatically choose certain EEG segments. Ideally, these segments would be the ones with an agreement with the other two experts and using them would lead to an NSDA with similar classification performance as when consensus annotations are used. To find such segments we used *confident learning*. In short, based on out-of-sample probabilities of EEG segments containing seizure activity, confident learning determines which segments may be incorrectly labelled. These segments are removed from the training set and the detector developed on the remaining data set with *clean labels* uses per-class weights to compensate for the removed segments. For further details regarding the method see appendix A.

The data set was preprocessed the same way as in paper V with one difference; here not only seizure segments overlapped but also the non-seizure ones. The total numbers of 16 s long EEG segments are shown in table 3.1. To obtain accurate out-of-sample probabilities we used Monte Carlo dropout [26], the same way as we applied it in papers IV and V. From the experiments in paper IV we observed the confidence in the seizure/non-seizure predictions can differ between the recordings and, therefore, to avoid removing segments with correct annotations we applied the confident learning on each recording separately. The NSDA trained on clean training data did not utilise dropout. For comparison reasons, we used the approach for consensus labels and individual human experts. We expected only a few segments to be removed in case consensus labels were used, given the three experienced scorers were in agreement. On the other hand, more segments were expected to be removed when Expert B annotations were used as this scorer disagreed with the other two scorers the most (table 3.1). All detectors were evaluated using consensus annotations, i.e. labels that are presumably correct.

Table 3.1. Number of seizure and non-seizure segments per scorer for the neonatal EEG data set. The numbers of exclusive segments for each scorer are given in parentheses.

	Expert A	Expert B	Expert C	Consensus
Seizure segments	10485 (332)	14241 (2188)	11139 (1052)	8563
Non-seizure segments	87014 (1559)	83132 (726)	85430 (1065)	80106

In table 3.2 there is a comparison of the detectors using all available training EEG segments and detectors using only segments that were according to the confident learning most likely correctly labelled. There were no noticeable improvements or degradation in the classification performance when only segments with clean seizure/non-seizure labels were used. The detectors utilising annotations from Expert A were closest to the classification performance of detectors developed using consensus labels. As expected from results in paper I, utilizing annotations from Expert B led to detectors with the lowest classification performance. These detectors especially showed poor performance on non-seizure segments and predicted many of them as seizures.

Table 3.2. Patient- and segment-based area under the curve (AUC), sensitivity (SE) and specificity (SP) for neonatal seizure detectors trained using all available EEG segments and only segments most likely correctly labelled according to the confident learning, referred to as clean labels. Patient-based metrics were first calculated on individual patients with seizures and then averaged. Segment-based metrics were calculated on concatenated recordings.

	Patient-based			Segment-based		
Annotations	AUC	SE [%]	SP [%]	AUC	SE [%]	SP [%]
Consensus	0.93	73.51	97.83	0.96	78.36	98.07
Clean consensus	0.92	71.06	97.44	0.97	83.58	97.62
Expert A	0.92	74.47	94.96	0.96	85.95	96.19
Clean Expert A	0.92	71.58	97.33	0.97	84.22	97.67
Expert B	0.91	75.54	89.05	0.94	83.66	91.97
Clean Expert B	0.92	76.91	91.71	0.93	80.56	93.77
Expert C	0.91	74.76	94.64	0.96	83.94	95.25
Clean Expert C	0.92	75.10	95.45	0.96	83.77	96.19

To explain the obtained results in table 3.2 we further investigated the segments removed from the training sets. We first analysed the segments with consensus annotations. On average across 38 NSDAs (one for each patient with seizures) only 1 % of the segments were removed; 5 % of seizure segments and 0.9 % of non-seizure segments. As expected, this shows that with confident learning only a few segments were removed and the training sets did not differ a lot from the ones using all segments. Consequently, the NSDAs were comparable.

Applying confident learning to segments annotated by individual experts resulted in removing 3 % of segments labelled by Expert A and Expert C, and 4 % labelled by Expert B. In these cases, a larger portion of segments were detected as potentially incorrectly annotated, especially seizure segments. In particular, 10 % (Expert A), 15 % (Expert B) and 12 % (Expert C) of seizure segments were removed and 2 % of non-seizure segments. However, differences between the NSDAs were still minor (table 3.2). In figure 3.1 we investigated the distribution of seizure segments, i.e. ratio of consensus, exclusive and segments with partial agreement (agreement of two scorers). Even though between 38 % and 45 % of exclusive seizure segments were detected as incorrectly labelled and were removed from the training sets, segments with disagreements still represented about $9-24$ % of the clean seizure segments. From the figure, it is observed that the distribution of the seizure segments did not change a lot and may partly explain why minor improvements were observed for detectors using only clean labels during the training.

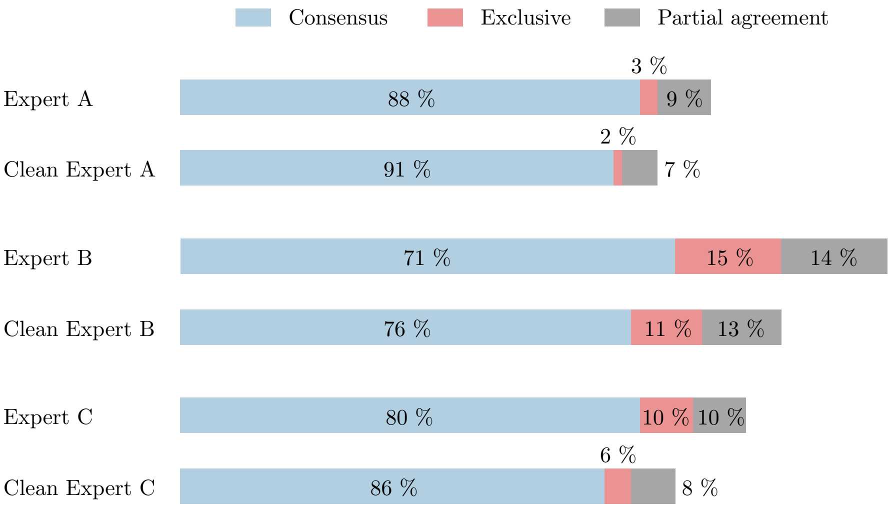


Figure 3.1. Comparison of distributions of seizure segments for each human expert before (top) and after (bottom) confident learning was applied. Percentages are averages across 38 training sets; one for each patient with seizures. For consensus seizures, all three human experts agreed on the annotation. Exclusive seizures indicate seizure annotation only by one expert. A partial agreement was achieved if one other scorer agreed on a seizure annotation.

To conclude, with confident learning we were able to detect the expected segments that might have incorrect labels, but, also some segments with likely correct annotations were removed from the training sets. Most differences between the data sets were among seizure segments. A comparable number of non-seizure segments had exclusive annotations to one expert (table 3.1), but these represent a very small portion of the non-seizure segments. Additionally, during training non-seizure segments were subsampled and segments with disagreements were not visible to the detector in each training epoch, i.e. each pass through the training set.

3.1.2 Training the NSDA with small amount of data

Strict data privacy regulations (e.g. Health Insurance Portability and Accountability Act (HIPAA) in the U.S., or the European General Data Protection Regulation (GDPR)) limit sharing EEG data between institutions and consequently restrict setting up large and diverse data sets required for the development of accurate seizure detectors. To overcome this issue in paper II, we used an ensemble of NSDAs. Each detector in the ensemble was trained on small/local data sets and the estimated seizure probabilities for new data were obtained by passing the EEG segments through all the detectors and aggregating the individual predictions. As demonstrated in figure 3.2 the design includes a trusted agent that takes care of obtaining the final predictions. Besides ensembling multiple predictions, the agent also protects the data to be predicted and detectors from malicious attacks [24, 141].

There are different ways of aggregating multiple predictions, we compared four approaches. First, the final prediction was the mean of R independent 0/1 (non-seizure/seizure) predictions, so-called *majority vote*. The second approach was similar to the first one, but instead of using raw 0/1 predictions, probabilities of input being a seizure were averaged. As the third

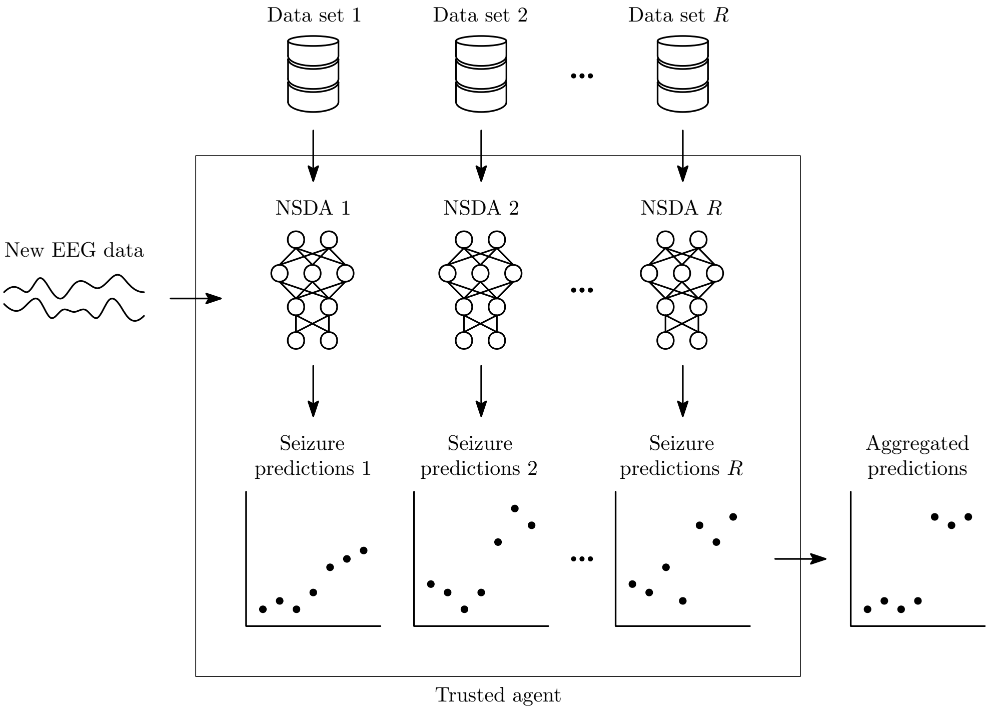


Figure 3.2. A schematic representation of the ensemble of NSDAs in practice. Each NSDA is trained on local data that can not be shared. Trained detectors are given to the trusted agent that then makes predictions on the data to be predicted.

aggregation scheme, we used the weighted average across the seizure predictions and more weight was put on the predictions from more accurate NSDAs. We aggregated predictions from $R-1$ detectors and data from a randomly chosen detector was used to determine the weights, i.e. to determine which of the detectors was more accurate [134]. Last, we used the Dawid-Skene method which alongside aggregated predictions estimates the accuracy of individual detectors for each class (sensitivity and specificity) [17].

Since our main source of neonatal EEG data with seizure annotations was the only publicly available data set [106], we simulated the situation where data are not allowed to be shared by splitting the recordings with seizure segments into $R = 3,4,\ldots,10$ smaller (local) data sets. Every time the same data was split, thus, the increased number of local data sets resulted in fewer recordings in each local set.

To further validate the obtained results in paper II, we trained an ensemble of SDAs for adults. The data was preprocessed and the detectors were trained the same way as in paper V. All patients in the training set were randomly split into approximately equally sized $R = 3,4,\ldots,10$ data sets representing the local sets unable to be shared. The experiment was repeated five times. The aggregation schemes were applied in the same manner as for the neonatal detectors in paper II. The Dawid-Skene method was applied to data from each patient separately and weights of the weighted mean scheme were determined based on data from one local SDA. This detector was not used in the ensemble.

As expected the performance of the local seizure detectors dropped with the increasing R

and fewer recordings in individual data sets (figure 3.3). The average classification performance of local detectors was also outperformed by all aggregation schemes. In paper II we observed the majority vote was slightly outperformed by the average seizure probability approach and was therefore not included in figure 3.3. With the Dawid-Skene method, most seizures were detected, but the number of falsely detected seizures was compromised by increasing R. For large values of R weighted mean performed best. With this approach, we were able to put low weights on detectors with very low classification performance. For $R = 3,4$ and 5, the local seizure detectors approached the classification performance of the *baseline* detector trained on all available recordings. In these cases, the aggregated predictions were comparable with predictions from the baseline detector. This suggests that EEG data does not need to be shared as long as there is a sufficient amount of data available in each institution and detectors are trained accordingly.

Comparing neonatal seizure detectors with adult ones in figure 3.3 we observe a smaller variance in specificity values for adult detectors. This can partly be explained by the use of patients without seizures in the training sets as well. Meaning more non-seizure data was present during the training and the detectors were able to learn the distinguishable patterns. This may further explain lower sensitivity values as proportionally there was more non-seizure data in the adult training sets than in the neonatal ones where only patients with seizures were included. On the other hand AUC values of the aggregation schemes were comparable between the data sets. The AUC of local adult seizure detectors did not decrease as rapidly as for the neonatal detectors with increasing R. In general, the training data of the adult EEG data set is larger, consequently, the local data sets contain more data and the local detectors were able to generalise better.

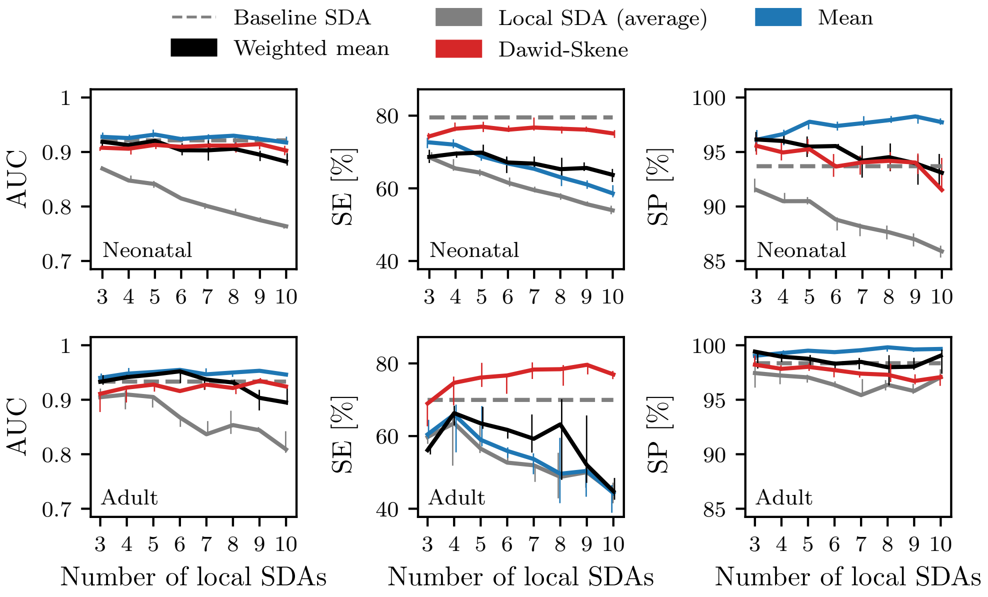


Figure 3.3. Area under the curve (AUC), sensitivity (SE) and specificity (SP) as a function of a number of local seizure detectors in neonatal (top) and adult (bottom) ensembles. The metrics are averages across patients with seizures. The solid lines show the medians of the runs and vertical lines denote interquartile ranges. The dashed line correspond to the detector trained on the union of data sets (baseline SDA). The NSDAs used for this figure were used in paper II.

3.2 Applicability of the NSDA in clinical settings

Neonates are admitted to the NICU for various reasons, e.g. in case they are born before 37 weeks of gestation, have a birth weight below 2.5 kg or have a condition that needs special care [5, 42, 97]. The condition, age [49, 52] and medications [48, 85] that are given to a neonate influence the EEG signals and cause high intra- and inter-patient variance among the signals. The variance may be further pronounced with recording equipment and used protocols in individual NICU. Therefore, in practice, an NSDA is given a large variety of input EEG signals and is still expected to perform well no matter the input. Additionally, due to the variability present in the signals, the detector is likely to encounter EEG data different from the one used for development. How well the detector works in different settings is addressed in papers III, IV and V, and in this section, we give a short summary and discussion of the obtained results. First, in paper III we address different NICU protocols by analysing the effect of reducing the number of available input EEG channels. The results indicate that the detector performs with satisfactory classification accuracy even when as few as three EEG channels are available. Second, in papers IV and V, the calibration of the NSDA is addressed by utilising methods that have worked in other domains. With these approaches, the detector is less confident in the seizure/non-seizure predictions and is consequently able to notify the user when the detections are for some reason unreliable. Further, for highly confident predictions, the NSDA is accurate in the vast majority of those.

3.2.1 Applicability of the NSDA in different recording settings

Despite the use of different EEG recording equipment in NICUs, the international standard (IEC 80601-2-26) guarantees a minimum signal quality and it limits the differences in EEG signals caused by recording equipment. Another factor that may contribute to differences in EEG signals is the number and placement of EEG electrodes. Typically, the electrodes are positioned according to the modified 10-20 system for small neonatal heads [99]; nonetheless, some differences may occur due to (minor) misplacements. The remaining differences in EEG signals are primarily influenced by individual patients. For instance, less cortical activity is expected for neonates born preterm compared to those born after completing 37 weeks of gestation [124]. Additionally, medications and treatments prescribed to patients in NICUs can impact cortical activity and lead to observable changes in EEG signals [85]. EEG signals are also prone to artefacts associated with other equipment present in the NICU (e.g. mechanical ventilation), patient movements and handling (e.g. diaper change and feeding) [119]. Therefore, while a minimum signal quality of raw EEG signals is guaranteed, there are differences in signals that cannot be eliminated through hardware or related standards.

Paper III addresses the effect of the number of EEG channels on a seizure detector and its related classification performance. The NSDA used in this work is capable of handling any number of channels. This is achieved by extracting features from each EEG channel separately and subsequently combining them using an attention layer that can process any number of feature vectors. We compared the classification performance of the NSDA on three different montages with 3, 8 and 18 channels. Unsurprisingly, the proportion of correctly detected seizures decreased when reduced montages were used as the seizure activity was not picked up by the used electrodes. For instance, when only 3 channels were utilized instead of 18, the proportion of detected seizure segments (sensitivity) dropped from 87 % to 69 %. A similar decrease has also been observed for human experts and their visual detection using reduced montages [104, 109]. These findings suggest that the NSDA detects a satisfactory number of seizures even when employing reduced montages.

Conversely, when utilizing a limited number of EEG electrodes/channels, they are typically positioned in areas that are less prone to artefacts [133] and can result in fewer falsely detected seizures. The findings from employing only a limited number of EEG segments containing artefacts indicate that just a few types of artefacts noticeably affect the NSDA and its potential seizure detection. Respiratory and ECG artefacts are examples of such artefacts. These artefacts are rhythmical in nature (like seizures) and therefore, can cause false seizure detections, i.e. incorrectly predicting a seizure when there is none. Figure 3.4 presents one such example. It shows a 16 s long EEG segment that was inaccurately classified as a seizure by the NSDA used in paper III. The attention coefficients were highest for the T6-O2 and Fz-Cz channels, both of which are clearly contaminated with an ECG artefact. This means that the detection was mainly based on channels contaminated with a rhythmic artefact. It was not observed that specific types of artefacts caused missed seizures.

To limit false detections caused by seizure-like artefacts, an artefact detector can be applied [19, 107, 133]. The straightforward application of an artefact detector could indicate uncertain predictions. Since ECG and respiratory signals are already normally recorded for every neonate, these signals may be incorporated into the detector [38] with e.g., multimodal learning [7]. If EEG signals look similar to ECG signals, this should indicate that the patterns visible in EEG signals are not the result of seizures. Another option would also be to analyse longer periods of EEG segments by using different architectures (e.g. transformers [128]). We know that seizure activity evolves over time, but ECG and respiratory signals have approximately the same frequency over time. By analysing longer periods of signals change in frequencies could be captured.

3.2.2 Applicability of the NSDA for different patients

From a closer investigation of the NSDA and its classification performance on individual recordings in figure 3.5, we realised that classification accuracy was comparable to human experts for most of them, however, it failed for a subset (12/38) of the recordings. For these, the fraction of detected seizures (sensitivity) was lower than ∼70 % and the fraction of correctly classified segments with non-seizure activity (specificity) was lower than ∼90 %. For 5 out of these 12 recordings AUC values were greater than 0.9 which indicates that the 0.5 threshold was most likely not the optimal choice. The low sensitivity of 2/12 recordings was related to low seizure burden. These recordings contain less than 1 min of seizure activity and it means that the detector might (partly) miss only one short seizure which led to low sensitivity values. Short seizures (e.g. shorter than 30 s) can be difficult to classify even for human experts [108]. Two recordings appeared to have low specificity. One of these recordings had annotated biological rhythms not related to the cortical activity and some of the incorrect seizure detections were due to artefacts like the one in figure 3.4. For some recordings, there was no obvious explanation. For these, we assumed the EEG signals were somehow different from the ones used in the training phase of the NSDA. In other words, the data set available for the development of the detector was not diverse enough to capture *all* relevant EEG patterns. The NSDA also returned probabilities close to one for all the EEG segments, i.e. it was equally confident in correct and incorrect predictions. Therefore the detector cannot be used to notify the user when the predictions are unreliable and potentially incorrect. This motivated the work done for paper IV and its extended version, paper V.

An NSDA is considered to be well-calibrated if the probabilities returned by the classifier reflect observed probabilities. A well-calibrated detector is therefore mostly correct in case

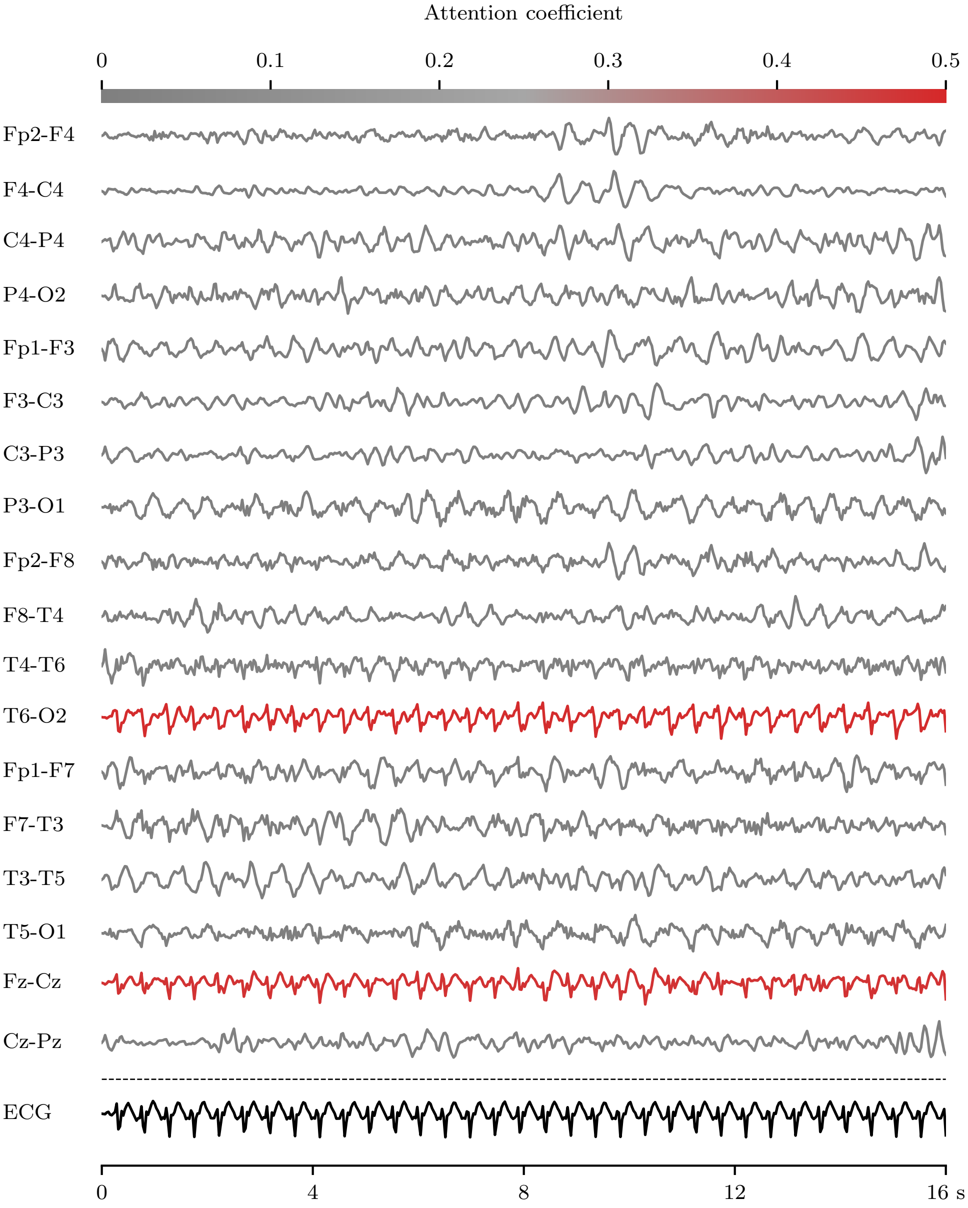


Figure 3.4. An example of an EEG segment that was incorrectly classified as a seizure due to the presence of an ECG artefact. The NSDA made the detection mainly based on channels T6-02 and Fz-Cz that clearly contain ECG artefact. The same preprocessing of all signals and the detector were used in paper III. The NSDA was trained using the 18-channel montage.

the confidence in prediction is close to one, and the predictions are completely unreliable if the confidence is close to 0.5. In an attempt to improve calibration, we tested Monte Carlo dropout [26] in paper IV. We studied calibration further in paper V where we compared temperature scaling [41], dropout [26], mixup [118] and deep ensembles [63]. The obtained results suggest that by using any of these approaches the detector is less confident in the seizure and non-seizure predictions, especially in incorrect ones. The detectors were also mostly correct for EEG segments with highly confident predictions. This was evident in the overconfidence error values, which were closer to zero in comparison with values calculated from the uncalibrated detectors. In figure 3.6 there is an example of seizure predictions made by an uncalibrated and calibrated NSDAs on one recording. For this recording, the experts were mostly in agreement about the annotations, except for the first seven seconds of the seizure. Both detectors, the uncalibrated and calibrated one, detected the annotated seizure, but the calibrated NSDA was less confident in the predictions and the returned seizure probabilities were closer to the 0.5 threshold. This holds also for the part where detectors were incorrect and missed a portion of the seizure.

In order for the detector to be more useful in practice and to reduce the review time of human experts, the detector should only pass a small part of the recordings on to the expert for further review. In our case, if we used a threshold of 0.9 for confident predictions, the NSDA with mixup would pass on to a human expert about 41 % of all EEG segments which was the largest portion of segments among the calibration approaches. Detectors utilizing ensembling and dropout would require human expertise for about 18 % and 21 % of the segments, respectively. The smallest portion of the segments would be reviewed for the uncalibrated NSDA, about 9 %. The ratios of uncertain predictions for each recording can be observed in figure 3.7. Assuming the human expert is always correct in seizure annotations, the uncalibrated model would be least improved by the visual inspection. We can also observe that deep ensemble and mixup falsely detected seizures with a confidence lower than 0.9 and most of them were eliminated with the help of an EEG expert, i.e. specificity values are close to 100 % for the majority of recordings.

Figure 3.7 shows that for some recordings and 0.9 confidence threshold, detectors utilizing calibration methods would ask the user to annotate more than 80 % of the whole recording. In an attempt to reduce this effort, we tried to re-train the classifier using *transfer learning*. Here, the feature extractor of the NSDA was left unchanged but the weights in the final linear layer (classification layer) were adapted for a specific patient/recording. The adaptation of the classification part of the DNN was based on only a few EEG segments to limit the interaction with the human expert. These experiments were however not successful. Figure 3.8 shows that even if the whole recording would be annotated and logistic regression would be used as a classifier, we would be unable to achieve *perfect* classification performance on all recordings. This means that the features extracted with a CNN were not descriptive enough to distinguish between seizure and non-seizure classes in a linear manner. This experiment shows, that a straightforward application of transfer learning is unlikely to resolve the issue. Further research is needed to find an appropriate adaptation method that improves the classification performance of a detector and does not require a lot of interaction with a human expert.

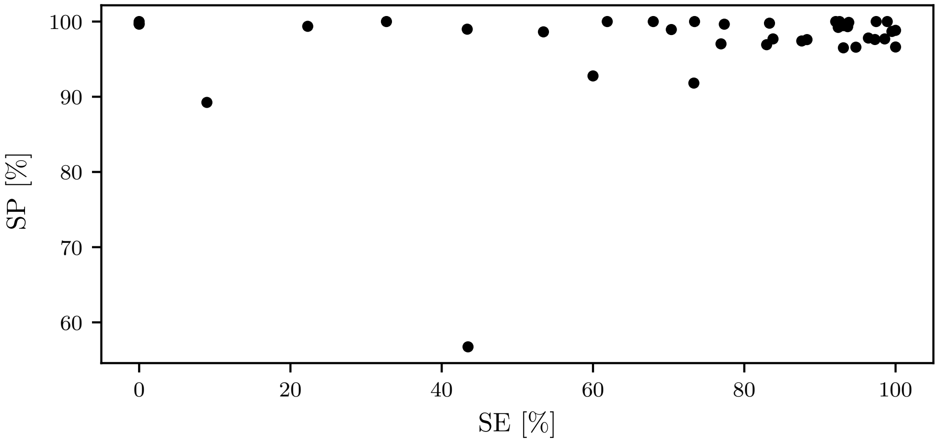


Figure 3.5. Sensitivity (SE) and specificity (SP) values for individual recording with seizure segments. The NSDA used for this figure was used in paper V as the uncalibrated detector.

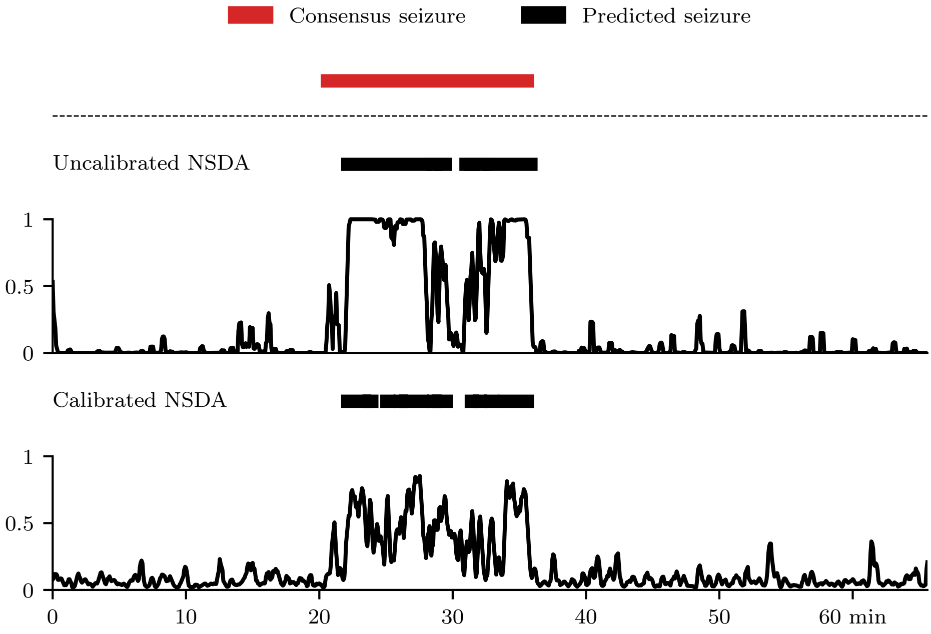


Figure 3.6. Seizure predictions for a single recording using both an uncalibrated and calibrated NSDA. The calibrated detector employed mixup. The black curves represent seizure probability estimates (confidence values), with the corresponding seizure predictions illustrated by the black blocks. In this recording, all three experts were in agreement, except for the initial 7 s *of the seizure (red block). The same detectors were used in paper V.*

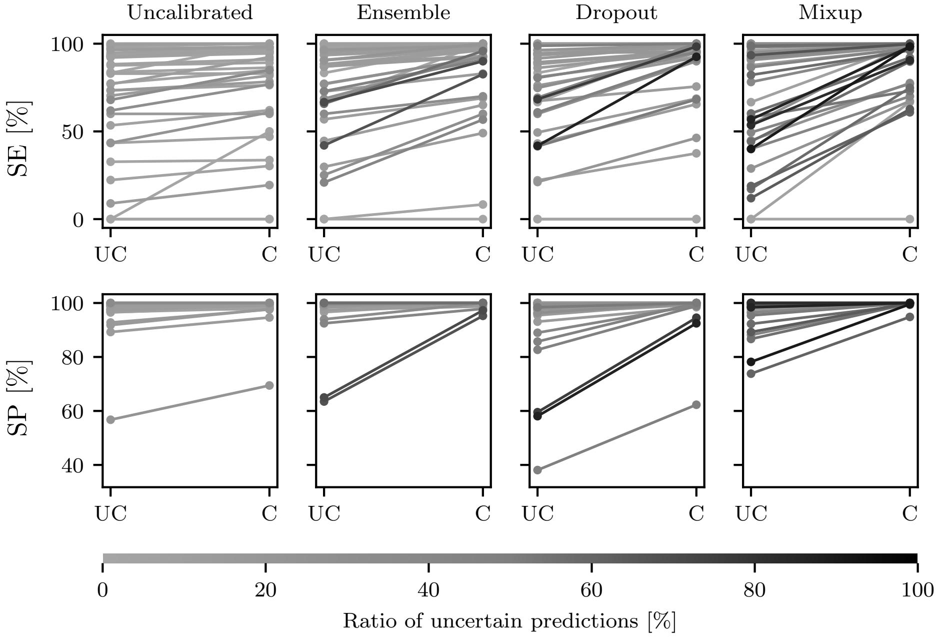


Figure 3.7. Sensitivity (SE) and specificity (SP) of uncorrected (UC) and corrected (C) NSDAs. Corrections are done by a human expert and it is assumed that corrections are always accurate. The threshold for confident/unconfident predictions is set to 0.9. Each line represents one recording/patient with at least one 16 s *long seizure segment. All the models were also used in paper V.*

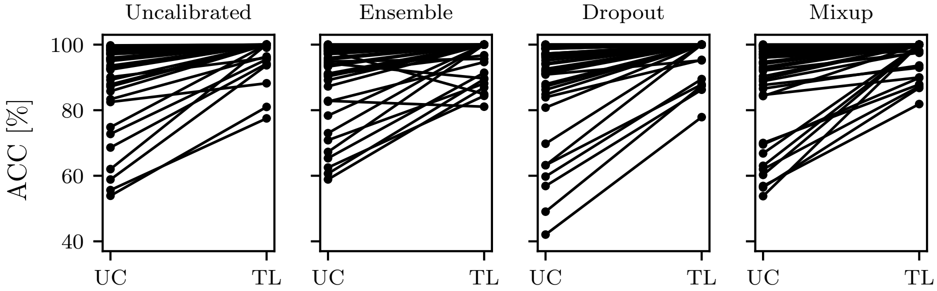


Figure 3.8. Accuracy (ACC) of uncorrected (UC) NSDA and NSDA with transfer learning (TL). A logistic regression classifier is used for transfer learning and takes extracted features (outputs of the attention layer) as inputs. For detectors utilizing ensemble and dropout, the inputs to the logistic regression classifier are the average features. Predictions from the detector using dropout were recalculated for this figure and are not exactly as the ones used in the paper V.

3.3 NSDAs in clinical practice

NSDAs have been in development for years and a few have been integrated into commercial EEG monitoring systems. In this section, we summarize and discuss results in paper VI related to an NSDA used in a Dutch hospital. Findings from the interview with NICU nurses indicate that an NSDA is a valuable asset in the NICU even though each automatic detection needs to be confirmed by visual inspection of the corresponding EEG signals. To further increase the value of an automatic seizure detector, the commercial EEG system should be upgraded according to the latest research. However, NSDAs are rarely integrated into the monitoring systems. In collaboration with Kvikna Medical ehf. we address this issue and update their cloud-based EEG system with an application programming interface (API) allowing easy and quick integration and evaluation of third-party NSDAs into a clinical environment.

3.3.1 User experience with a commercial NSDA

In order to obtain user insights about an NSDA, we interviewed the primary users, 20 NICU nurses. Nurses are always present at the bedside and interact with an EEG monitoring system the most. The interview took place at Wilhelmina Children's Hospital (WKZ), University Medical Centre Utrecht (UMCU), which specialises in neonatal neurology. The overreaching aim of the study was to collect feedback about the EEG monitoring system currently in use and obtain suggestions for improvements. At the time of the interviews, the hospital employed an Olympic Brainz Monitor (Natus Medical Inc., USA) for continuous EEG monitoring with a built-in NSDA, called RecogniZe. The detector is intended to detect seizures of term neonates. It takes three-channel EEG (P3-P4, C3-P3 and C4-P4) as input and makes a decision based on the regularity of the signals [54].

Despite many falsely detected seizures (mostly due to artefacts), the nurses found the NSDA useful in identifying interesting parts of long EEG recordings. A yellow bar appears on the screen for every automatic detection and the users can retrospectively inspect the corresponding EEG by clicking on these bars. The bars are visible from a distance and allow the nurses to glance at the monitor from their workstations or when they are passing by the monitor. This is a favourable feature of the EEG system since each nurse is usually responsible for 2 – 3 neonates and cannot devote all of their attention to one monitor.

Each automatically detected seizure is first verified by the nurse responsible for the patient. Less experienced nurses or nurses in doubt due to ambiguity present in the EEG signals, seek assistance from colleagues. In case the nurses suspect a seizure, a neonatologist is called and takes care of an appropriate treatment plan (e.g. anti-seizure medication) if seizures are confirmed. To conclude, the nurses regularly use the NSDA despite its imperfections and mainly use it as a guide for identifying points of interest in EEG recordings.

Our findings in paper VI agree with Sharpe et al. [98]. Both studies suggest that commercially available NSDAs are useful in clinical practice even if each automatic detection needs to be verified by an EEG expert. Interviewed users were, in general, satisfied with the seizure detection rate, however, the NSDAs falsely detected many seizures due to artefacts present in the EEG signals. Reviewing each automatically detected seizure is therefore important to limit the over-diagnosis of seizures. This is a particular concern since anti-seizure medication, such as phenobarbital [65], have undesired side effects and can cause neurodegeneration [129].

In paper VI we found that the nurses are in most cases able to recognise seizure EEG patterns and distinguish them from artefactual ones. They are then capable of filtering out some of the automatically falsely detected seizures and the neonatalogist is not called for every single detection. However, WKZ specialises in neonatal neurology and NICU nurses receive additional EEG training at UMCU where they learn about the application of electrodes and interpretation of the signals. Such training is not part of formal nursing education but would be of significant clinical benefit as EEG is getting recommended more, not only for seizure detection but also for its prognostic value [95], and having an EEG expert available at the bedside 24/7 is unfeasible [98].

3.3.2 Deployment of the NSDA in a clinical practice

Despite decades of academic research on NSDAs and evidence from clinical practice that automatic detectors are a useful tool [98], there are currently only two commercially available EEG monitoring systems specialized for neonates with integrated detectors, the Olympic Brainz Monitor (Natus Medical Inc., USA) and nëo CFM (ANT Neuro b.v., Germany). One of the main reasons why NSDAs are not routinely integrated into commercial EEG systems is their bad performance and lack of demonstration on real-world clinical data that would convince manufacturers of EEG systems to incorporate the detectors into their monitors.

For research purposes, it would be useful if existing EEG monitoring systems would be compatible with any third-party NSDA. Under the appropriate regulations, this would enable researchers to properly validate the detectors, obtain realistic values of performance metrics and improve the detectors accordingly [23, 31, 132]. In collaboration with Kvikna Medical ehf. we designed and implemented an API allowing quick and easy integration of the detectors into their cloud-based Stratus EEG (Kvikna Medical ehf., Iceland). The interface consists of two main parts (figure 3.9): initialisation (Init API) and review (Review API). In the initialisation phase, the location of the NSDA (in the form of an internet protocol (IP) address) is given to the Stratus EEG server and the detector is provided with the credentials required to access the data. After this step is completed, the detector can get the data to be processed through the Review API. This API also enables sending the results of the detector back to the Stratus EEG server which then takes care of displaying the results. In practice, this lets the users of the Stratus EEG monitoring system perform automatic detections with an NSDA of their choice without the need to switch to a new or upgraded monitoring system.

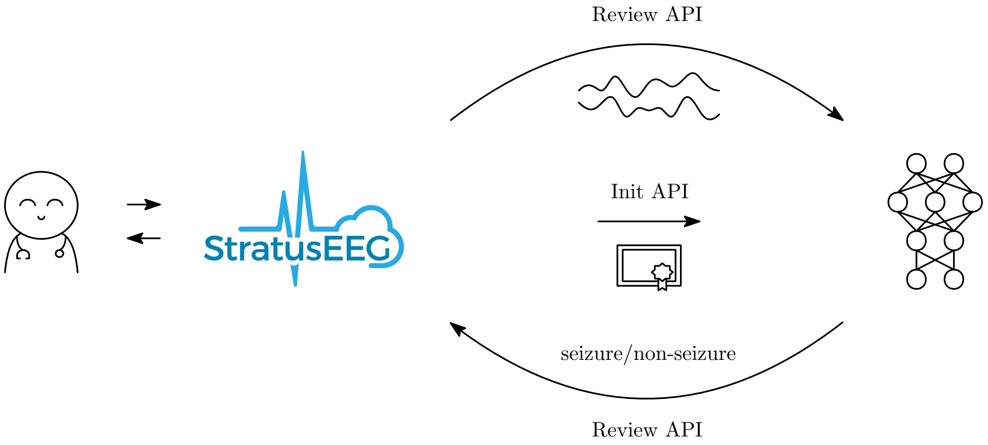


Figure 3.9. Schematic representation of the Stratus EEG API enabling integration of a third-party automatic tool, e.g. an NSDA.

The way the API is designed, any automatic tool with numeric output can be integrated, i.e. the interface is not limited to seizure detectors. Utilizing the Google remote procedure call (gRPC) framework for implementing the API adds even more flexibility to the approach as developers can implement the tools in different programming languages like Python, C/C++, and Java. That is, the tools do not need to be written in the same programming language as Stratus EEG.

While the API-oriented framework offers several advantages, it also presents specific administration and technical challenges. Establishing a clear agreement between users and algorithm owners is required to protect the data and define explicit terms regarding algorithm usage. Additionally, incorporating a third party into the monitoring system introduces an additional potential source of errors. To ensure overall accurate outputs and a good end-user experience, algorithm owners must maintain a stable internet connection, manage the overall workload efficiently, and keep the tool synchronized with the API.

4 Summary and future perspectives

Mortality among neonates has decreased significantly in the past decades, and now the emphasis is on improving the lives of everyone through brain-focused care. EEG is one of the non-invasive techniques to continuously monitor the neonatal brain that gives insights into the current state of the brain and can also be used as a prognostic tool. Interpretation of EEG signals provides information on the current sleep/wake state, discontinuity/continuity of brain activity and presence of abnormalities such as seizures. Seizures are the most common neurological emergency in the first four weeks after birth and can cause permanent brain damage and associated brain impairments later in life. If seizures are detected and treated (e.g. with anti-seizure medication) early, these damages may be (partly) prevented. EEG is required for the detection of neonatal seizures as they are typically clinically silent and visible just on EEG recordings. However, few NICUs have the needed EEG expertise constantly available at the bedside and a reliable automatic seizure detector that takes EEG as input would be of great clinical significance. In this work, we improve an NSDA based on deep learning with the main focus on real-world challenges and clinical usefulness.

NSDAs have been in development for about 30 years and human-level performance has yet to be reached. One of the reasons making the development challenging is the complexity of neonatal EEG signals. Interpretation of the signals is highly ambiguous even for experienced human experts and as a consequence, the annotations are subjective to each scorer and may have mistakes. These mistakes can noticeably affect the development of seizure detectors (paper I). To limit incorrect labels agreement between multiple experts can be used, but is often too expensive. To automatically select EEG segments with likely correct labels before training the ML-based detector, we propose to use confident learning. With this approach, a large portion of EEG segments with disagreements between the experts are removed, but also a substantial portion of segments with complete agreement are removed. Some valuable information is therefore removed from the training set and classification performance is not noticeably improved. Since mistakes in seizure annotations are inevitable, future research is needed to further address this issue. Besides removing EEG segments prior to training of seizure detector, one could also use methods addressing label noise during training of seizure detectors [101].

EEG signals are also highly diverse between the neonates, depending on their age, condition, medication that is given to them, and other factors, e.g. recording protocols. Many hours of EEG recordings from multiple individuals are therefore required to learn the relevant patterns and achieve seizure detectors with human-level classification performance. However, the amount of data is limited by the time-consuming annotation process and strict data privacy regulations. To overcome the data privacy issues, we propose to use an ensemble of locally developed NSDAs (paper II). The results on neonatal and adult EEG suggest that if a sufficient amount of data is available at each institution, the ensemble can perform as well as a detector trained on the union of the locally available data sets. In our study, local detectors are trained on a subset of one data set. Future research is needed to verify the results with data sets from various sources, as each data set may have specific properties [135].

Due to the complexity and diversity of EEG signals among patients, it can be expected that the seizure detector will in practice encounter patterns that were not used during the training. To avoid mistakes in such cases, it would be preferred if the detector notifies the user that the input EEG signals look somehow different and the predictions are not completely reliable. Incorrect predictions can be costly since untreated seizures may cause permanent brain damage, but over-diagnosis and treatment with anti-seizure medications also have undesired side effects and can negatively influence the developing brain. We show that utilising calibration techniques that have been effective in other domains leads to detectors less confident in predictions, especially in incorrect predictions (papers IV and V). The detector can then inform the user about uncertain predictions. It remains unclear how useful such detectors would be in clinical practice. For instance, how probabilities would be visualized alongside raw seizure/non-seizure predictions while preserving the visibility of automatic detections from a distance and what the threshold for certain predictions would be. And finally, would such a detector reduce the reviewing time while maintaining the classification performance? Studies on ML-tools working with pathologists for histopathologic interpretation suggest that reviewing time can be reduced and interpretation more accurate [58, 103]. However, for examples that are difficult to classify, pathologists tend to heavily rely on the automatic tool. This can lead to a decrease in accuracy in case the automatic classification is incorrect [58].

The architecture of the applied NSDA can make a prediction for EEG segments with any number of channels (montage). We show the seizures picked up by a specific montage get detected (paper III). In case a larger number of channels are used, more seizures get detected but also more artefacts are picked up. Some artefacts, e.g. ECG, are rhythmic and can be misdiagnosed as seizures. To limit false detections, an artefact detector can be applied [19, 107, 133] or artefacts can be addressed by the architecture of the seizure detector [7, 128].

Artefacts also cause many falsely detected seizures in commercially available NSDAs (paper VI). However, nurses at WKZ still find the detector useful. It remains unclear whether detectors would be useful in NICUs that do not specialize in neonatal neurology and lack the required expertise. Despite the advances in ML techniques, which have led to automatic neonatal seizure detectors approaching human-level performance, their implementation in new clinical environments should be done with special care to avoid misdiagnoses. One potential approach is to ensure the availability of external expertise when needed. To facilitate this, remote access to the EEG monitor should be granted.

A Confident learning

Incorrect annotations, known as *label noise*, are present in almost every data set, even widely used benchmark data sets in computer vision, natural language and audio processing have incorrectly annotated examples [80]. If incorrect labels are present during the development of a classifier, its performance is often compromised [25] and evaluation is not reliable [80]. The issue can be addressed by selecting the data with probably correct labels before the training or evaluating of a classifier, using regularisation techniques preventing the classifier from overfitting the training set and making the classifier or objective function to be optimised, robust to label noise [101]. Confident learning [81] is a methodology that detects examples with potentially incorrect labels, i.e. we aim to remove incorrectly annotated seizure or non-seizure EEG segments from the training (and test) sets.

Let X be a data set with given (noisy) labels $\tilde{y}$, true underlying labels y^* and the labels are one of c classes. Confident learning aims to detect $x \in X$ for which the given label may be incorrect ($\tilde{y} \neq y^*$). For this approach, out-of-sample predictions and accompanied probabilities on X are required. To obtain these, a model (detector) with parameters θ is used. In case such a pre-trained model is not available, k-fold cross-validation may be used. Here we use $k = 3$.

There are three main steps in confident learning [81]. First, based on out-of-sample probabilities we estimate *confident joint* $C_{\tilde{y},y^*}$ which represents an unnormalised joint distribution between $\tilde{y}$ and y^*. The diagonal elements of $C_{\tilde{y},y^*}$ correspond to the instances that are correctly annotated with high confidence. On the other hand, the off-diagonal elements are associated with likely incorrectly annotated examples. In particular, these instances are labelled with class, say i, but the out-of-sample prediction equals class j ($j \neq i$) and the corresponding probability exceeds a class-specific threshold t_j. In our study, the off-diagonal elements would be the number of EEG segments that are likely mistakenly annotated as seizures or non-seizures.

Formally, $C_{\tilde{y},y^*}$ is a $c \times c$ matrix with the following elements,

$$C_{\tilde{y},y^*}[i][j] = \Bigg| \Bigg\{ x \in X_{\tilde{y}=i} : \hat{p}(\tilde{y} = j; x; \theta) \geq t_j \text{ and } \\ j = \underset{l \in [c]: \hat{p}(\tilde{y}=l;x;\theta)}{\arg\max} \; \hat{p}(\tilde{y} = l; x; \theta) \Bigg\} \Bigg|; \quad i,\ j \in [c],$$

where $|\cdot|$ denotes the size of a set, $[c] = \{1, 2, \dots, c\}$, $\hat{p}$ is an out-of-sample probability and $X_{\tilde{y}=i}$ is a subset of data X with given annotations equal to i. Per-class threshold t_j is calculated as the average out-of-sample probability of examples labelled as j belonging to class j,

$$t_j = \frac{1}{|X_{\tilde{y}=j}|} \sum_{x \in X_{\tilde{y}=j}} \hat{p}\left(\tilde{y} = j; x; \theta\right); \quad j \in [c].$$

In the second step of confident learning, potentially incorrect instances are removed from the data set X. There are several possibilities for this step, we choose the one suggested in [81] and remove instances corresponding to the off-diagonal elements of matrix $C_{\tilde{y},y^*}$. By doing this EEG segments most likely incorrectly annotated are removed, but segments close to the decision boundary (0.5) typically remain in the data sets.

Last, in case the data set is used as a training set of a classifier, the classifier is trained with a per-class weighted objective function (e.g. cross entropy) to compensate for the removed examples,

$$w_i = \frac{1}{\hat{p}(\tilde{y} = i|y^* = i)}; \quad i \in [c].$$

The closer the conditional probability $\hat{p}(\tilde{y} = i|y^* = i)$ is to zero, the more instances labelled as class i are not likely correctly labelled. Consequently, more data labelled as class i correspond to the off-diagonal elements of $C_{\tilde{y},y^*}$ (is removed from the data set) and more weight is given to the remaining examples. On the other hand, if the probability is close to one, very few examples labelled as class i are removed from the training set and no extra weight is given to these instances.

Confident learning has been used before for the correction of the seizure annotations [140]. In comparison with our work, Zhang et al. apply the approach to the EEG signals obtained with a wearable device and not high-end laboratory equipment, they make the assumption that no seizure segments were missed and empirically define a threshold that detects EEG segments with possibly incorrect labels.

Confident learning can be used with the open-source framework Cleanlab, available at `github.com/cleanlab/cleanlab`.

References

[1] Rehan Ahmed et al. "Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel." In: *Computers in Biology and Medicine* 82 (2017), pp. 100–110. DOI: 10.1016/j.compbiomed.2017.01.017.

[2] Amir H. Ansari et al. "A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants." In: *Journal of Neural Engineering* 17.1 (2020), p. 016028. DOI: 10.1088/1741-2552/ab5469.

[3] Amir H. Ansari et al. "Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor." In: *Clinical Neurophysiology* 127.9 (2016), pp. 3014–3024. DOI: 10.1016/j.clinph.2016.06.018.

[4] Amir H. Ansari et al. "Neonatal seizure detection using deep convolutional neural networks." In: *International Journal of Neural Systems* 29.04 (2019), p. 1850011. DOI: 10.1142/S0129065718500119.

[5] Bhagat Baghel, Anurup Sahu, and K. Vishwanadham. "Pattern of admission and outcome of neonates in a NICU of tribal region Bastar, India." In: *International journal of medical research professionals* 2 (2016), pp. 147–50. DOI: 10.21276/ijmrp.2016.2.6.029.

[6] Shahina Bano, Vikas Chaudhary, and Umesh C. Garga. "Neonatal hypoxic-ischemic encephalopathy: A radiological review." In: *Journal of pediatric neurosciences* 12.1 (2017), p. 1. DOI: 10.4103/1817-1745.205646.

[7] Khaled Bayoudh et al. "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets." In: *The Visual Computer* (2021), pp. 1–32. DOI: 10.1007/s00371-021-02166-7.

[8] Stella T. Björkman et al. "Seizures are associated with brain injury severity in a neonatal model of hypoxia–ischemia." In: *Neuroscience* 166.1 (2010), pp. 157–167. DOI: 10.1016/j.neuroscience.2009.11.067.

[9] Geraldine B. Boylan, Nathan Stevenson, and Sampsa Vanhatalo. "Monitoring neonatal seizures." In: *Seminars in Fetal and Neonatal Medicine* 18.4 (2013). Neonatal seizures, pp. 202–208. DOI: 10.1016/j.siny.2013.04.004.

[10] Jeffrey W. Britton et al. *Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants*. 2016.

[11] Jan Brogger et al. "Visual EEG reviewing times with SCORE EEG." In: *Clinical Neurophysiology Practice* 3 (2018), pp. 59–64. DOI: 10.1016/j.cnp.2018.03.002.

[12] Sarah G. Buttle et al. "Continuous electroencephalography monitoring for critically ill neonates: A Canadian perspective." In: *Canadian Journal of Neurological Sciences* 46.4 (2019), pp. 394–402. DOI: 10.1017/cjn.2019.36.

[13] Abdullah Caliskan and Suleyman Rencuzogullari. “Transfer learning to detect neonatal seizure from electroencephalography signals.” In: *Neural Computing and Applications* 33.18 (2021), pp. 12087–12101. DOI: 10.1007/s00521-021-05878-y.

[14] Patrick Celka and Paul Colditz. “A computer-aided detection of EEG seizures in infants: A singular-spectrum approach and performance comparison.” In: *IEEE Transactions on Biomedical Engineering* 49.5 (2002), pp. 455–462. DOI: 10.1109/10.995684.

[15] Robert R. Clancy. “Interictal sharp EEG transients in neonatal seizures.” In: *Journal of Child Neurology* 4.1 (1989), pp. 30–38. DOI: 10.1177/088307388900400105.

[16] Aengus Daly et al. “Towards deeper neural networks for neonatal seizure detection.” In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2021, pp. 920–923. DOI: 10.1109/EMBC46164.2021.9629485.

[17] Alexander P. Dawid and Allan M. Skene. “Maximum likelihood estimation of observer error-rates using the EM algorithm.” In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28.1 (1979), pp. 20–28. DOI: 10.2307/2346806.

[18] Deborah Discenza. “Neuro-NICUs: Nurturing the tiniest of brains.” In: *Neonatal Network* 34.5 (2015), pp. 291–293. DOI: 10.1891/0730-0832.34.5.291.

[19] Piotr J. Durka et al. “A simple system for detection of EEG artifacts in polysomnographic recordings.” In: *IEEE Transactions on Biomedical Engineering* 50.4 (2003), pp. 526–528. DOI: 10.1109/TBME.2003.809476.

[20] Johanna Eicher et al. “A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models.” In: *BMC Medical Informatics and Decision Making* 20.1 (2020), pp. 1–14. DOI: 10.1186/s12911-020-1041-3.

[21] Raffaele Falsaperla et al. “aEEG vs cEEG’s sensivity for seizure detection in the setting of neonatal intensive care units: A systematic review and meta-analysis.” In: *Acta Paediatrica* 111.5 (2022), pp. 916–926. DOI: 10.1111/apa.16251.

[22] Jean Feng et al. “Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare.” In: *npj Digital Medicine* 5.1 (2022), p. 66. DOI: 10.1038/s41746-022-00611-y.

[23] U.S. Food and Drug Administration. *Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan*. 2021.

[24] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures.” In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS ’15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

[25] Benoit Frenay and Michel Verleysen. “Classification in the presence of label noise: A survey.” In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5 (2014), pp. 845–869. DOI: 10.1109/TNNLS.2013.2292894.

[26] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning.” In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. Proceedings of Machine Learning Research. PMLR, 2016, pp. 1050–1059.

[27] Gaetano Gargiulo et al. “A new EEG recording system for passive dry electrodes.” In: *Clinical Neurophysiology* 121.5 (2010), pp. 686–693. DOI: 10.1016/j.clinph.2009.12.025.

[28] Aisling A. Garvey et al. “Does early cerebral near-infrared spectroscopy monitoring predict outcome in neonates with hypoxic ischaemic encephalopathy? A systematic review of diagnostic test accuracy.” In: *Neonatology* 119.1 (2021), pp. 1–9. DOI: 10.1159/000518687.

[29] Erin Gibson et al. “EEG variability: Task-driven or subject-driven signal of interest?” In: *NeuroImage* 252 (2022), p. 119034. DOI: 10.1016/j.neuroimage.2022.119034.

[30] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks.” In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Vol. 15. Proceedings of Machine Learning Research. PMLR, 2011, pp. 315–323.

[31] Camila González et al. “Lifelong nnU-Net: a framework for standardized medical continual learning.” In: *Scientific Reports* 13.1 (2023), p. 9381. DOI: 10.1038/s41598-023-34484-2.

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[33] Jean Gotman. “Automatic detection of epileptic seizures.” In: *Presurgical Assessment of the Epilepsies with Clinical Neurophysiology and Functional Imaging*. Vol. 3. Handbook of Clinical Neurophysiology. Elsevier, 2003, pp. 155–165. DOI: 10.1016/S1567-4231(03)03012-0.

[34] Jean Gotman et al. “Automatic seizure detection in the newborn: Methods and initial evaluation.” In: *Electroencephalography and Clinical Neurophysiology* 103.3 (1997), pp. 356–362. DOI: 10.1016/S0013-4694(97)00003-9.

[35] Artur Gramacki and Jarosław Gramacki. “A deep learning framework for epileptic seizure detection based on neonatal EEG signals.” In: *Scientific Reports* 12.1 (2022), p. 13010. DOI: 10.1038/s41598-022-15830-2.

[36] Alexandre Gramfort et al. “MEG and EEG data analysis with MNE-Python.” In: *Frontiers in Neuroscience* 7 (2013). DOI: 10.3389/fnins.2013.00267.

[37] Barry R. Greene et al. “A comparison of quantitative EEG features for neonatal seizure detection.” In: *Clinical Neurophysiology* 119.6 (2008), pp. 1248–1261. DOI: 10.1016/j.clinph.2008.02.001.

[38] Barry R. Greene et al. “Combination of EEG and ECG for improved automatic neonatal seizure detection.” In: *Clinical Neurophysiology* 118.6 (2007), pp. 1348–1359. DOI: 10.1016/j.clinph.2007.02.015.

[39] Sukhi Grewal and Jean Gotman. “An automatic warning system for epileptic seizures recorded on intracerebral EEGs.” In: *Clinical Neurophysiology* 116.10 (2005), pp. 2460–2472. DOI: 10.1016/j.clinph.2005.05.020.

[40] Jiuxiang Gu et al. “Recent advances in convolutional neural networks.” In: *Pattern Recognition* 77 (2018), pp. 354–377. DOI: 10.1016/j.patcog.2017.10.013.

[41] Chuan Guo et al. "On calibration of modern neural networks." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1321–1330.

[42] Georg Hansmann. "Absolute and relative indications for neonatal transport and NICU admission." In: *Neonatal Emergencies*. Cambridge University Press, 2009, pp. 179–180. DOI: 10.1017/CBO9781139010467.033.

[43] Charles R. Harris et al. "Array programming with NumPy." In: *Nature* 585.7825 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[44] Hamid Hassanpour, Mostefa Mesbah, and Boualem Boashash. "Time–frequency based newborn EEG seizure detection using low and high frequency signatures." In: *Physiological Measurement* 25.4 (2004), p. 935. DOI: 10.1088/0967-3334/25/4/012.

[45] Trevor Hastie et al. *The elements of statistical learning: Data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[46] Kaiming He et al. "Deep residual learning for image recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[47] Tim Hermans et al. "A multi-task and multi-channel convolutional neural network for semi-supervised neonatal artefact detection." In: *Journal of Neural Engineering* 20.2 (2023), p. 026013. DOI: 10.1088/1741-2552/acbc4b.

[48] Gregory L. Holmes and Faye Korteling. "Drug Effects on the Human EEG." In: *American Journal of EEG Technology* 33.1 (1993), pp. 1–26. DOI: 10.1080/00029238.1993.11080427.

[49] Richard A. Hrachovy and Eli M. Mizrahi. *Atlas of neonatal electroencephalography*. Springer Publishing Company, 2015.

[50] David H. Hubel and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." In: *The Journal of physiology* 160.1 (1962), p. 106. DOI: 10.1113/jphysiol.1962.sp006837.

[51] Lucia Hug et al. "National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: A systematic analysis." In: *The Lancet Global Health* 7.6 (2019), e710–e720. DOI: 10.1016/S2214-109X(19)30163-9.

[52] Aatif M. Husain. "Review of neonatal EEG." In: *American Journal of Electroneurodiagnostic Technology* 45.1 (2005), pp. 12–35. DOI: 10.1080/1086508X.2005.11079505.

[53] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning." In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2127–2136.

[54] Natus Medical Incorporated. *K123079: Olympic Brainz Monitor. 510K Summary*. 2013.

[55] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 448–456.

[56] Dmitry Y. Isaev et al. “Attention-based network for weak labels in neonatal seizure detection.” In: *Proceedings of machine learning research* 126 (2020), pp. 479–507.

[57] Justin M. Johnson and Taghi M. Khoshgoftaar. “Survey on deep learning with class imbalance.” In: *Journal of Big Data* 6.1 (2019), pp. 1–54. DOI: 10.1186/s40537-019-0192-5.

[58] Amirhossein Kiani et al. “Impact of a deep learning assistant on the histopathologic classification of liver cancer.” In: *npj Digital Medicine* 3.1 (2020), p. 23. DOI: 10.1038/s41746-020-0232-8.

[59] Diederik P. Kingma and Jimmy Ba. *Adam: A method for stochastic optimization*. 2017. arXiv: 1412.6980.

[60] Masaomi Kitayama et al. “Wavelet analysis for neonatal electroencephalographic seizures.” In: *Pediatric Neurology* 29.4 (2003), pp. 326–333. DOI: 10.1016/S0887-8994(03)00277-7.

[61] Johannes Koren et al. “Systematic analysis and comparison of commercial seizure-detection software.” In: *Epilepsia* 62.2 (2021), pp. 426–438. DOI: 10.1111/epi.16812.

[62] Mustafa A. Kural et al. “Accurate identification of EEG recordings with interictal epileptiform discharges using a hybrid approach: Artificial intelligence supervised by human experts.” In: *Epilepsia* 63.5 (2022), pp. 1064–1073. DOI: 10.1111/epi.17206.

[63] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.” In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[64] Marie D. Lamblin and Anne de Villepin Touzery. “EEG in the neonatal unit.” In: *Neurophysiologie Clinique/Clinical Neurophysiology* 45.1 (2015), pp. 87–95. DOI: 10.1016/j.neucli.2014.11.007.

[65] Vi T. Le et al. “Neonatal antiepileptic medication treatment patterns: A decade of change.” In: *American Journal of Perinatology* 38.05 (2021), pp. 469–476. DOI: 10.1055/s-0039-1698457.

[66] Mikhail Lebedev, Ioan Opris, and Manuel F. Casanova. *Augmentation of brain function: Facts, fiction and controversy. Volume I: Brain-machine interfaces*. Frontiers Media SA, 2018, pp. 394–398.

[67] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning.” In: *nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[68] Yann LeCun et al. “Handwritten digit recognition with a back-propagation network.” In: *Advances in Neural Information Processing Systems*. Vol. 2. Morgan-Kaufmann, 1989.

[69] A. Liu et al. “Detection of neonatal seizures through computerized EEG analysis.” In: *Electroencephalography and Clinical Neurophysiology* 82.1 (1992), pp. 30–37. DOI: 10.1016/0013-4694(92)90179-L.

[70] Aileen Malone et al. “FC21.3 Ability of medical personnel to accurately differentiate neonatal seizures from non-seizure movements.” In: *Clinical Neurophysiology* 117.Supplement 1 (2006), p. 1. DOI: 10.1016/j.clinph.2006.06.070.

[71] Scott M. McKinney et al. “International evaluation of an AI system for breast cancer screening.” In: *Nature* 577.7788 (2020), pp. 89–94. DOI: 10.1038/s41586-019-1799-6.

[72] Saeed M. Moghadam et al. “An automated bedside measure for monitoring neonatal cortical activity: A supervised deep learning-based electroencephalogram classifier with external cohort validation.” In: *The Lancet Digital Health* 4.12 (2022), e884–e892. DOI: 10.1016/S2589-7500(22)00196-0.

[73] Saeed M. Moghadam et al. “Sleep state trend (SST), a bedside measure of neonatal sleep state fluctuations based on single EEG channels.” In: *Clinical Neurophysiology* 143 (2022), pp. 75–83. DOI: 10.1016/j.clinph.2022.08.022.

[74] Lakshmi Nagarajan, Soumya Ghosh, and Linda Palumbo. “Ictal electroencephalograms in neonatal seizures: Characteristics and associations.” In: *Pediatric Neurology* 45.1 (2011), pp. 11–16. DOI: 10.1016/j.pediatrneurol.2011.01.009.

[75] Soundharya Nagasubramanian, Banu Onaral, and Robert R. Clancy. “On-line neonatal seizure detection based on multi-scale analysis of EEG using wavelets as a tool.” In: *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 3. 1997, pp. 1289–1292. DOI: 10.1109/IEMBS.1997.756611.

[76] Vinod Nair and Geoffrey Hinton. “Rectified linear units improve restricted Boltzmann machines.” In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Omnipress, 2010, pp. 807–814.

[77] Michael A. Navakatikyan et al. “Seizure detection algorithm for neonates based on wave-sequence analysis.” In: *Clinical Neurophysiology* 117.6 (2006), pp. 1190–1203. DOI: 10.1016/j.clinph.2006.02.016.

[78] Alexandru Niculescu Mizil and Rich Caruana. “Predicting good probabilities with supervised learning.” In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML ’05. Association for Computing Machinery, 2005, pp. 625–632. DOI: 10.1145/1102351.1102430.

[79] Jeremy Nixon et al. “Measuring calibration in deep learning.” In: *CVPR Workshops*. Vol. 2. 7. 2019.

[80] Curtis Northcutt, Anish Athalye, and Jonas Mueller. *Pervasive label errors in test sets destabilize machine learning benchmarks*. 2021. arXiv: 2103.14749.

[81] Curtis Northcutt, Lu Jiang, and Isaac Chuang. “Confident learning: Estimating uncertainty in dataset labels.” In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 1373–1411. DOI: 10.1613/jair.1.12125.

[82] Mark E. O’Sullivan et al. “Development of an EEG artefact detection algorithm and its application in grading neonatal hypoxic-ischemic encephalopathy.” In: *Expert Systems with Applications* 213 (2023), p. 118917. DOI: 10.1016/j.eswa.2022.118917.

[83] Alison O'Shea et al. "Investigating the Impact of CNN Depth on Neonatal Seizure Detection Performance." In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 5862–5865. DOI: 10.1109/EMBC.2018.8513617.

[84] Alison O'Shea et al. "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture." In: *Neural Networks* 123 (2020), pp. 12–25. DOI: 10.1016/j.neunet.2019.11.023.

[85] Rawad Obeid and Tammy N. Tsuchida. "Treatment effects on neonatal EEG." In: *Journal of Clinical Neurophysiology* 33.5 (2016), pp. 376–381. DOI: 10.1097/WNP.0000000000000300.

[86] Benedetta Olmi et al. "Automatic detection of epileptic seizures in neonatal intensive care units through EEG, ECG and video recordings: A survey." In: *IEEE Access* 9 (2021), pp. 138174–138191. DOI: 10.1109/ACCESS.2021.3118227.

[87] Adam Paszke et al. "PyTorch: An imperative style, high-performance deep learning library." In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[88] Elena Pavlidis, Rhodri O. Lloyd, and Geraldine B. Boylan. "EEG - A valuable biomarker of brain injury in preterm infants." In: *Developmental Neuroscience* 39.1-4 (2017), pp. 23–35. DOI: 10.1159/000456659.

[89] Serena Pellegrin et al. "Neonatal seizures: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data." In: *Vaccine* 37.52 (2019), p. 7596. DOI: 10.1016/j.vaccine.2019.05.031.

[90] Ronit M. Pressler et al. "The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE task force on neonatal seizures." In: *Epilepsia* 62.3 (2021), pp. 615–628. DOI: 10.1111/epi.16815.

[91] Khadijeh Raeisi et al. "A class-imbalance aware and explainable spatio-temporal graph attention network for neonatal seizure detection." In: *International Journal of Neural Systems* (2023), p. 2350046. DOI: 10.1142/s0129065723500466.

[92] Khadijeh Raeisi et al. "A graph convolutional neural network for the automated detection of seizures in the neonatal EEG." In: *Computer Methods and Programs in Biomedicine* 222 (2022), p. 106950. DOI: 10.1016/j.cmpb.2022.106950.

[93] Sumit A. Raurale et al. "Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time–frequency distributions." In: *Journal of Neural Engineering* 18.4 (2021), p. 046007. DOI: 10.1088/1741-2552/abe8ae.

[94] Mary A. Ryan et al. "An introduction to neonatal EEG." In: *The Journal of perinatal & neonatal nursing* 35.4 (2021), pp. 369–376. DOI: 10.1097/JPN.0000000000000599.

[95] Amanda G. Sandoval Karamian and Courtney J. Wusthoff. "Current and future uses of continuous EEG in the NICU." In: *Frontiers in Pediatrics* (2021), p. 1254. DOI: 10.3389/fped.2021.768670.

[96] Nina Schwalbe and Brian Wahl. "Artificial intelligence and the future of global health." In: *The Lancet* 395.10236 (2020), pp. 1579–1586. DOI: 10.1016/S0140-6736(20)30226-9.

[97] Gauri S. Shah et al. “Clinical profile and outcome of neonates admitted to neonatal intensive care unit (NICU) at a tertiary care centre in Eastern Nepal.” In: *Journal of Nepal Paediatric Society* 33.3 (2013), pp. 177–181.

[98] Cynthia M. Sharpe et al. “Assessing the feasibility of providing a real time response to seizures detected with continuous long term neonatal EEG monitoring.” In: *Journal of Clinical Neurophysiology* 36.1 (2019), p. 9. DOI: `10.1097/WNP.0000000000000525`.

[99] Renée A. Shellhaas et al. “The American Clinical Neurophysiology Society’s guideline on continuous electroencephalography monitoring in neonates.” In: *Journal of Clinical Neurophysiology* 28.6 (2011), pp. 611–617. DOI: `10.1097/WNP.0b013e31823e96d7`.

[100] Liesbeth S. Smit et al. “Neonatal seizure monitoring using non-linear EEG analysis.” In: *Neuropediatrics* 35.06 (2004), pp. 329–335. DOI: `10.1055/s-2004-830367`.

[101] Hwanjun Song et al. “Learning from noisy labels with deep neural networks: A survey.” In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), pp. 1–19. DOI: `10.1109/TNNLS.2022.3152527`.

[102] Beena G. Sood, Kathleen McLaughlin, and Josef Cortez. “Near-infrared spectroscopy: Applications in neonates.” In: *Seminars in Fetal and Neonatal Medicine* 20.3 (2015), pp. 164–172. DOI: `10.1016/j.siny.2015.03.008`.

[103] David F. Steiner et al. “Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer.” In: *The American Journal of Surgical Pathology* 42.12 (2018), p. 1636. DOI: `10.1097/PAS.0000000000001151`.

[104] Nathan Stevenson, Leena Lauronen, and Sampsa Vanhatalo. “The effect of reducing EEG electrode number on the visual interpretation of the human expert for neonatal seizure detection.” In: *Clinical Neurophysiology* 129.1 (2018), pp. 265–270. DOI: `10.1016/j.clinph.2017.10.031`.

[105] Nathan Stevenson, Karoliina T. Tapani, and Sampsa Vanhatalo. “Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert.” In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 5991–5994. DOI: `10.1109/EMBC.2019.8857367`.

[106] Nathan Stevenson et al. “A dataset of neonatal EEG recordings with seizure annotations.” In: *Scientific Data* 6 (), p. 190039. DOI: `10.1038/sdata.2019.39`.

[107] Nathan Stevenson et al. “Artefact detection in neonatal EEG.” In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 926–929. DOI: `10.1109/EMBC.2014.6943743`.

[108] Nathan Stevenson et al. “Interobserver agreement for neonatal seizure detection using multichannel EEG.” In: *Annals of Clinical and Translational Neurology* 2.11 (2015), pp. 1002–1011. DOI: `10.1002/acn3.249`.

[109] Moritz Tacke et al. “Effects of a reduction of the number of electrodes in the EEG montage on the number of identified seizure patterns.” In: *Scientific Reports* 12.1 (2022). DOI: `10.1038/s41598-022-08628-9`.

[110] Joshua D. Tao and Amit Mathur. “Using amplitude-integrated EEG in neonatal intensive care.” In: *Journal of Perinatology* 30.1 (2010), S73–S81. DOI: `10.1038/jp.2010.93`.

[111] Karoliina T. Tapani, Sampsa Vanhatalo, and Nathan Stevenson. "Time-varying EEG correlations improve automated neonatal seizure detection." In: *International Journal of Neural Systems* 29.04 (2019), p. 1850030. DOI: 10.1142/S0129065718500302.

[112] Maria L. Tataranno et al. "Precision medicine in neonates: A tailored approach to neonatal brain injury." In: *Frontiers in Pediatrics* 9 (2021), p. 634092. DOI: 10.3389/fped.2021.634092.

[113] Hasan Tekgul et al. "Electroencephalography in neonatal seizures: Comparison of a reduced and a full 10/20 montage." In: *Pediatric Neurology* 32.3 (2005), pp. 155–161. DOI: 10.1016/j.pediatrneurol.2004.09.014.

[114] Andriy Temko and Gordon Lightbody. "Detecting neonatal seizures with computer algorithms." In: *Journal of Clinical Neurophysiology* 33.5 (2016), pp. 394–402. DOI: 10.1097/WNP.0000000000000295.

[115] Andriy Temko et al. "EEG-based neonatal seizure detection with support vector machines." In: *Clinical Neurophysiology* 122.3 (2011), pp. 464–473. DOI: 10.1016/j.clinph.2010.06.034.

[116] Andriy Temko et al. "Performance assessment for EEG-based neonatal seizure detectors." In: *Clinical Neurophysiology* 122.3 (2011), pp. 474–482. DOI: https://doi.org/10.1016/j.clinph.2010.06.035.

[117] Michal Teplan. "Fundamental of EEG Measurement." In: *Measurment Science Review* 2 (2002).

[118] Sunil Thulasidasan et al. "On mixup training: Improved calibration and predictive uncertainty for deep neural networks." In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[119] Mona C. Toet and Petra M. A. Lemmers. "Brain monitoring in neonates." In: *Early Human Development* 85.2 (2009), pp. 77–84. DOI: 10.1016/j.earlhumdev.2008.11.007.

[120] Sana Tonekaboni et al. "What clinicians want: Contextualizing explainable machine learning for clinical end use." In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. Vol. 106. Proceedings of Machine Learning Research. PMLR, 2019, pp. 359–380.

[121] Cristina Uria Avellanal, Neil Marlow, and Janet M. Rennie. "Outcome following neonatal seizures." In: *Seminars in Fetal and Neonatal Medicine*. Vol. 18. 4. Elsevier. 2013, pp. 224–232. DOI: 10.1016/j.siny.2013.01.002.

[122] "Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy." In: *Computers in Biology and Medicine* 145 (2022), p. 105399. DOI: 10.1016/j.compbiomed.2022.105399.

[123] Guido Van Rossum and Fred L. Drake. *Python 3 reference manual*. CreateSpace, 2009. ISBN: 1441412697.

[124] Sampsa Vanhatalo and Kai Kaila. "Development of neonatal EEG activity: From phenomenology to physiology." In: *Seminars in Fetal and Neonatal Medicine*. Vol. 11. 6. Elsevier. 2006, pp. 471–478. DOI: 10.1016/j.siny.2006.07.008.

[125] Sampsa Vanhatalo et al. "Why monitor the neonatal brain—that is the important question." In: *Pediatric Research* 93.1 (2023), pp. 19–21. DOI: 10.1038/s41390-022-02040-9.

[126] Gabriel Fernando Todeschi Variane et al. "Current status and future directions of neuromonitoring with emerging technologies in neonatal care." In: *Frontiers in Pediatrics* 9 (2022). DOI: 10.3389/fped.2021.755144.

[127] Chakrapani Vasudevan and Malcolm Levene. "Epidemiology and aetiology of neonatal seizures." In: *Seminars in Fetal and Neonatal Medicine*. Vol. 18. 4. Elsevier. 2013, pp. 185–191. DOI: 10.1016/j.siny.2013.05.008.

[128] Ashish Vaswani et al. "Attention is all you need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[129] Carmen Verwoerd et al. "Efficacy of Levetiracetam and Phenobarbital as first-line treatment for neonatal seizures." In: *Journal of Child Neurology* 37.5 (2022), pp. 401–409. DOI: 10.1177/08830738221086l0.

[130] Pauli Virtanen et al. "SciPy 1.0: Fundamental algorithms for scientific computing in Python." In: *Nature methods* 17.3 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[131] K. Visalini, Saravanan Alagarsamy, and Deivanayagam Nagarajan. "Neonatal seizure detection using deep belief networks from multichannel EEG data." In: *Neural Computing and Applications* (2023), pp. 1–11. DOI: 10.1007/s00521-023-08254-0.

[132] Kerstin N. Vokinger, Stefan Feuerriegel, and Aaron S. Kesselheim. "Continual learning in medical devices: FDA's action plan and beyond." In: *The Lancet Digital Health* 3.6 (2021), e337–e338. DOI: 10.1016/S2589-7500(21)00076-5.

[133] Lachlan Webb et al. "Automated detection of artefacts in neonatal EEG with residual neural networks." In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106194. DOI: 10.1016/j.cmpb.2021.106194.

[134] David H. Wolpert. "Stacked generalization." In: *Neural Networks* 5.2 (1992), pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.

[135] Lichao Xu et al. "Cross-dataset variability problem in EEG decoding with deep learning." In: *Frontiers in human neuroscience* 14 (2020), p. 103. DOI: 10.3389/fnhum.2020.00103.

[136] Elissa Yozawitz. "Neonatal seizures." In: *New England Journal of Medicine* 388.18 (2023), pp. 1692–1700. DOI: 10.1056/NEJMra2300188.

[137] Kun H. Yu, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare." In: *Nature Biomedical Engineering* 2.10 (2018), pp. 719–731. DOI: 10.1038/s41551-018-0305-z.

[138] Aston Zhang et al. *Dive into deep learning*. 2023. arXiv: 2106.11342.

[139] Chiyuan Zhang et al. "Understanding deep learning (still) requires rethinking generalization." In: *Communications of the ACM* 64.3 (2021), pp. 107–115. DOI: 10.1145/3446776.

[140] Jingwei Zhang et al. "Automatic annotation correction for wearable EEG based epileptic seizure detection." In: *Journal of Neural Engineering* 19.1 (2022), p. 016038. DOI: 10.1088/1741-2552/ac54c1.

[141] Yuheng Zhang et al. "The secret revealer: Generative model-inversion attacks against deep neural networks." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 253–261.

Paper I

Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network

Ana Borovac, Steinn Gudmundsson, Gardar Thorvardsson, and Thomas P. Runarsson

NeurIPS Data-Centric AI workshop, 2021

Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network

Ana Borovac[*]
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
University of Iceland
Reykjavik, Iceland
anb48@hi.is

Steinn Guðmundsson
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
University of Iceland
Reykjavik, Iceland
steinng@hi.is

Gardar Thorvardsson
Kvikna Medical ehf.
Reykjavik, Iceland
gardar@kvikna.com

Thomas Philip Runarsson
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
University of Iceland
Reykjavik, Iceland
tpr@hi.is

Abstract

Neonatal seizures are common among infants and can be detected with an electroencephalogram (EEG). The EEG signals are complex time-series using multiple channels. Human domain experts are often in disagreement when labelling neonatal seizure data. Only few studies will include labels from multiple experts, as annotating hours of EEG recordings is time consuming and expensive. In this study we investigated the differences in performance of a deep-learning-based neonatal seizure detector trained using single expert labelling versus data labelled using the consensus of multiple experts. Results indicate that there are differences even when the experts are in minor disagreement. We find that excluding ambiguously labeled data is important when training a neonatal seizure detector.

1 Introduction

Seizures are common among infants, with a prevalence of 1 – 5 per thousand live births [4]. Since untreated seizures can cause brain damage [1], it is paramount to detect them early. Seizure detection in infants is complicated by the fact that the majority of seizures cannot be observed clinically [2]. The

[*] Kvikna Medical ehf., Reykjavik, Iceland

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

current gold standard for neonatal seizure detection (NSD) is a multi-channel electroencephalogram (EEG) recording with simultaneous video, analyzed by a human expert [14]. The frequency and duration of seizures within an EEG are of clinical interest.

EEGs are time-series that represent the electrical activity of the brain. Neonatal EEG recordings are usually obtained with 4 – 20 electrodes that are placed on the scalp and last from a few hours to days. Analysis of an EEG requires extensive training and is time consuming which hampers widespread use. Automating the procedure is therefore of obvious clinical significance. The measurements have high inter- and intra-patient variability, the EEG is highly dependant on the age of the neonate, its condition [7, 8] and medication [6, 12]. Non-cerebral artifacts such as heartbeat, breathing and infant care frequently contaminate the signal and may mimic seizure activity. Due to the complexity of neonatal EEG signals, human experts are often in disagreement [11], in particular when seizures are short in duration [15].

Even though human experts provide the gold standard neonatal seizure labels, label noise is likely to be present in the training data which can have a negative effect on the performance of a machine learning model [18]. To the best of our knowledge there are only a few studies in the field of neonatal seizure detection addressing label noise by utilizing multiple human-expert labels [11, 13, 15, 17]. In this work we compare five strategies for utilizing labels from multiple human experts in the training of a NSD based on a deep convolutional neural network.

2 Methods

The data set used in the experiments contains segments from 79 neonatal EEG recordings, each approximately 1 hour in length, and accompanying labels from three human experts with 1 sec resolution [16]. The recordings contain 19 channels sampled at 256 Hz that were combined in a longitudinal montage (a frequently used pairwise combination of channels). The segments were split into 16 sec long blocks with 12 sec overlap. The signals were filtered with a 6th order Chebyshev Type 2 band-pass filter with cut-off frequencies of 0.5 Hz and 32 Hz, down-sampled to 32 Hz and standardised so that the mean and standard deviation were zero and one, respectively. Each 16 sec interval was labeled as a seizure or a non-seizure interval per human expert (A, B or C), the majority vote and consensus amongst experts, resulting in five sets of labelings. Ambiguous segments, i.e. segments that were partly labeled as seizure and partly as non-seizure, were excluded. Figure 1 illustrates scoring for a typical EEG segment and the total number of seizure/non-seizure segments is given in table 1. Non-seizure segments were approximately 8 times as many as the seizure segments. The non-seizure segments were therefore randomly sub-sampled to obtain balanced training sets. One network (NSD) was trained for each of the five labelings in table 1.

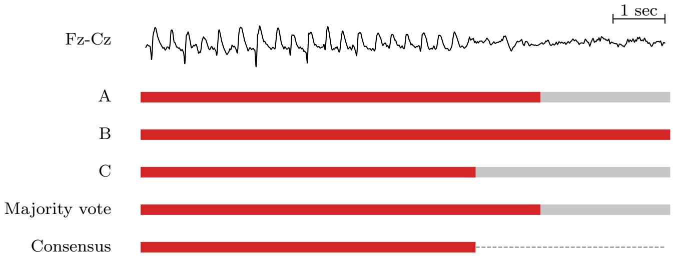


Figure 1: 10 sec EEG segment (channel Fz-Cz), labeling from scorers A, B and C, majority vote and consensus labels. Seizure areas are annotated with red, non-seizure with grey and ignored parts with dashed grey line.

A convolutional neural network proposed by Stevenson et al. [17] was used as a feature extractor. It consists of 10 convolutional layers with 32 filters of size 3 and one convolutional layer with 2 filters of size 3. Each convolutional layer is followed by a batch normalization layer and ReLU activation. Before the fourth, seventh and tenth convolutional layers, average pooling is applied with filters of size 8, 4 and 2 respectively. The stride was set to 3 for all three pooling layers. The feature extractor

Table 1: The total number of seizure and non-seizure segments available for each labeling; human experts (A, B and C), majority vote and consensus labels. The number of seizure and non-seizure segments exclusive to each expert are in parentheses.

Labeling	Seizure	Non-seizure
A	10482 (332)	85075 (619)
B	14170 (2129)	81266 (401)
C	11127 (1043)	83511 (394)
Majority vote	11658	84847
Consensus	8560	78260

is followed by an attention layer [9] and a fully connected layer with two output neurons and softmax activation.

Cross entropy was used as a loss function and the model parameters were optimized using the Adam optimizer with a mini-batch size of 128. The learning rate was set to 0.001 in the beginning and halved every 10 epochs. The model was trained for 40 epochs. Experiments using 30 and 50 epochs gave similar results (data not shown). A fixed number of epochs was used during training due to the prohibitive computational cost of using leave-one-patient-out cross-validation for parameter tuning.

Each of the five models were tested on labelings from experts A, B and C to investigate whether a model trained on labels from a single expert, under- or over-performs models trained on labels from the other experts in any significant way. The models were also tested on the consensus labels. The models were evaluated by leaving one subject out at a time to avoid train-test set overlap. There are 38 patients with at least one 16 sec long consensus seizure segment in the data set [16] and the results report below are based on data from these 38 patients. Cohen's kappa (κ) was used as the performance metric instead of ROC AUC since the test set was highly imbalanced and clinical utility of a NSD does not necessarily follow from a high AUC value [9].

The code used in the experiments was written in Python using PyTorch 1.7.1 and executed on a NVIDIA GeForce GTX 1080 Ti GPU.

3 Results and discussion

The main results are presented in figure 2. The figure shows that all the models performed poorly (i.e. low kappa values) on a small subset of patients. The poor performance is partly caused by the relatively small training set and high inter-patient variability. Some of the recordings have very few seizure or non-seizure segments which means that the performance metric is very sensitive to predictions from these segments.

Experts often disagree on the exact start and end times of seizures. They disagree also on seizures that are shorter than 30 sec in duration [15]. The consensus set excludes these segments, resulting in seizure segments that are in a sense “clean”. This appears to be beneficial since the model trained on the consensus labels performs best overall (figure 2). The mean kappa values are between 0.52 and 0.61 for the NSD trained with consensus data.

The NSD trained with labels from expert B performs worst, irrespective of the test set. Table 1 shows that this expert labeled 27 % - 35 % more segments as seizures than experts A and C. Some of these additional seizure segments are confusing the classifier, leading to an increased number of false seizure predictions. This led to higher sensitivity and lower specificity (table 2).

Training on labels from expert A resulted in a model that performed the best, out of the three models trained on labels from a single expert. Expert A annotated the least number of exclusive segments (table 1) and agreed with at least one of the other two experts for most parts of the EEG recordings.

4 Conclusion

The experiments show that NSD performance can depend strongly on the expert responsible for scoring the EEG, as the results for expert B clearly show. The results from expert B also show

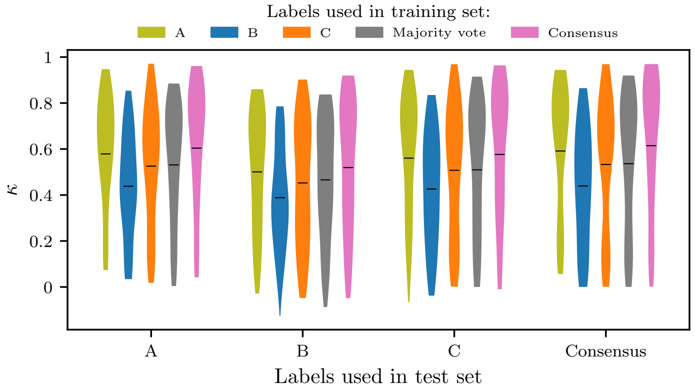


Figure 2: Comparison of Cohen's kappa (κ) values of models trained using different labels illustrated by the different colours. Results are compared with different test labels. Solid lines denote the mean values.

Table 2: Mean sensitivity and specificity values for different training/test labels.

	Test labels							
	Sensitivity [%]				Specificity [%]			
Training labels	A	B	C	Consensus	A	B	C	Consensus
A	76.77	67.55	75.93	80.51	89.96	91.22	90.90	92.28
B	79.12	71.41	77.38	80.74	81.93	82.73	82.87	83.85
C	75.94	66.85	73.69	78.30	88.04	88.76	88.92	90.08
Majority vote	78.68	70.19	76.36	80.80	86.91	87.96	87.89	89.16
Consensus	75.15	66.19	73.51	78.47	91.62	92.61	92.37	93.68

significant differences compared to the model using the majority vote in the training set. Improvement in classifier performance due to using majority vote of multiple domain experts has previously been observed in a study on prostate cancer classification [10].

When labels from multiple experts are available, using consensus labels can reduce label noise and improve the overall accuracy of the NSD. This is in agreement with previous findings on other types of data [18]. It further indicates that if the data labels are close to being noise-free, a clinically relevant NSD can be obtained even when the training set is relatively small. For comparison, kappa values calculated between the human experts over the entire data set were in the range 0.63 to 0.73.

Models trained on labels from a single expert did not result in models that captured the criteria the experts used to identify seizure segments. Explanations include the model architecture not capturing all the information an expert uses to determine the absense/presence of seizures. When scoring an EEG, experts frequently inspect segments that occur earlier or later in the recording. This behaviour is not captured by the convolutional network used here. Another explanation could be inattentional blindness [3]. However, there does not exist an absolute truth in EEG recordings, comparable to biopsies in skin cancer detection [5] and mistakes can not be easily confirmed.

To conclude, when using labels from one human expert it must be kept in mind that the labels are subjective to the expert and the performance of a model is highly dependent on the expert labelling the data. Therefore, when training a NSD it is important to reduce the label noise by excluding segments with ambiguous labels.

Acknowledgments and Disclosure of Funding

Authors thank Nathan J Stevenson for sharing the architecture of the neural network feature extractor and the pre-processing of data. This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 813483.

References

[1] Stella T Björkman, Stephanie M Miller, Stephen E Rose, Christopher Burke, and Paul B Colditz. Seizures are associated with brain injury severity in a neonatal model of hypoxia–ischemia. *Neuroscience*, 166(1): 157–167, 2010.

[2] Geraldine B Boylan, Nathan J Stevenson, and Sampsa Vanhatalo. Monitoring neonatal seizures. In *Seminars in Fetal and Neonatal Medicine*, volume 18, pages 202–208. Elsevier, 2013.

[3] Trafton Drew, Melissa L-H Võ, and Jeremy M Wolfe. The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological science*, 24(9):1848–1853, 2013.

[4] Hannah C Glass, Courtney J Wusthoff, Renée A Shellhaas, Tammy N Tsuchida, Sonia Lomeli Bonifacio, Malaika Cordeiro, Joseph Sullivan, Nicholas S Abend, and Taeun Chang. Risk factors for EEG seizures in neonates treated with hypothermia: a multicenter cohort study. *Neurology*, 82(14):1239–1244, 2014.

[5] Achim Hekler, Jakob N Kather, Eva Krieghoff-Henning, Jochen S Utikal, Friedegund Meier, Frank F Gellrich, Julius Upmeier zu Belzen, Lars French, Justin G Schlager, Kamran Ghoreschi, et al. Effects of label noise on deep learning-based skin cancer classification. *Frontiers in Medicine*, 7:177, 2020.

[6] Gregory L Holmes and Faye Korteling. Drug effects on the human EEG. *American Journal of EEG Technology*, 33(1):1–26, 1993.

[7] Richard A Hrachovy and Eli M Mizrahi. *Atlas of neonatal electroencephalography*. Springer Publishing Company, 2015.

[8] Aatif M Husain. Review of neonatal EEG. *American journal of electroneurodiagnostic technology*, 45(1): 12–35, 2005.

[9] Dmitry Yu Isaev, Dmitry Tchapyjnikov, C Michael Cotten, David Tanaka, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and David Carlson. Attention-based network for weak labels in neonatal seizure detection. *Proceedings of machine learning research*, 126:479, 2020.

[10] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[11] Aileen Malone, C Anthony Ryan, Anthony Fitzgerald, Louise Burgoyne, Sean Connolly, and Geraldine B Boylan. Interobserver agreement in neonatal seizure identification. *Epilepsia*, 50(9):2097–2101, 2009.

[12] Rawad Obeid and Tammy N Tsuchida. Treatment effects on neonatal EEG. *Journal of Clinical Neurophysiology*, 33(5):376–381, 2016.

[13] 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.

[14] Ronit M Pressler, Maria Roberta Cilio, Eli M Mizrahi, Solomon L Moshé, Magda L Nunes, Perrine Plouin, Sampsa Vanhatalo, Elissa Yozawitz, Linda S de Vries, Kollencheri Puthenveettil Vinayan, et al. The ilae classification of seizures and the epilepsies: Modification for seizures in the neonate. position paper by the ilae task force on neonatal seizures. *Epilepsia*, 62(3):615–628, 2021.

[15] Nathan J Stevenson, Robert R Clancy, Sampsa Vanhatalo, Ingmar Rosén, Janet M Rennie, and Geraldine B Boylan. Interobserver agreement for neonatal seizure detection using multichannel EEG. *Annals of clinical and translational neurology*, 2(11):1002–1011, 2015.

[16] Nathan J Stevenson, Karoliina Tapani, Leena Lauronen, and Sampsa Vanhatalo. A dataset of neonatal EEG recordings with seizure annotations. *Scientific data*, 6:190039, 2019.

[17] Nathan J Stevenson, Karoliina Tapani, and Sampsa Vanhatalo. Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5991–5994. IEEE, 2019.

[18] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3):177–210, 2004.

Paper II

Ensemble learning using individual neonatal data for seizure detection

Ana Borovac, Steinn Gudmundsson, Gardar Thorvardsson, Saeed M. Moghadam, Päivi Nevalainen, Nathan Stevenson, Sampsa Vanhatalo, and Thomas P. Runarsson

IEEE Journal of Translational Engineering in Health and Medicine, 2022

HEALTHCARE

IEEE Journal of Translational Engineering in Health and Medicine

Received 7 April 2022; revised 6 July 2022; accepted 18 August 2022.
Date of publication 23 August 2022; date of current version 30 August 2022.

Digital Object Identifier 10.1109/JTEHM.2022.3201167

Ensemble Learning Using Individual Neonatal Data for Seizure Detection

ANA BOROVAC[1,2], STEINN GUDMUNDSSON[1], GARDAR THORVARDSSON[2],
SAEED M. MOGHADAM [3], (Graduate Student Member, IEEE), PÄIVI NEVALAINEN[3,4],
NATHAN STEVENSON[5], (Member, IEEE), SAMPSA VANHATALO[3], AND THOMAS P. RUNARSSON[1]

[1]Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, 107 Reykjavik, Iceland
[2]Kvikna Medical ehf., 110 Reykjavik, Iceland
[3]BABA Center, Pediatric Research Center, Department of Physiology, University of Helsinki, 00014 Helsinki, Finland
[4]HUS Diagnostic Center, Epilepsia Helsinki and Department of Clinical Neurophysiology, New Children's Hospital, Helsinki University Hospital, 00029 Helsinki, Finland
[5]Brain Modelling Group, QIMR Berghofer Medical Research Institute, Herston, QLD 4006, Australia

CORRESPONDING AUTHOR: A. BOROVAC (anb48@hi.is)

This work was supported in part by the Sigrid Juselius Foundation, and in part by the European Union's Horizon 2020 Research and Innovation Programme under Agreement 813483.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Research Review Board of the HUS Diagnostic Center.

ABSTRACT Objective: Sharing medical data between institutions is difficult in practice due to data protection laws and official procedures within institutions. Therefore, most existing algorithms are trained on relatively small electroencephalogram (EEG) data sets which is likely to be detrimental to prediction accuracy. In this work, we simulate a case when the data can not be shared by splitting the publicly available data set into disjoint sets representing data in individual institutions. Methods and procedures: We propose to train a (local) detector in each institution and aggregate their individual predictions into one final prediction. Four aggregation schemes are compared, namely, the majority vote, the mean, the weighted mean and the Dawid-Skene method. The method was validated on an independent data set using only a subset of EEG channels. Results: The ensemble reaches accuracy comparable to a single detector trained on all the data when sufficient amount of data is available in each institution. Conclusion: The weighted mean aggregation scheme showed best performance, it was only marginally outperformed by the Dawid–Skene method when local detectors approach performance of a single detector trained on all available data. Clinical impact: Ensemble learning allows training of reliable algorithms for neonatal EEG analysis without a need to share the potentially sensitive EEG data between institutions.

INDEX TERMS Convolutional neural network, distributed learning, ensemble learning, neonatal EEG, seizure detection algorithm.

I. INTRODUCTION

Seizures are common during perinatal period [1], and management of neonatal seizures requires timely detection and treatment to reduce ensuing brain damage [2]. The current gold standard for neonatal seizure detection is visual analysis by a human expert using a full-montage video electroencephalogram (EEG) [3]. Since such service is rarely available in neonatal intensive care units (NICUs), there is an urgent clinical need for automated neonatal seizure detection algorithm (NSDA) with human expert level accuracy.

Early automated NSDAs were based on *features*, quantitative descriptors of short, e.g. $10-16$ sec long, EEG segments and expert-defined threshold decision rules [4], [5], [6]. Hard-coded thresholds were later replaced by statistical techniques, such as linear discriminant analysis [7], support vector machines (SVMs) [8], [9], [10] and neural networks [11]. Recently, promising results have been obtained using convolutional neural networks (CNNs) [12], [13], [14].

Deep neural networks (DNNs) generally require a large amount of training data [15]. However, building a large and diverse enough neonatal EEG data set with high quality seizure annotations is time consuming, ambiguous [16], [17] and often limited due to strict regulations (e.g. the Privacy Rule of the U.S. Health Insurance Portability and Accountability Act (HIPAA), or the European General Data Protection Regulation (GDPR)) making data sharing between institutions difficult, if not impossible [18], [19]. Challenges in sharing data have triggered growing interest in distributed approaches to statistical learning [20].

One approach that requires minimal sharing of information is model ensembling, i.e. models are trained locally at each institution and predictions on new data are aggregated (ensembled) from predictions made by the local models. This requires sharing only the models across the network of institutions rather than sharing the potentially sensitive, original biosignals. However, the procedures in model sharing need to be planned so that they mitigate the impact of possible inadvertent leaks of training data through a model [21], [22]. One solution to this problem is to have a *trusted agent* in charge of the models and an aggregation procedure. Compared to the federated learning [23], ensembling does not require communication between the institutions during the training phase (which may be difficult to set up) and it does not require the institutions to use the same model architecture. One institution could e.g. use a DNN, another an SVM and a third a decision tree classifier.

Once predictions on new data have been made there are a number of techniques by which they can be ensembled. If predictions are accompanied by probabilities they can be averaged [24], [25], if not, a commonly used method for label aggregation is to simply select the most frequent label, referred to as *majority vote* in the following. One could also put more weight on some predictions if they are a priori more trustworthy, otherwise, an estimate of each annotator performance can be used [26], [27], [28]. Dawid and Skene [29] used an expected maximization (EM) algorithm [30] to estimate annotator performance and provide consensus labels.

Ensemble learning has previously been used in neonatal seizure detection. In [31] stacking is used where different model types trained on the same data are combined. In [32] three identical NSDAs are trained on the same EEG data but using labels from different experts. In this work we use ensemble learning on disjoint data sets, to simulate the situation where institutions train NSDAs on locally available data. Depending on the training data available at each institution and its similarity to new data to be labelled, the local NSDAs are expected to vary in performance. The main contribution and novelty of this work is in the discovery of how such locally trained models can be aggregated with the aim of achieving performance comparable to a single state-of-the-art NSDA trained on the union of all local training data sets. For aggregation we compared the majority vote, the mean, the weighted mean (via stacking) and the Dawid–Skene expected maximization algorithm. We show that the weighted mean outperforms the other methods if the NSDAs in the ensemble are trained on very few patients and Dawid-Skene marginally outperforms the other methods when the local NSDAs are not much worse than the state-of-the-art NSDA. The NSDAs and ensembles are further validated on an independent data set consisting of more than 2100 hours of EEG recorded from a small subset of the channels used to train the classifiers.

II. METHODS AND PROCEDURES

Multiple local models, referred to as *local NSDAs* in the following, are trained on disjoint subsets of multi-channel EEG recordings, simulating a scenario where several hospitals train NSDAs individually, without sharing patient data. The trained detectors are then shared with a trusted agent. To classify a short EEG segment from a new patient as seizure/non-seizure, the trusted agent sends the segment through all the local NSDAs and the predictions are aggregated using one of the following schemes: majority vote, mean, weighted mean or the Dawid–Skene method. The methodology is summarized in figure 1.

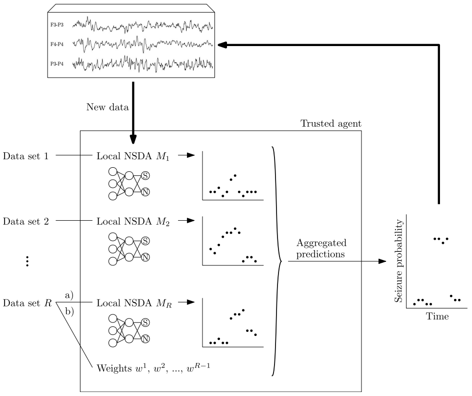


FIGURE 1. A schematic diagram of the proposed method. Each data set is used to train a local NSDAs or weights that are shared with a trusted agent. The trusted agent makes predictions on new data. Seizure predictions for new data are obtained a) by aggregating predictions made by *R* NSDAs using the majority vote, the mean or the Dawid-Skene method, or, b) by aggregating predictions made by *R* − 1 local NSDAs using the weighted mean (weights are learned on the *R*th data set).

For local NSDAs, we used DNNs which take EEG segments as input. The networks share the same architecture but have different network weights since they were trained on disjoint training sets.

A. AGGREGATION SCHEMES

In the following we consider a binary classification problem where the classes are labeled 0 and 1. Let D be a set of N predictions from R independent models

$$D = \left\{ \left(p_1^1, p_1^2, \ldots, p_1^R\right), \ldots, \left(p_N^1, p_N^2, \ldots, p_N^R\right) \right\},$$

where p_i^j is the estimated probability of model j of instance i belonging to class 1. By setting a threshold between the classes to 0.5, the predicted label of model j of instance i is given by

$$y_i^j = \begin{cases} 1; & \text{if } p_i^j \geq 0.5, \\ 0; & \text{otherwise.} \end{cases}$$

A simple way to aggregate multiple predictions for instance i, when models do not output their confidence (e.g. class probabilities), is to use majority vote, i.e. select the most frequent label. Here we use the mean of predicted labels,

$$\mu_i^{MV} = \frac{1}{R}\sum_{j=1}^{R} y_i^j; \quad i \in \{1, 2, \ldots, N\}. \tag{1}$$

When the models output class probabilities, which is e.g. the case when the models correspond to the neural networks, the predictions can be aggregated by taking the mean probability,

$$\mu_i^{M} = \frac{1}{R}\sum_{j=1}^{R} p_i^j; \quad i \in \{1, 2, \ldots, N\}. \tag{2}$$

As some of the models might perform better than others, a weighted mean can be used to emphasize the more accurate models. To get the final prediction in a range between 0 and 1, we used logistic regression,

$$\mu_i^{WM} = \sigma\left(\sum_{j=1}^{R} w^j p_i^j\right); \quad i \in \{1, 2, \ldots, N\}, \tag{3}$$

where $\sigma(x) = {}^{1}/_{(1+e^{-x})}$. The weights for w^j are learned on a held out data set (see section II-D).

The fourth aggregation method evaluated here is the Dawid–Skene method. The method estimates the sensitivity and specificity of each model, together with consensus predictions μ^{DS}. For details of the method see appendix A. To predict the absence/presence of seizures from the above aggregation schemes, a threshold of 0.5 is used.

B. DATA

The EEG data used to train the NSDAs is a publicly available data set containing 79 approximately one hour long neonatal EEG recordings, measured with 19 Ag/AgCl electrodes positioned according to the 10-20 system [33]. An 18 channel montage is used, i.e. we derive channels Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz and Cz-Pz. The recordings are annotated by three EEG experts where each second in a recording is annotated as a seizure or non-seizure. We refer to this data set as 18-channel DS below.

The second, proprietary, data set (the 3-channel DS) consisting of EEG recordings of 28 neonates, is used as a held out test set to evaluate the aggregation schemes in a real world setting, i.e. detectors are trained on the 18-channel DS and tested on this data set. The data set is also used in [34] and is a subset of the data set used in [35]. Institutional Research Review Board of the HUS diagnostic center approved the use of this data, including a waiver of consent due to the study's retrospective and observational nature. Each recording spans from 19 hours to 7 days. The recordings were obtained using 4 needle electrodes (F3, F4, P3 and P4) with a common reference, instead of the full set of 19 electrodes used in the training data set. Neonatal recordings are typically performed with this reduced electrode set to allow easier maintenance in a long duration brain monitoring [36]. The three bipolar derivations (F3-P3, F4-P4 and P3-P4) are used for both two human expert annotators and as the detectors input.

Additional attributes of the data sets are given in table 2 in appendix B.

Each EEG recording is cut into 16 sec long segments with 12 sec overlap. Out of the 79 (28) recordings in 18-channel DS (the 3-channel DS), 38 (24) contain at least one seizure longer than 16 sec identified by three (two) human experts, meaning each of these recordings contain at least one consensus seizure segment. Segments containing more than 1 sec of zero voltage interval in at least one channel (disconnected electrode or pause in the recording) are left-out from the training and test sets. The signals are filtered with a 6th order Chebyshev Type 2 band-pass filter with cut-off frequencies of 0.5 Hz and 16 Hz, down-sampled to 32 Hz and rescaled to 16-bit integers. This is similar to the preprocessing in [10] and [13].

C. NEONATAL SEIZURE DETECTION ALGORITHM

Each NSDA is a neural network consisting of three components; a feature extractor, an attention layer and an output layer. The feature extractor is a CNN from [37]. The features are extracted from each EEG channel separately and are combined into a single feature channel by the attention layer [13]. The attention layer is used since expert labels are not specific to individual channels and neonatal seizures tend to be partial [3], i.e. localized in a small area of the brain and therefore only present in a subset of the recorded channels. The attention layer is also independent of the number of input feature channels making the detector independent of the number of recorded EEG channels. The output layer is a fully connected layer with two output nodes representing the two classes. A detailed description of the network architecture is given in appendix C.

To compare the aggregation schemes to current state-of-the-art NSDAs, we trained a neural network using all the recordings in the 18-channel DS containing at least one consensus seizure longer than 16 sec (P). This NSDA is referred to as the *baseline NSDA* in the following.

The local NSDAs use the same neural network architecture as the baseline NSDA but differ in the data used for training. The patients in P (patients containing a consensus seizure) are partitioned into $k = 3, 4, \ldots, 10$ subsets representing data sets in individual institutions. Partitioning is random such that each patient is in exactly one subset and there are at least three patients in every subset. The union of the k subsets is then P, the data set used as a training set for the baseline NSDA. By excluding patients without consensus seizures we ensure each subset has patients with seizures and eliminate the varying number of EEGs with normal brain activity in individual subsets, making the analysis more straightforward. As there can be a big difference between the training set sizes, we obtain local NSDAs with different generalisation strengths and consequently with different performance strengths on unseen data. This is expected in practice. Even though the acquisition equipment is subject to international standards and the electrodes are positioned according to the 10-20 system, the EEG signals may vary considerably depending on the patient cohorts as the signals differ between neonates of different ages and conditions [38], [39]. Therefore, the detectors are expected to perform differently on unseen data.

D. TRAINING

After partitioning the training set, each NSDA (baseline NSDA and local NSDAs) is trained on 16 sec long EEG segments corresponding to the consensus seizures and non-seizure segments. To avoid complications due to class imbalance [13], [40], the training sets are balanced prior to training by sub-sampling the non-seizure segments. Segments with disagreements between the human experts and partly seizure/non-seizure segments are not included in the training sets. Cross entropy is used as the loss function. The Adam optimizer is used to optimize the network weights using an initial learning rate of 0.001 which is then halved every 10 epochs. The NSDAs are trained for 30 epochs with a mini-batch size of 32. Hyper-parameters, learning rate and number of epochs, are tuned empirically, from observing the behavior of the loss function during the training of the baseline NSDA. A small mini-batch size is chosen due to a small amount of data used in some local NSDAs. For the weighed mean aggregation scheme, the weights $w^j, j \in \{1, 2, \ldots, R\}$, are learned using a stacking classifier [28]. A logistic regression classifier is trained using the data from one randomly selected local NSDA in each experiment. This local NSDA is not used in an ensemble for making predictions on a test patient. Therefore, non-overlapping data sets are used for training the local NSDAs and the logistic regression classifier. Also, the training data of the local NSDAs would not need to be shared in practice as the input of the logistic regression classifier is just a set of seizure probabilities estimated by the local NSDAs and these can be provided by the trusted agent.

All the deep learning code used in the experiments is implemented using PyTorch 1.7.1 [41] and run on an NVIDIA GTX 1080 Ti GPU. For logistic regression, we use the scikit-learn [42] implementation with default hyper-parameters. The code is available at github.com/anaborovac/Distributed-NSDA.

E. PERFORMANCE

To avoid overlap between training and test data when evaluating classifier performance on the 18-channel DS, leave-one-subject-out cross-validation is used. This entailed training 38 baseline NSDAs, 38 sets of local NSDAs and 38 sets of logistic regression classifiers, leaving out data from one subject (patient) at a time. The experiment is repeated 10 times, resulting in $10 \cdot 38 \cdot (3+4+\cdots+10) = 19760$ local NSDAs and $10 \cdot 38 \cdot (1+1+\cdots+1) = 10 \cdot 38 \cdot 8 = 3040$ logistic regression classifiers.

Data from each left-out patient is sent through the corresponding baseline NSDA and local NSDAs. Predictions from the baseline NSDAs are compared to human expert labels to obtain performance metrics. Predictions from the local NSDAs are first aggregated using one of the aforementioned aggregation schemes: majority vote (1), mean (2), weighted mean (3) and the Dawid–Skene method (appendix A) to obtain the final predictions and these are then compared to human expert labels.

Two sets of performance metrics are calculated, metrics based on the success/failure in classifying individual 16 sec long segments, and event-based metrics which indicate whether a seizure is detected at all, or whether a seizure is falsely reported. The segment-based metrics are sensitivity (SE), specificity (SP) and the area under the receiver operating characteristic curve (AUC). These metrics are calculated from segments without disagreements between human experts and segments with either seizure either non-seizure activity for the whole segment duration. The event-based metrics are seizure detection rate (SDR), false detections per hour (FD/h) and the mean false detection duration (MFDD) [43]. A consensus seizure is considered to be detected if it is detected at any point in time and a seizure is considered as a false detection if it did not overlap with any (consensus or not) seizure labelled by the human experts. Definitions of the metrics are provided in appendix D. Metrics calculated on each patient separately are summarized by their means and medians.

Before the event-based metrics are calculated a post-processing step is in order since segments overlap. Besides a few segments at the beginning and end of each recording, for each 4 sec long segment there are 4 overlapping 16 sec long segments. Prediction for a 4 sec segment is obtained by averaging predictions from overlapping 16 sec long segments [44], [45]. Seizures with duration less than 10 sec are excluded and considered normal brain activity as by definition seizures are longer than 10 sec [46].

For studying the segment-based level of agreement between the local NSDAs we use Gwet's first-order agreement coefficient (AC1) [47]. Compared to the often used Cohen's (Fleiss') κ [13], [48], [49], Gwet's AC1 is less prone to the paradoxes associated with highly imbalanced data [50], [51].

Performance on the 3-channel DS is evaluated in the same manner as for the 18-channel DS, i.e. the metrics are calculated for each patient separately and then summarized with the mean and the median. The baseline NSDA is trained using all 38 patients in P (no patients are left-out), and the union of the training sets for the local NSDAs also contain all 38 patients in P. This results in additional $1 + 10 \cdot (3 + 4 + \cdots + 10) = 521$ NSDAs and $10 \cdot (1 + 1 + \cdots + 1) = 10 \cdot 8 = 80$ logistic regression classifiers.

III. RESULTS

To assess the clinical usefulness of the aggregation schemes they are compared to a baseline NSDA which is trained on data from all 38 patients in P (in a leave-one-subject-out setting for evaluation on the 18-channel DS). The baseline NSDA thus corresponds to the situation where a single agent has access to all the training data (P), a situation which is expected to be favorable compared to aggregating predictions from multiple models trained on disjoint subsets of the same data.

A. BASELINE NSDA

Table 1 compares the performance of the baseline detector to other NSDAs found in the literature. All detectors are neural networks and were trained or tested using the 18-channel DS. The difference between the mean (0.92) and median (0.98) AUC values for the baseline NSDA calculated on the 18-channel DS is mainly due to the presence of respiratory and heart rate artefacts and low seizure burden in some of the recordings.

TABLE 1. Comparison of the area under the curve (AUC) values found in the literature. Each reference uses a different proprietary data set. All NSDAs, except [13], were trained using the 18-channel DS. Superscript L denotes leave-one-subject-out testing and superscript C denotes AUC value on concatenated recordings from the data set.

		AUC	
		18-channel DS	Proprietary DS
Isaev et al. [13]	mean	0.92	0.97^{L}
O'Shea et al. [14]	mean	0.96^{C}	0.99^{L}
Stevenson et al. [49]	median	0.99^{L}	
Baseline NSDA	median	0.98^{L}	0.93
	mean	0.92^{L}	0.92

The performance of an NSDA on an independent test set is usually worse than performance estimates obtained from a held out training data. Such a decrease can be attributed to several factors, including differences in patient cohorts, seizure prevalence, the number of available EEG channels, the human experts that annotated the EEG [48], and training data not representing the general population. For example, the mean AUC decreased from 0.97 to 0.92 in [13] and from 0.99 to 0.96 in [14]. We observe a similar drop in performance when the baseline detector was tested on a proprietary the 3-channel DS. Detailed validation of the NSDA performance is available in table 3 in appendix E.

In summary, the baseline NSDA gives comparable results to the state-of-the-art NSDAs and performs well on recordings which include only a small subset of the channels used in training.

B. AGGREGATION SCHEMES

Here we evaluate the different aggregation schemes and compare them to the baseline NSDA and to the average performance of the local NSDAs. If the baseline performance can be reached with an aggregation scheme, it would indicate that the data does not need to be shared during the training of an NSDA to obtain a detector with state-of-the-art performance. The four aggregation schemes, majority vote, mean, weighted mean and the Dawid–Skene method were evaluated on the 18-channel DS and the 3-channel DS for $k = 3, 4, \ldots, 10$ local NSDAs. Results for the majority vote are not shown since in all cases majority vote was slightly outperformed by the mean aggregation scheme (see figure 7 in appendix E).

With an increasing number of local NSDAs the average performance of an individual detector gradually gets worse (figure 2). This is explained by the fact that the number of patients behind each local NSDA is becoming smaller since the total number of patients in the combined training sets is constant (37 for the 18-channel DS and 38 for the 3-channel DS). Consequently there is an increased risk of overfitting in individual detectors. The size of the local training sets is quantified with the mean median number of patients in the training set. E.g., if four local NSDAs are used and the mean median is 8.1, then on average there are at least nine patients in the training of two of the local NSDAs.

Figure 2 shows that the AUC, seizure detection rate and false detection rate behave similarly across both data sets for all the aggregation schemes, but there is considerably more variability for the 3-channel DS. All the aggregation schemes give AUC values that are similar to the baseline value. However, the aggregation schemes differ in terms of seizure detection rate and false detections per hour.

Figure 3 shows the seizure probability estimates returned by local NSDAs for an hour-long recording, together with probability estimates obtained with the ensemble methods. All the aggregation schemes result in AUC close to one, although they detect only 3 out of 7 consensus seizures. The missed seizures are short in duration and they are clearly visible in the figure (as white bands) since the corresponding probabilities are higher than for the non-seizure segments.

The SDR in figure 2 behaves similarly for both data sets. For all values of k tested, the Dawid–Skene method is comparable to the baseline NSDA, while for the mean and the weighted mean aggregation schemes, fewer seizures were detected with an increased number of local NSDAs. Recall that when there are few NSDAs, each NSDA detects almost as many seizures as the baseline detector. The mean aggregation scheme performed slightly worse than the weighted mean and both performed notably worse than the Dawid–Skene method for more than four local detectors. Moreover, the average SDR of the local NSDAs is comparable to the values corresponding to the mean aggregation scheme. With the weighted mean a larger number of seizures are detected for $k \geq 8$ ($k = 10$) on the 18-channel DS (3-channel DS), for smaller k the mean and the weighted mean aggregation schemes return comparable seizure detection rates.

Moreover, in figure 2 we observe that all aggregation schemes result in a lower number of FD/h than the average local NSDA. The average FD/h of the local NSDAs are noticeably higher for the 3-channel DS than for the

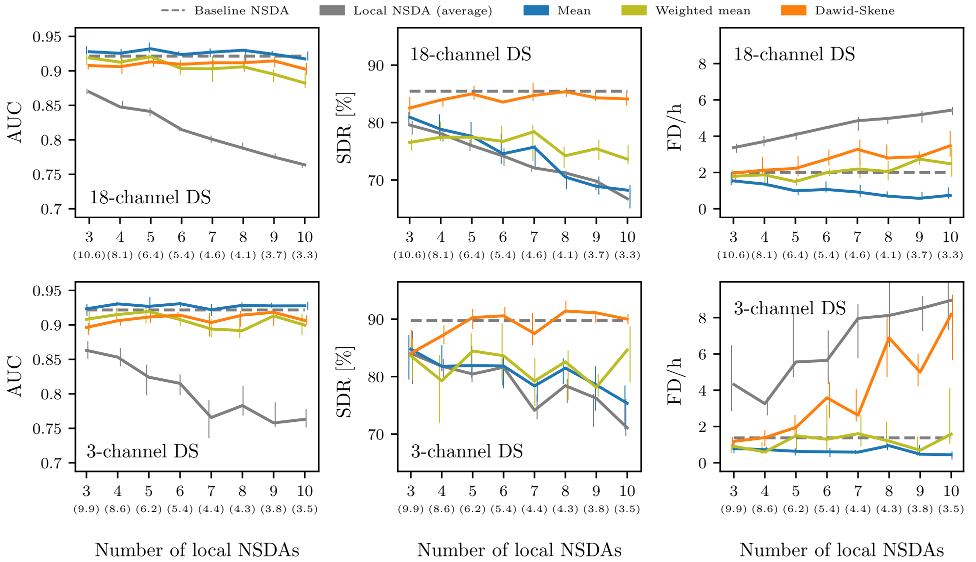


FIGURE 2. Average area under the curve (AUC), seizure detection rate (SDR) and false detections per hour (FD/h) as a function of the number of local NSDAs used in the aggregation schemes. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines. The grey dashed line represents the average metric of the baseline NSDA. The average (across ten runs) mean median number of patients in each NSDA is shown in parentheses.

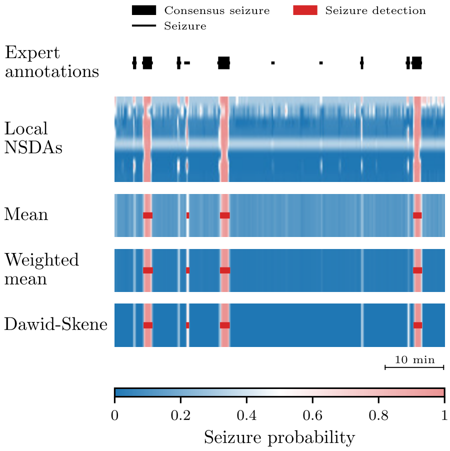


FIGURE 3. An example of aggregated predictions from eight local NSDAs. The area under the curve is 1.0 for the mean and the weighted mean and 0.99 for the Dawid–Skene method. All aggregation schemes detect 42.9 % of consensus seizures and they do not falsely detect any seizure.

18-channel DS. One possible explanation is that the recordings in the 3-channel DS are much longer and on average just 3.5 % of a recording corresponds to a seizure activity. The mean aggregation scheme has a lower false detection rate than the baseline NSDA and the FD/h decreases steadily with increasing number of local NSDAs. This may be a result of low level of agreement between the local NSDAs for the large k (figure 5 in appendix E). So, even though an individual local NSDA falsely detects a large number of seizures, the aggregated prediction filtered them out or was below the 0.5 threshold. This may on the other hand caused problems with the Dawid–Skene method, i.e. the FD/h increased slowly on the 18-channel DS and rapidly on the 3-channel DS with increasing number of local NSDAs. In contrast, the logistic regression classifier determining the weights for the weighted mean aggregation scheme successfully detected local NSDAs with high/low false detection rate for all k tested.

We observed low false detection rates for the mean and weighted mean aggregation schemes and therefore investigated whether the false detections are short or long in duration. We did not observe big differences between the aggregation schemes (10 - 30 sec) and different values of local NSDAs (figure 6 in appendix E).

To summarise, all aggregation schemes tested here perform better than the average local NSDA and are comparable to the baseline NSDA for $k \in \{3, 4\}$. This shows that the overfitting by local models noted earlier is offset by aggregating their predictions. This is in line with published reports on ensemble methods such as Random Forests which aggregate predictions from multiple models individually overfitting the data. The decrease in performance for larger values of k is mainly a result of training the local NSDAs on smaller training sets that do not capture the general population. The (weighted) mean aggregation scheme detects fewer seizures than the baseline detector, however the false detection rate is comparable, if not lower. The Dawid–Skene method successfully detects the same number of seizures as the baseline NSDA for any number of local NSDAs, but the false detection rate is compromised for $k \geq 6$. Predictions obtained with the Dawid–Skene are difficult to explain [52], [53], only a few local NSDAs with poor performance may have caused unexpected and undesired aggregated prediction [54].

IV. CONCLUSION

In this work we have shown that an NSDA based on a convolutional neural network together with an attention layer can accurately detect seizures, even if the data is obtained with different types of electrodes (scalp vs needle) and significantly lower number of channels than it was used for training. All the performance metrics of the NSDAs unsurprisingly dropped when training sets contained data from only a few patients. For aggregation of such NSDAs the weighted mean aggregation scheme performed best. Compared to the Dawid–Skene method, it successfully detected local NSDAs with high false detection rates and seizure detection rate was not as compromised as it was for the mean aggregation scheme. When a larger number of patients was included in the training of individual local NSDAs, i.e. when the number of local NSDAs was few, the Dawid–Skene method marginally outperformed the other aggregation schemes. It had a higher seizure detection rate and the false detections per hour was comparable to the (weighted) mean aggregation scheme. Independent of the number of local NSDAs, the majority vote was slightly outperformed by the mean aggregation scheme and all aggregation schemes performed better than the average individual (local) NSDA.

The experiments suggest that data does not need to be shared between institutions. It takes approx. 15 seconds to process one hour of 18-channel EEG with 10 local detectors, which is fast enough to be used in an online setting in the clinic. By utilizing GPU optimized code in the preprocessing steps and a fast version of the Dawid-Skene aggregation method [55], one hour of EEG could be processed in less than 2 seconds.

To confirm the findings reported here in a real-world setting, data from multiple institutions would be required. A large data set would also allow a detailed study on the number of local NSDAs needed to reach the desirable classification performance and whether a mixture of different types of NSDAs improves or degrades the overall performance.

APPENDIX A
DAWID-SKENE METHOD

The Dawid–Skene method was initially used to estimate the performance of human annotators [29]. Here the method is used to estimate the performance of models (local NSDAs) and obtain consensus judgement amongst them. The method is as follows. From a given set D of model predictions, the task is to estimate consensus labels $\{\mu_i\}_{i=1}^N$, the sensitivity α^j and specificity β^j of predictive model $j \in \{1, 2, \ldots, R\}$. Let Y_e denote the multivariate random variable

$$Y_e = (Y_1^1, Y_1^2, \ldots, Y_1^R, \ldots, Y_N^1, Y_N^2, \ldots, Y_N^R),$$

where random variable Y_i^j denotes the label given to instance i by model j. Furthermore, let T_i denote a random variable corresponding to the true label of instance i for which

$$P[T_i = 1] = t_i = t; \quad i \in \{1, 2, \ldots, N\}.$$

Assuming that model labels are independent and that conditional probability of Y_i^j on T_i follows Bernoulli distribution with parameters α^j and β^j, respectively:

$$\begin{aligned} a_i &= P_\alpha\left[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 1\right] \\ &= \prod_{j=1}^{R} (\alpha^j)^{y_i^j} (1-\alpha^j)^{1-y_i^j}; \quad i \in \{1, 2, \ldots, N\}, \\ b_i &= P_\beta\left[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 0\right] \\ &= \prod_{j=1}^{R} (\beta^j)^{1-y_i^j} (1-\beta^j)^{y_i^j}; \quad i \in \{1, 2, \ldots, N\}. \end{aligned}$$

To simplify the notation, let $\theta = (t, \alpha, \beta)$ denote the parameters to be estimated. Assuming that instances are sampled independently, the likelihood function for Y_e is [29], [56]:

$$\begin{aligned} P_\theta[Y_e] &= \prod_{i=1}^{N} P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R] \\ &= \prod_{i=1}^{N} \Bigg(\underbrace{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 1]}_{a_i} \underbrace{P_\theta[T_i = 1]}_{t} \\ &\qquad + \underbrace{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 0]}_{b_i} \underbrace{P_\theta[T_i = 0]}_{1-t} \Bigg) \\ &= \prod_{i=1}^{N} \left(a_i t + b_i(1-t)\right). \end{aligned} \tag{4}$$

Dawid and Skene used the EM algorithm to identify a local maximum of the likelihood function. The true labels are estimated by maximizing the likelihood function using estimated values for the sensitivity and specificity of each annotator, and the prior probability of class 1 (t), i.e. seizure. The algorithm has two main steps [29].

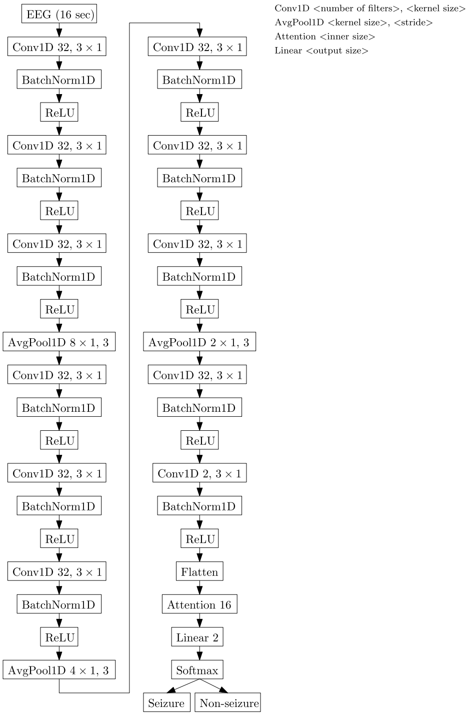


FIGURE 4. Architecture of the NSDA with a total of 29352 learnable parameters. Other parameters were set to default PyTorch values.

Expectation step: calculate the expected value of a true label knowing labels made by predictive models,

$$
\begin{aligned}
\mu_i &= \mathbb{E}[T_i | Y_i^1, Y_i^2, \ldots, Y_i^R] \\
&= P_\theta[T_i = 1 | Y_i^1, Y_i^2, \ldots, Y_i^R] \\
&= \frac{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 1] P_\theta[T_i = 1]}{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R]} \\
&\qquad \text{(Bayes' theorem)} \\
&= \frac{a_i t}{a_i t + b_i(1-t)}; \quad i \in \{1, 2, \ldots, N\}. \qquad (5)
\end{aligned}
$$

Maximization step: estimate t, α^j and β^j that maximize the likelihood function (4),

$$
t = \frac{\sum_{i=1}^N \mu_i}{N}, \qquad (6)
$$

$$
\alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}; \quad j \in \{1, 2, \ldots, R\}, \qquad (7)
$$

$$
\beta^j = \frac{\sum_{i=1}^N (1-\mu_i)(1-y_i^j)}{\sum_{i=1}^N (1-\mu_i)}; \quad j \in \{1, 2, \ldots, R\}. \qquad (8)
$$

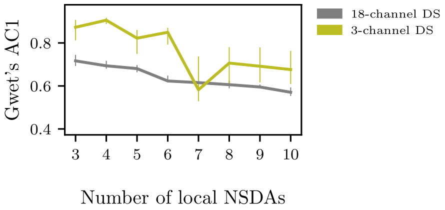


FIGURE 5. Average Gwet's AC1 between local NSDAs for 18-channel DS and 3-channel DS. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines.

TABLE 2. A summary of the data sets used in the study. Numbers inside parentheses represent standard deviation. Means for recordings are calculated across patients containing at least one consensus seizure longer than 16 sec (duration of one EEG segment).

	18-channel DS	3-channel DS
Number of patients	79	28
Number of patients with consensus seizures ≥ 16 sec	38	24
Gestational age (weeks)	39.3 (2.1)	39.2 (2.0)
Number of derived EEG channels	18	3
Total recordings duration (hours)	111.9	2149.4
Mean recording duration (hours)	1.4 (0.6)	76.4 (35.8)
Total number of consensus seizures	344	1387
Total duration of consensus seizures (hours)	11.0	65.3
Mean duration of consensus seizures (minutes)	1.9 (2.7)	2.8 (6.0)
Mean fraction of recording containing seizures (%)	31.8 (26.4)	5.3 (5.6)
Mean fraction of recording containing consensus seizures (%)	19.1 (20.9)	3.5 (4.0)

In the special case when all the μ_i's are either 0 or 1, then t is the estimated ratio of positive instances and α^j (β^j) is an estimated ratio of correctly predicted positive (negative) examples by expert j, i.e. the estimated sensitivity (specificity) of expert j.

```
Input: $D$, $\epsilon = 10^{-5}$, $k_{max} = 5000$
Output: $\mu^{DS}$
  initialize $\mu^{DS} = \mu^M$
  compute $\theta^{(0)}$ using equations (6), (7) and (8)
  $k = 0$
  repeat
    k = k + 1
    compute $\mu^{DS}$ using equation (5)
    compute $\theta^{(k)}$ using equations (6), (7) and (8)
  until $|\log P_{\theta^{(k-1)}}[Y_e] - \log P_{\theta^{(k)}}[Y_e]| < \epsilon \quad or \quad k \geq k_{max}$
```

APPENDIX B
DATA INFORMATION

See table 2.

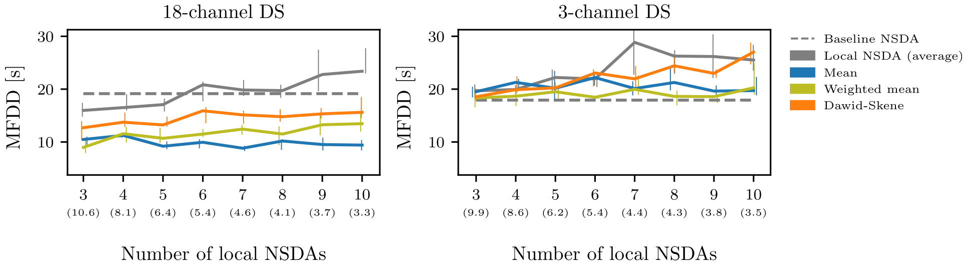


FIGURE 6. Average mean false detection duration (MFDD) as a function of the number of local NSDAs used in the aggregation schemes. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines. The grey dashed line represents the average MFDD of the baseline NSDA.

APPENDIX C ARCHITECTURE OF THE NSDA

In this work the NSDAs are deep neural networks consisted of three components, a feature extractor [37], an attention layer [13] and an output layer (figure 4). We used PyTorch implementation of layers for the feature extractor and for the output layer. Using PyTorch notation, the attention layer was implemented as follows. If an input to the attention layer is of size (N, C_{in}, L) then the output is of size (N, L) and can be described as

$$\mathrm{out}(N_i) = \sum_{k=0}^{C_{in}-1} a_k \mathrm{input}(N_i, k);$$

$$a_k = \frac{\exp\left(w^T \tanh\left(V \mathrm{input}(N_i, k)^T\right)\right)}{\sum_{j=0}^{C_{in}-1} \exp\left(w^T \tanh\left(V \mathrm{input}(N_i, j)^T\right)\right)},$$

where $V \in \mathbb{R}^{L\times \text{<inner size>}}$ and $w \in \mathbb{R}^{L\times 1}$ are learnable parameters.

APPENDIX D PERFORMANCE METRICS

A. SEGMENT-BASED METRICS

Segment-based metrics were calculated based on 16 sec long EEG segments. A true positive (TP) is a correctly predicted seizure segment, a true negative (TN) is a correctly predicted non-seizure segment, a false positive (FP) is an incorrectly predicted non-seizure segment and a false negative (FN) is an incorrectly predicted seizure segment.

- Sensitivity (ratio of correctly predicted seizure intervals):

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \cdot 100.$$

- Specificity (ratio of correctly predicted non-seizure intervals):

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \cdot 100.$$

- Area under the receiver operating characteristics curve (AUC). The receiver operating characteristics curve describes SE depending on 1-SP.

TABLE 3. Accuracy of the baseline model. Area under the curve (AUC), sensitivity (SE), specificity (SP), seizure detection rate (SDR), false detections per hour (FD/h) and mean false detection duration (MFDD) are computed as the mean and median over all the patients with seizures.

		Segment-based metrics		
		AUC	SE [%]	SP [%]
18-channel DS	median	0.98	90.46	97.21
	mean	0.92	79.52	93.69
3-channel DS	median	0.93	78.00	98.23
	mean	0.92	70.54	97.40
		Event-based metrics		
		SDR [%]	FD/h	MFDD [s]
18-channel DS	median	100.0	0.91	12.00
	mean	85.45	1.99	19.15
3-channel DS	median	95.55	0.97	15.82
	mean	89.77	1.37	17.92

B. EVENT-BASED METRICS

Event-based metrics are in comparison with the segment-based metrics focused on each predicted seizure and not just 16 sec long segments. Three event-based metrics were used [43]:

- Seizure detection rate (SDR):

$$\mathrm{SDR} = \frac{\mathrm{DS}}{\mathrm{CS}} \cdot 100,$$

where DS is a number of detected consensus seizures and CS is a number of consensus seizures. A seizure was considered to be detected if it was detected at any time of its duration.

- False detections per hour (FD/h):

$$\mathrm{FD/h} = \frac{\mathrm{IDS}}{\mathrm{D}},$$

where IDS is a number of incorrectly detected seizures and D is duration of data in hours. A seizure was considered to be incorrectly detected if it was not overlapping with any seizure annotated by the experts.

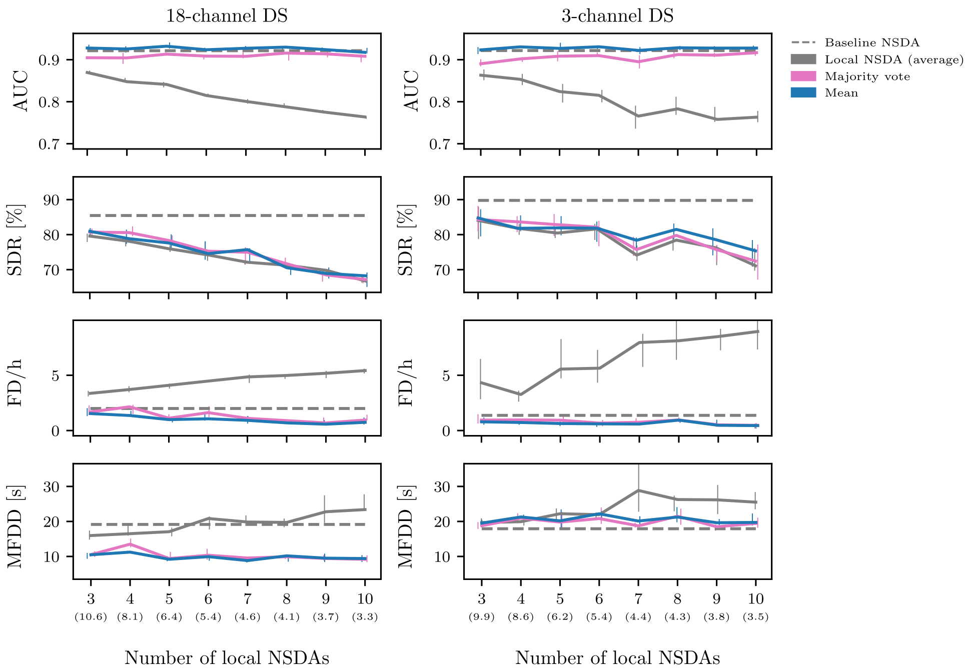


FIGURE 7. Average area under the curve (AUC), seizure detection rate (SDR), false detections per hour (FD/h) and false detection duration (MFDD) as a function of the number of local NSDAs used in the aggregation schemes. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines. The grey dashed line represents the average metric of the baseline NSDA. The average (across ten runs) mean median number of patients in each NSDA is shown in parentheses.

- Mean false detection duration (MFDD):

$$\text{MFDD} = \begin{cases} 0; & \text{if IDS} = 0 \\ \frac{\text{DIDS}}{\text{IDS}}; & \text{otherwise} \end{cases},$$

where DIDS is a sum of durations of incorrectly detected seizures in seconds and IDS is a number of incorrectly detected seizures.

APPENDIX E
ADDITIONAL RESULTS

See table 3 and figures. 6 and 7.

REFERENCES

[1] H. C. Glass *et al.*, "Risk factors for EEG seizures in neonates treated with hypothermia: A multicenter cohort study," *Neurology*, vol. 82, no. 14, pp. 1239–1244, Apr. 2014.

[2] S. T. Björkman, S. M. Miller, S. E. Rose, C. Burke, and P. B. Colditz, "Seizures are associated with brain injury severity in a neonatal model of hypoxia–ischemia," *Neuroscience*, vol. 166, no. 1, pp. 157–167, Mar. 2010.

[3] R. M. Pressler *et al.*, "The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE task force on neonatal seizures," *Epilepsia*, vol. 62, no. 3, pp. 615–628, Mar. 2021.

[4] A. Liu, J. S. Hahn, G. P. Heldt, and R. W. Coen, "Detection of neonatal seizures through computerized EEG analysis," *Electroencephalogr. Clin. Neurophysiol.*, vol. 82, no. 1, pp. 30–37, 1992.

[5] S. Nagasubramanian, B. Onaral, and R. Clancy, "On-line neonatal seizure detection based on multi-scale analysis of EEG using wavelets as a tool," in *Proc. 19th Annu. Int. Conf. IEEE Eng. Med. Biol. Society. Magnificent Milestones Emerg. Opportunities Med. Eng.*, vol. 3, Jun. 1997, pp. 1289–1292.

[6] M. A. Navakatikyan, P. B. Colditz, C. J. Burke, T. E. Inder, J. Richmond, and C. E. Williams, "Seizure detection algorithm for neonates based on wave-sequence analysis," *Clin. Neurophysiol.*, vol. 117, no. 6, pp. 1190–1203, Jun. 2006.

[7] B. R. Greene, S. Faul, W. P. Marnane, G. Lightbody, I. Korotchikova, and G. B. Boylan, "A comparison of quantitative EEG features for neonatal seizure detection," *Clin. Neurophysiol.*, vol. 119, no. 6, pp. 1248–1261, 2008.

[8] R. Ahmed, A. Temko, W. P. Marnane, and G. Lightbody, "Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel," *Comput. Biol. Med.*, vol. 82, pp. 100–110, Mar. 2017.

[9] A. H. Ansari *et al.*, "Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor," *Clin. Neurophysiol.*, vol. 127, no. 9, pp. 3014–3024, Sep. 2016.

[10] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 464–473, Mar. 2011.

[11] H. Hassanpour, M. Mesbah, and B. Boashash, "Time–frequency based newborn EEG seizure detection using low and high frequency signatures," *Physiol. Meas.*, vol. 25, no. 4, p. 935, 2004.

[12] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *Int. J. Neural Syst.*, vol. 29, no. 4, May 2019, Art. no. 1850011.

[13] D. Y. Isaev *et al.*, "Attention-based network for weak labels in neonatal seizure detection," *Proc. Mach. Learn. Res.*, vol. 126, p. 479, Aug. 2020.

[14] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture," *Neural Netw.*, vol. 123, pp. 12–25, Mar. 2020.

[15] Y. Bengio, I. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2016.

[16] A. Malone, C. A. Ryan, A. Fitzgerald, L. Burgoyne, S. Connolly, and G. B. Boylan, "Interobserver agreement in neonatal seizure identification," *Epilepsia*, vol. 50, no. 9, pp. 2097–2101, Sep. 2009.

[17] N. J. Stevenson, R. R. Clancy, S. Vanhatalo, I. Rosén, J. M. Rennie, and G. B. Boylan, "Interobserver agreement for neonatal seizure detection using multichannel EEG," *Ann. Clin. Translational Neurol.*, vol. 2, no. 11, pp. 1002–1011, Nov. 2015.

[18] J. Eicher, R. Bild, H. Spengler, K. A. Kuhn, and F. Prasser, "A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–14, Dec. 2020.

[19] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 13, no. 3, pp. 1–207, 2019.

[20] M. Kirienko *et al.*, "Distributed learning: A reliable privacy-preserving strategy to change multicenter collaborations using AI," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 48, no. 12, pp. 1–14, 2021.

[21] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.

[22] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 253–261.

[23] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[24] K. Chang *et al.*, "Distributed deep learning networks among institutions for medical imaging," *J. Amer. Med. Inform. Assoc.*, vol. 25, pp. 945–954, Aug. 2018.

[25] A. Tuladhar, S. Gill, Z. Ismail, and N. D. Forkert, "Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling," *J. Biomed. Informat.*, vol. 106, Jun. 2020, Art. no. 103424.

[26] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, "Domain-weighted majority voting for crowdsourcing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 163–174, Jan. 2018.

[27] T. Tian and J. Zhu, "Max-margin majority voting for learning from crowds," in *Proc. Nips*, 2015, pp. 1621–1629.

[28] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[29] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Statist. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.

[31] Y. Pan, H. Li, L. Liu, Q. Li, X. Hou, and B. Dong, "AEEG signal analysis with ensemble learning for newborn seizure detection," in *Proc. Int. Workshop Multiscale Multimodal Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 76–84.

[32] M. A. Tanveer, M. J. Khan, H. Sajid, and N. Naseer, "Convolutional neural networks ensemble model for neonatal seizure detection," *J. Neurosci. Methods*, vol. 358, Jul. 2021, Art. no. 109197.

[33] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizure annotations," *Sci. Data*, vol. 6, no. 1, Mar. 2019, Art. no. 190039.

[34] K. T. Tapani, P. Nevalainen, S. Vanhatalo, and N. J. Stevenson, "Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy," 2022, *arXiv:2202.12023*.

[35] P. Nevalainen, M. Metsäranta, S. Toiviainen-Salo, T. Lönnqvist, S. Vanhatalo, and L. Lauronen, "Bedside neurophysiological tests can identify neonates with stroke leading to cerebral palsy," *Clin. Neurophysiol.*, vol. 130, no. 5, pp. 759–766, May 2019.

[36] G. B. Boylan, N. J. Stevenson, and S. Vanhatalo, "Monitoring neonatal seizures," *Seminars Fetal Neonatal Med.*, vol. 18, no. 4, pp. 202–208, Aug. 2013.

[37] A. OrShea, G. Lightbody, G. Boylan, and A. Temko, "Investigating the impact of CNN depth on neonatal seizure detection performance," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 5862–5865.

[38] R. A. Hrachovy and E. M. Mizrahi, *Atlas of Neonatal Electroencephalography*. Cham, Switzerland: Springer, 2015.

[39] A. M. Husain, "Review of neonatal EEG," *Amer. J. Electroneurodiagnostic Technol.*, vol. 45, no. 1, pp. 12–35, Mar. 2005.

[40] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[41] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[43] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. B. Boylan, "Performance assessment for EEG-based neonatal seizure detectors," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 474–482, Mar. 2011.

[44] T. M. Ingolfsson *et al.*, "Towards long-term non-invasive monitoring for epilepsy via wearable EEG devices," 2021, *arXiv:2106.08008*.

[45] U. Pale, T. Teijeiro, and D. Atienza, "Systematic assessment of hyperdimensional computing for epileptic seizure detection," 2021, *arXiv:2105.00934*.

[46] T. N. Tsuchida *et al.*, "American clinical neurophysiology society standardized EEG terminology and categorization for the description of continuous EEG monitoring in neonates: Report of the American clinical neurophysiology society critical care monitoring committee," *J. Clin. Neurophysiol.*, vol. 30, no. 2, pp. 161–173, 2013.

[47] K. L. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Oxford, U.K.: Advanced Analytics, 2014.

[48] A. Borovac, S. Gudmundsson, G. Thorvardsson, and T. P. Runarsson, "Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network," in *Proc. NeurIPS*, Dec. 2021, pp. 1–5.

[49] N. Stevenson, K. Tapani, and S. Vanhatalo, "Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5991–5994.

[50] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. The problems of two paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543–549, Jan. 1990.

[51] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples," *BMC Med. Res. Methodol.*, vol. 13, no. 1, pp. 1–7, Dec. 2013.

[52] S. Ibrahim, X. Fu, N. Kargas, and K. Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[53] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3537–3580, 2016.

[54] C. Miao, Q. Li, L. Su, M. Huai, W. Jiang, and J. Gao, "Attack under disguise: An intelligent data poisoning attack mechanism in crowdsourcing," in *Proc. World Wide Web Conf.*, 2018, pp. 13–22.

[55] V. B. Sinha, S. Rao, and V. N. Balasubramanian, "Fast dawid-skene: A fast vote aggregation scheme for sentiment classification," 2018, *arXiv:1803.02781*.

[56] V. C. Raykar *et al.*, "Learning from crowrds," *J. Mach. Learn. Res.*, vol. 11, no. 4, pp. 1–26, 2010.

Paper III

Neonatal seizure detection algorithms: The effect of channel count

Ana Borovac, Thomas P. Runarsson, Gardar Thorvardsson, and Steinn Gudmundsson

Current Directions in Biomedical Engineering, 2022

DE GRUYTER Current Directions in Biomedical Engineering 2022;8(2): 604-607

Ana Borovac*, Thomas Philip Runarsson, Gardar Thorvardsson, and Steinn Gudmundsson

Neonatal seizure detection algorithms: The effect of channel count

https://doi.org/10.1515/cdbme-2022-1154

Abstract: The number of electrodes used to acquire neonatal EEG signals varies between institutions. Therefore, tools for automatic EEG analysis, such as neonatal seizure detection algorithms, need to be able to handle different electrode montages in order to find widespread use. The aim of this study was to analyse the effect of montage on neonatal seizure detector performance. A full 18-channel montage was compared to reduced 3- and 8-channel montages using a convolutional neural network for seizure detection. Sensitivity decreased by 10 – 18 % for the reduced montages while specificity was mostly unaffected. Electrode artefacts and artefacts associated with biological rhythms caused incorrect classification of non-seizure activity in some cases, but these artefacts were filtered out in the 3-channel montage. Other types of artefacts had little effect. Reduced montages result in some reduction in classifier accuracy, but the performance may still be acceptable. Recording artefacts had a limited effect on detection accuracy.

Keywords: Seizure detection, neonatal EEG, reduced montage

1 Introduction

A neonatal electroencephalogram (EEG) is used for continuous monitoring of the state of the brain and for making patient prognosis [1]. Monitoring is usually done by placing 2 – 12 electrodes on the scalp and measuring voltage difference across electrodes [2]. The signals are frequently visualized and processed as bipolar derivations (channels), the set of which is called a *montage*. In [3, 4] it was shown that a full 19-electrode setup (18-channels) enabled detection of a larger number of seizures compared to a reduced montage, however a reduced montage may still produce clinically useful information. Fewer electrodes are preferred in practice since such setups can be applied more quickly and are easier to maintain for the duration of the recordings, some of which may last several days.

Due to scarcity of neonatal EEG experts [5] and time consuming analysis, there is an interest in developing tools for automatic EEG analysis. Such tools could make the analysis faster, more accessible, prompt wider use of EEG in NICUs, and eventually lead to improved patient care. In order for those tools to find widespread use, they need to be able to handle a variable number of EEG channels since recording protocols differ between institutions.

In this study the effects of using 3- and 8-channels in a neonatal seizure detection algorithm (NSDA) were analysed by comparing them to a full 18-channel montage. The effects of artefacts on NSDA accuracy were also investigated.

2 Methods

A publicly available EEG data set [6] containing 18-channel EEG recordings from 79 neonates and seizure annotations (labeling) by three human experts was used in this study. This data set has been studied in the context of NSDAs in the past [7–10]. The annotations do not specify which channels contain seizures, only that a seizure is present in one or more of the channels. Therefore an additional set of channel-by-channel seizure annotations made by a forth annotator [11] was also used. This second set of annotations was only used to analyse the location of (un)detected seizures with respect to the reduced 3- and 8-channel montages. Channel-by-channel artefact annotations of the same data were obtained from [12]. The artefact annotations span approx. 36 % of the recordings and are divided into six categories: clean EEG, device interference artefacts, electromyography (EMG) artefacts, movement artefacts, electrode artefacts and artefacts associated with noncortical biological (cardiac or respiratory) rhythms. These annotations were used to analyse which types of artefacts may cause incorrect classifications. Only segments belonging to one of the six categories were used for the artefact analysis.

Following [7], each EEG recording was band-pass filtered (0.5 – 16 Hz), down-sampled to 32 Hz and cut into 16 s long segments with 12 s overlap. Individual segments were standardised so that the mean was zero and the standard deviation was one. Three montages were used in this study, an 18-channel montage (figure 1), an 8-channel montage (Fp1-T3, T3-O1, Fp1-C3, C3-O1, Fp2-C4, C4-O2, Fp2-T4, T4-O2) and a 3-channel montage (F3-P3, F4-P4 and F3-F4).

The seizure detector was a convolutional neural network [13] combined with an attention layer [10] to handle a

***Corresponding author: Ana Borovac,** University of Iceland and Kvikna Medical ehf., Reykjavik, Iceland, e-mail: anb48@hi.is
Thomas Philip Runarsson, Steinn Gudmundsson, University of Iceland, Reykjavik, Iceland
Gardar Thorvardsson, Kvikna Medical ehf., Reykjavik, Iceland

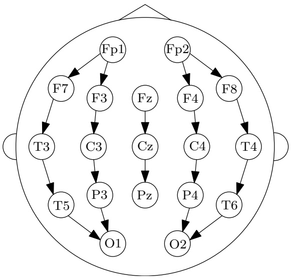


Fig. 1: Full 18-channel (double banana) montage. Electrodes are positioned according to the 10-20 system.

variable number of channels during the training and testing (i.e. when classifications/predictions are made) as described in [7]. The NSDA was trained in a leave-one-subject-out cross-validation setting. The training set contained an equal number of consensus seizure and non-seizure segments (segments where all experts were in agreement). To determine seizure/non-seizure segments, the human expert labels were used instead of the channel-by-channel labels. The network weights were optimised using the Adam optimiser with an initial learning rate of 0.001 and halving the learning rate every 10 epochs. Learning was done with a mini-batch of size 128 and stopped after 30 epochs [7]. The area under curve, sensitivity and specificity were used to evaluate classifier performance. Following [8, 9], predictions from the left-out subjects were concatenated into a single array and compared to consensus labels. By concatenating predictions, it was possible to use data from all the recordings, including those that do not contain any seizures.

3 Results

The influence of montage on detector performance was analysed by training and testing on all of the three montages. In all cases the labels correspond to the scoring done by the three experts who utilized the 18-channel montage. As most neonatal seizures are partial, i.e. they affect only a small part of the brain [14], they appear only in a subset of channels of the full 18-channel montage. Therefore, some seizures may not be picked up by the 3- or 8-channel montages and a choice of the montage is expected to influence the training and the classification of the NSDA.

As expected, for all training montages, sensitivity increased with an increasing number of channels in the test set (table 1). For the 3- and 8-channel montages, some seizure segments were incorrectly classified as non-seizures, since the seizure activity was not picked up by the montage. This is further illustrated in figures 2A and 2B which include the seizure segments picked up by the 3- and 8-channel montages. The detector performance on such segments is comparable to the 18-channel montage, especially for detectors with clinically useful specificity (defined here as 95 % or above). On the rest of the segments (figures 2C and 2D) the performance drops noticeably in comparison to the 18-channel montage.

Tab. 1: Comparison of the area under the curve (AUC), sensitivity (SE), and specificity (SP) values for 3-, 8- and 18-channel montages used for training and testing. Metrics were calculated on concatenated recordings. All the data was used, irrespective of seizure and artefact annotation.

	Testing montage		
	A: 3-channel		
Training montage	**AUC**	**SE [%]**	**SP [%]**
3-channel	0.90	77	91
8-channel	0.89	72	92
18-channel	0.90	69	95
	B: 8-channel		
	AUC	**SE [%]**	**SP [%]**
3-channel	0.89	81	87
8-channel	0.90	78	91
18-channel	0.89	77	93
	C: 18-channel		
	AUC	**SE [%]**	**SP [%]**
3-channel	0.92	86	85
8-channel	0.92	85	89
18-channel	0.94	87	93

Table 1C shows that using the 18-channel montage for training resulted in an NSDA with the highest performance accuracy when the detector was tested using 18 channels. There are two key differences with respect to the other two training montages. First, the 18-channel montage has more data for training than the 3- and 8-channel montages since each channel carries information used in training the NSDA. Second, the data was annotated using the 18-channel montage and therefore some seizure activity was not picked up by the 3- and 8-channel montages. This led to incorrectly labelled training segments and likely contributed to a decrease in specificity [15].

Moreover, a marginal increase in specificity with decreasing number of channels used in the test set is observed. In [12] it was shown that the 3-channel montage is less susceptible to artefacts than the other two. We therefore investigated whether artefact contamination leads to an increase in incorrectly classified segments.

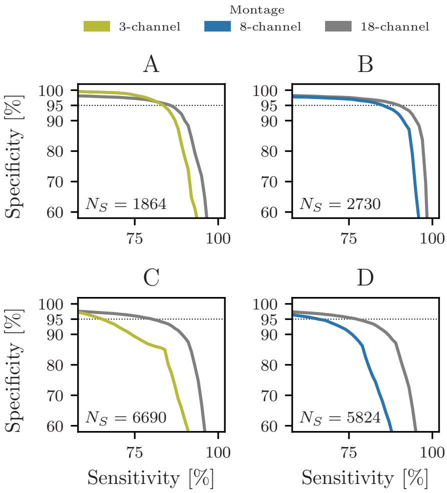


Fig. 2: Panels A and B include seizure segments picked up by the 3- and 8-channel montages. Panels C and D include the remaining seizure segments. N_S denotes the number of seizure segments. All consensus non-seizure segments were included, irrespective of artefact annotation. The horizontal line represents a lower bound for clinically useful specificity (95 %). In all cases the NSDA was trained using the 18-channel montage.

Table 2 shows sensitivity and specificity values for both the clean segments and the segments with artefacts. Only channels annotated as clean or containing an artefact were used as inputs to the NSDA (a subset of the 18-channel montage). For the seizure segments, a seizure annotation on the corresponding channels was required since it is unreasonable to expect a seizure prediction when seizure activity is absent in the input. Due to the limited number of seizure segments with an artefact and seizure annotation on the same channels, it is difficult to draw definite conclusions. The specificity values show that the NSDA performs similarly on clean inputs as for inputs containing device, EMG or movement artefacts. A 5 – 10 % lower accuracy was obtained for segments containing electrode and biological rhythm artefacts. Overall, with the exception of electrode and biological rhythm artefacts, artefacts have a minor effect on the NSDA. A likely explanation is that a relatively large number of artefacts is included in the training set (approx. 30 % [12]).

A comparison of segments containing artefacts and segments with a clean annotation on all channels is provided in figure 3. A marginal difference is observed for the 8- and 18-channel montages in the clinically significant region whereas the 3-channel montage appears to be less affected by artefacts. This could be due to the fact that F3/F4 and P3/P4 electrodes are in general less sensitive to artefacts than the prefrontal and temporal electrodes which are included in the other montages [12]. It should be noted however that these observations are based on a limited amount of data.

Tab. 2: Sensitivity and specificity using only a subset of the 18-channel montage annotated as clean or containing an artefact. The seizure segments correspond to the seizure channel-by-channel annotation. Number of segments for each case are inside parentheses. The NSDA was trained using the 18-channel montage.

	Sensitivity [%]	Specificity [%]
Clean	81 (108)	95 (17121)
Device interference	67 (3)	95 (1552)
EMG	70 (56)	96 (14630)
Movement	100 (11)	96 (4225)
Electrode artefact	73 (26)	89 (3979)
Biological rhythm	100 (10)	84 (1747)

4 Conclusion

Training an NSDA on a reduced montage resulted in a detector with a lower classification accuracy compared to a classifier trained on the full montage. This can be attributed to incorrectly labelled segments and a smaller training set. Furthermore, testing the NSDA on a reduced montage led to a lower but still clinically useful seizure detection rate. This is in line with studies analysing the effect of the number of channels on scoring by human experts [3, 4, 16–18]. Finally, the effects of artefacts on classification accuracy were investigated. We conclude that the majority artefacts appear to have only a small effect. Electrode artefacts and artefacts associated with biological rhythms caused some non-seizure segments to be incorrectly classified as seizure segments. These may in some instances be avoided by using the 3-channel montage. However, future studies with more data are required to confirm the observations.

Author Statement

Research funding: The European Union's Horizon 2020 research and innovation programme under grant agreement No 813483. Conflict of interest: Authors state no conflict of interest.

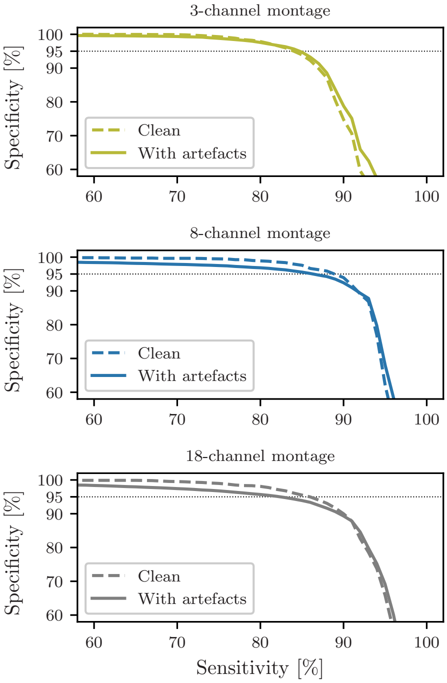


Fig. 3: Specificity as a function of sensitivity for 3-, 8- and 18-channel montages. The dashed lines represent 2525 non-seizure segments annotated as clean and the solid lines represent 23734 non-seizure segments where one or more channels contained an artefact. Seizure segments picked up by each montage were used to calculate sensitivity. The horizontal line represents a lower bound for clinically useful specificity (95 %). In all cases the NSDA was trained using the 18-channel montage.

References

[1] Britton JW, Frey LC, Hopp JL, Korb P, Koubeissi MZ, Lievens WE, Pestana-Knight EM, and St Louis EK. Electroencephalography (EEG): an introductory text and atlas of normal and abnormal findings in adults, children, and infants. 2016.

[2] Shellhaas RA, Chang T, Tsuchida T, Scher MS, Riviello JJ, Abend NS, Nguyen S, Wusthoff CJ, and Clancy RR. The American Clinical Neurophysiology Society's guideline on continuous electroencephalography monitoring in neonates. *Journal of clinical neurophysiology*, 28(6):611–617, 2011.

[3] Stevenson NJ, Lauronen L, and Vanhatalo S. The effect of reducing EEG electrode number on the visual interpretation of the human expert for neonatal seizure detection. *Clinical Neurophysiology*, 129(1):265–270, 2018.

[4] Tekgul H, Bourgeois BFD, Gauvreau K, and Bergin AM. Electroencephalography in neonatal seizures: comparison of a reduced and a full 10/20 montage. *Pediatric neurology*, 32(3):155–161, 2005.

[5] Ryan MA, Mathieson S, Dempsey E, and Boylan G. An introduction to neonatal EEG. *The Journal of perinatal & neonatal nursing*, 35(4):369–376, 2021.

[6] Stevenson NJ, Tapani K, Lauronen L, and Vanhatalo S. A dataset of neonatal EEG recordings with seizure annotations. *Scientific data*, 6(1):1–8, 2019.

[7] Borovac A, Gudmundsson S, Thorvardsson G, and Runarsson TP. Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network. In *The NeurIPS 2021 Data-Centric AI Workshop*, December 2021.

[8] Stevenson NJ, Tapani K, and Vanhatalo S. Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5991–5994. IEEE, 2019.

[9] O'Shea A, Lightbody G, Boylan G, and Temko A. Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks*, 123:12–25, 2020.

[10] Isaev DY, Tchapyjnikov D, Cotten CM, Tanaka D, Martinez N, Bertran M, Sapiro G, and Carlson D. Attention-based network for weak labels in neonatal seizure detection. *Proceedings of machine learning research*, 126:479, 2020.

[11] Tapani K and Stevenson NJ. Neonatal_seizure_detection. github.com/ktapani/Neonatal_Seizure_Detection, 2019.

[12] Webb L, Kauppila M, Roberts JA, Vanhatalo S, and Stevenson NJ. Automated detection of artefacts in neonatal EEG with residual neural networks. *Computer Methods and Programs in Biomedicine*, 208:106194, 2021.

[13] O'Shea A, Lightbody G, Boylan G, and Temko A. Investigating the impact of CNN depth on neonatal seizure detection performance. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5862–5865. IEEE, 2018.

[14] Pressler RM, Cilio MR, Mizrahi EM, Moshé SL, Nunes ML, Plouin P, Vanhatalo S, Yozawitz E, de Vries LS, Puthenveettil Vinayan K, et al. The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE Task Force on Neonatal Seizures. *Epilepsia*, 62(3):615–628, 2021.

[15] Frénay B and Verleysen M. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[16] Bye A and Flanagan D. Spatial and temporal characteristics of neonatal seizures. *Epilepsia*, 36(10):1009–1016, 1995.

[17] Shellhaas RA and Clancy RR. Characterization of neonatal seizures by conventional EEG and single-channel EEG. *Clinical Neurophysiology*, 118(10):2156–2161, 2007.

[18] Tacke M, Janson K, Vill K, Heinen F, Gerstl L, Reiter K, and Borggraefe I. Effects of a reduction of the number of electrodes in the EEG montage on the number of identified seizure patterns. *Scientific Reports*, 12(1):1–7, 2022.

Paper IV

Calibration of automatic seizure detection algorithms

Ana Borovac, Thomas P. Runarsson, Gardar Thorvardsson, and Steinn Gudmundsson

2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), 2022

Calibration of Automatic Seizure Detection Algorithms

A. Borovac[1,2], T. P. Runarsson[1], G. Thorvardsson[2] and S. Gudmundsson[1]

1. Faculty of Ind. Eng., Mech. Eng. and Comput. Sci., University of Iceland, Reykjavik, Iceland
2. Kvikna Medical ehf., Reykjavik, Iceland
{anb48, tpr, steinng}@hi.is, gardar@kvikna.com

***Abstract*— An EEG seizure detection algorithm employed in a clinical setting is likely to encounter many EEG segments that are difficult to classify due to the complexity of EEG signals and small data sets frequently used to train seizure detectors. The detectors should therefore be able to notify the clinician when they are uncertain in their predictions and they should also be accurate for confident predictions. This would enable the clinician to focus mainly on the parts of the recording where confidence in predictions is low. Here we analyse the calibration of neonatal and adult seizure detection algorithms based on a convolutional neural network in terms of how well the output seizure/non-seizure probabilities estimate the corresponding empirical frequencies. We found that the detectors turned out to be overconfident, in particular when incorrectly predicting seizure segments as non-seizure segments. The calibration of both detectors, measured in terms of expected calibration error and overconfidence error, was improved noticeably with the use of Monte Carlo dropout. We find that a straightforward application of dropout during training and classification leads to a noticeable improvement in the calibration of EEG seizure detectors based on a convolutional neural network.**

***Keywords*— electroencephalogram, automatic seizure detection, uncertainty, calibration**

I. Introduction

Seizures are common in the neonatal period [1], as well as in later stages of life [2]. Neonatal seizures should be detected and treated promptly as they often have an underlying brain injury [3]. In adulthood, the seizures may have a major impact on the quality of life and can be life threatening [4]. The current gold standard of seizure detection is a video electroencephalogram (EEG) observed by a human expert. Since EEG recordings frequently span hours to days, are prone to artefacts [5] and have high inter- and intra-patient variance [6, 7], scoring EEG recordings is time-consuming and requires special expertise that is not always available [8].

To speed up the analysis of EEG and make it more widely available, a significant effort has gone into the development of automated (neonatal) seizure detection algorithms (SDAs) [9, 10]. Designing and training SDAs with human-level performance is difficult for two main reasons. First, there is usually only a small amount of data available for training. Second, EEG signals are complex which makes seizure annotation difficult; even human experts with years of experience are often in disagreement [11, 12]. As a result, it may be expected that automatic classification would be difficult for some of the EEG segments. Algorithms that output confidence levels, in addition to seizure/non-seizure labels, are therefore desirable [13, 14]. EEG segments where confidence in prediction is low can then be passed on to the clinician for review. Furthermore, by directing the attention of the clinician to the parts of the recording where uncertainty is highest, manual scoring becomes more efficient.

Modern SDAs are based on deep neural networks (DNNs) [9, 10]. Output class probabilities may be interpreted as confidence estimates, where probabilities close to one would indicate high confidence and probabilities close to ½ would indicate low confidence in the seizure/non-seizure predictions. To the best of our knowledge, it has not been investigated how accurate such confidence estimates are in this setting. A classifier is considered to be well-*calibrated* if the confidence estimates are close to the empirical frequencies. In [15, 16] confidence estimates for a support vector machine classifier were obtained by a version of Platt scaling [17] and Becher et al. [16] additionally estimated confidence levels with trust scores [18].

Guo et al. [19] claim that DNNs are often poorly calibrated and overconfident in their predictions, despite achieving good classification performance. Hein et al. [20] show that DNNs employing the ReLU activation function can be overconfident in predictions for data far away from the training data. In recent years, various methods have been proposed for improving the calibration of DNNs [21, 22]. They include post-processing methods such as isotonic regression [23], conformal prediction [24] and Platt scaling [17], as well as methods that modify the training process such as mixup [25, 26], modelling probability distributions of class probabilities with Dirichlet distributions [27] and the use of dropout during training *and* prediction [28].

In this work, we analyse the calibration of an SDA based on a convolutional neural network and show that the detector is overconfident in its predictions, in particular for seizure segments. The calibration is improved noticeably, without degrading classification performance, by using a simple dropout strategy. The analysis is done on publicly available neonatal and adult EEG data sets.

978-1-6654-7029-2/22/$31.00 ©2022 IEEE

II. Methods

Data

The neonatal EEG was taken from a data set with 79 recordings [29] and processed as described in [30]. Briefly, the recordings were recorded with 19 electrodes with a common reference and from these signals, a bipolar longitudinal (double banana) montage with 18 channels was derived. The same montage was used by the human experts annotating the recordings [29]. The recordings were cut into 16 sec long segments with 12 sec overlap [31]. We included only segments where all three human annotators were in agreement [30]. Each signal was filtered with a 6th order Chebyshev Type 2 filter with band-pass 0.5 – 16 Hz and down-sampled from 256 Hz to 32 Hz [31]. This frequency band was selected since the cortical activity of neonates normally lies in this range [32–34]. The signals were then normalised to mean zero and standard deviation one [35–37]. Approximately 10 % of the segments contain seizures.

The adult EEG was taken from the TUH EEG seizure corpus, version 1.5.4 [38]. The set of recordings with averaged reference was used together with a bipolar temporal central parasagittal montage with 22 channels. The same montage was used by the human experts annotating the recordings. The training, validation and test sets contain recordings from 297, 41 and 41 patients, respectively. The pre-processing was similar to the neonatal data. Labelled signals were included if the manual annotation had a confidence value of one. The signals were cut into 16 sec long segments. There is an overlap of 12 sec for the seizure segments to make the set of seizures larger. The signals were filtered with a 0.5 – 25 Hz band-pass filter [39], down-sampled to 50 Hz and normalised in the same way as before. The fraction of segments containing seizures in the three data sets is 12 %.

Seizure Detection Algorithm

The SDA from [40] was used for both data sets in a binary setting (seizure/non-seizure). The detector is based on a DNN that uses multi-channel EEG as input. The network extracts features from each channel separately with a convolutional neural network [41] and combines the feature vectors into a single feature vector with an attention layer [42]. This is followed by a fully connected layer with two output nodes and a softmax activation function which provides confidence estimates for the classification. Because of the difference in sampling rates, the input size differs between the neonatal and adult data sets, resulting in a different number of features extracted from each channel (24 for the neonatal EEG vs. 44 for the adult EEG). Consequently, the numbers of parameters in the attention and fully connected layers are different. The neonatal detector has 29352 learnable parameters while the adult detector has 29712.

The training of the detector followed [30]. The neonatal (adult) detector was trained for 30 (50) epochs with the Adam optimizer, with an initial learning rate of 0.001 which was then halved every 10 epochs. Mini-batches were of size 128. Since the data sets were highly imbalanced, each mini-batch was balanced, i.e. there were 64 seizure segments and 64 non-seizure segments in each mini-batch. Hence, each epoch contained all the available seizure segments and an equal number of randomly selected non-seizure segments.

Dropout

Dropout is a simple and widely used regularization technique for improving the generalisation of DNNs [43]. With dropout, nodes are omitted at random from the network with fixed probability p during training, together with their connections. This prevents hidden nodes in the network from relying too much on other hidden nodes to correct their mistakes, which in turn reduces overfitting. In the typical setting (*standard dropout*), dropout is only used during training in order to reduce the amount of computations in the testing phase. In *Monte Carlo dropout*, T forward passes are performed with dropout enabled in the prediction phase and the predictions are averaged. It has been observed empirically that this can give a slight improvement in prediction accuracy over simple dropout. Due to the extra computational cost, Monte Carlo dropout is infrequently used for this purpose but it has the additional benefit of providing probability estimates that are better calibrated than those obtained with standard dropout [44]. The connection between Monte Carlo dropout and model uncertainty is provided in [28] where Monte Carlo dropout is interpreted as approximate Bayesian inference in deep Gaussian processes. Dropout with probability $p = 0.1$ was used for all nodes in the convolutional and attention layers (the nodes in the input layer were excluded) but for the nodes in the fully connected layer $p = 0.5$ [28, 43]. The average of $T = 10$ softmax predictions was used to obtain final probability estimates (averaging over a larger number of predictions gave similar results).

Performance Evaluation

The classification performance of the SDAs was evaluated with the area under the curve (AUC), sensitivity (SE) and specificity (SP).

The confidence of a single prediction is defined as the highest softmax output of the detector. For binary classification tasks, the confidence values, therefore, lie between 0.5 and 1. The calibration was evaluated with the expected calibration error [19],

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|, \tag{1}$$

and overconfidence error which gives high weight to confident but wrong predictions, a situation that is of particular concern in medical applications [25],

$$\mathrm{OE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \mathrm{conf}(B_k) \cdot \max(\mathrm{conf}(B_k) - \mathrm{acc}(B_k), 0), \tag{2}$$

where the confidence values have been partitioned into K equally sized bins (here $K = 5$), B_k is the set of segments where the confidence level falls into bin k, $|B_k|$ is the number of segments in bin k, $\mathrm{acc}(B_k)$ is the portion of correctly classified segments in bin k, $\mathrm{conf}(B_k)$ is the average confidence of segments in bin k and N is the total number of segments.

Leave-one-subject-out cross-validation was used for the evaluation of the neonatal SDA. The adult SDA was evaluated on a separate test set.

III. Results and Discussion

Detectors employing Monte Carlo dropout are referred to as *calibrated* in the following and they are compared to detectors that were trained without using any dropout during training and prediction (*not calibrated*).

Table 1 shows the performance of the neonatal and adult SDAs on the two data sets, averaged across patients with seizures. Inter-patient variability is quite high, in particular for the sensitivity metric. The variability of the specificity metric is lower for the adult SDA. This indicates that there are some patients that are difficult to classify in both data sets and for those, it would be preferred to obtain uncertain predictions rather than certain incorrect predictions. Such property of a detector may in the future also increase the trust of the clinicians using the system [13, 14].

Table 1. Mean area under the curve (AUC), sensitivity (SE) and specificity (SP) across the patients with at least one seizure segment. The standard deviations are shown in parentheses.

Neonatal SDA	**AUC**	**SE [%]**	**SP [%]**
Uncalibrated	0.90 (0.15)	76.02 (29.91)	93.80 (14.54)
Calibrated	0.93 (0.11)	78.39 (28.65)	95.24 (10.09)
Adult SDA			
Unalibrated	0.90 (0.14)	66.47 (32.15)	95.70 (4.85)
Calibrated	0.89 (0.16)	70.27 (32.29)	93.63 (7.05)

While the SDA architecture was designed for neonatal EEG it nevertheless gives fairly good results on the adult data set. For comparison, the best DNN architecture (out of 15 tested) reported in [45] has an AUC of 0.92, sensitivity 83 % and specificity 85 % on the TUH data set, with the caveat that [45] used a slightly different testing procedure. Detectors with high specificity (e.g., above 90 %) are often preferred in the online clinical setting to avoid frequent disruption due to false detections.

Table 1 shows that the classification performance (in terms of AUC, sensitivity and specificity) of the calibrated neonatal SDA is marginally better than for the uncalibrated detector, while the adult uncalibrated and calibrated SDAs perform similarly. This is in line with previous studies which applied Monte Carlo dropout for the classification of (medical) images [46–48].

Even though the performance of the uncalibrated and calibrated SDAs are similar in terms of the average AUC, sensitivity and specificity metrics, they can differ considerably in predictions on individual recordings. Figure 1 shows predictions for a single neonatal recording. Since the prediction confidence corresponds to the highest softmax output, the confident seizure predictions have a seizure probability close to one and confident non-seizure predictions have a seizure probability close to zero. The detector without calibration is confident in false seizure predictions for a big portion of the recording, but the predictions of the calibrated SDA which have high seizure probability are in agreement with the three human annotators that labelled the data set. Three out of seven consensus seizures are clearly detected and two additional seizures can quickly be identified by inspecting the areas with high uncertainty (figure 1), i.e. with seizure probability around 0.5.

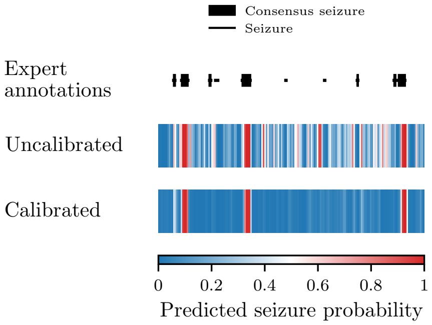
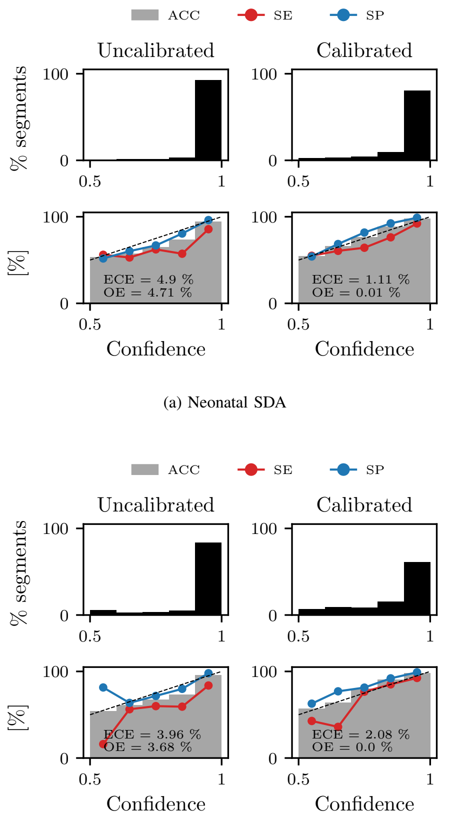


Figure 1. An example of predictions made by the uncalibrated and calibrated neonatal SDAs for a 56 min long neonatal recording. The recording contains seven seizures where all three human experts were in agreement and three additional seizures were labelled by at least one of the experts.

The calibration of the SDAs is further analyzed in figure 2. The uncalibrated neonatal (adult) detector predicts about 93 % (83 %) of the examples with confidence close to one. The reliability diagrams for this



(a) Neonatal SDA

(b) Adult SDA

Figure 2. Neonatal (a) and adult (b) SDAs without calibration (left) and with calibration (right). Confidence histograms (black) show the fraction of predictions with a given confidence value and reliability diagrams (grey) show the expected accuracy as a function of confidence value. Deviations from the dashed lines represent miscalibration. Accuracy (ACC), sensitivity (SE), specificity (SP), expected calibration error (ECE) and overconfidence error (OE).

case show that the confidence levels do not reflect the true accuracy, as indicated by deviation from the dashed lines. The deviation is clearly lower when calibration is applied and this is also reflected in the expected calibration error. In addition, the uncalibrated detectors are overconfident in their predictions, seizure predictions in particular, which results in a high overconfidence error. The error drops to 0.01 % and 0.0 % for the calibrated neonatal and adult detectors, respectively. Low overconfidence error would consequently allow the user to trust predictions with a high (e.g., > 0.9) confidence level. In other words, highly confident predictions are almost always correct and the risk of false detection or a missed detection is low.

Figure 3 illustrates the relationship between classifier performance and mean confidence levels for individual patients in the adult data set. The calibrated SDA estimates for sensitivity and specificity are closer to the true values (dashed lines) and more importantly, the confidence estimates for the difficult examples are much lower than for the patients on which the SDA performs almost perfectly. Similar observations are made also on the neonatal data (data not shown).

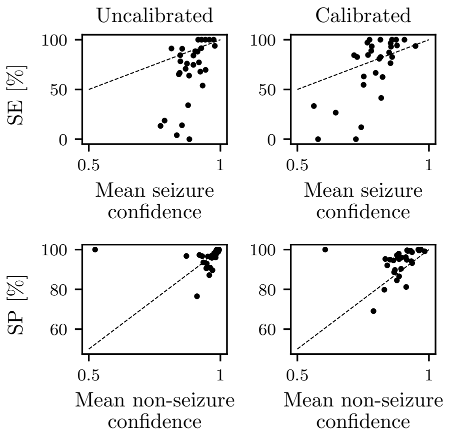


Figure 3. Each dot represents a patient from the adult data set. Mean seizure (non-seizure) confidence is the average confidence level of segments predicted as seizures (non-seizures). These two values are estimates for sensitivity (SE) and specificity (SP).

IV. Conclusion

In this work, we have shown that an SDA based on a DNN architecture optimised for neonatal seizure detection, can be retrained on adult EEG data to provide a reasonably accurate classifier for adult EEG. However, despite good classification performance, neonatal and adult detectors were overconfident in the predictions which may reduce user trust [13, 14]. Our results demonstrate that dropout [28] improves calibration, in particular for the seizure segments. A well-calibrated detector can notify the user when it is not confident in its predictions and leave the decision to the user. This allows the user to focus quickly on the parts of the recording where the automatic detection is uncertain.

As suggested in [46, 47] Monte Carlo dropout may not perform well in case of a distribution shift. In our case, the shift can be attributed to the different age groups, recording equipment and protocols. Therefore, further research is needed to investigate the influence of a distribution shift on the calibration of an SDA.

Dropout may also be combined with mixup training [25] and post-processing schemes such as Platt scaling [17] in the future. More work is also needed to find out how to present the output of calibrated detectors in intuitive and informative ways in the clinical setting.

Acknowledgements

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 813483.

References

[1] F. Pisani, C. Facini, E. Bianchi, G. Giussani, B. Piccolo, and E. Beghi, "Incidence of neonatal seizures, perinatal risk factors for epilepsy and mortality after neonatal seizures in the province of Parma, Italy," *Epilepsia*, vol. 59, no. 9, pp. 1764–1773, 2018.

[2] R. Kobau, F. Gilliam, and D. J. Thurman, "Prevalence of self-reported epilepsy or seizure disorder and its associations with self-reported depression and anxiety: results from the 2004 Healthstyles Survey," *Epilepsia*, vol. 47, no. 11, pp. 1915–1921, 2006.

[3] C. Vasudevan and M. Levene, "Epidemiology and aetiology of neonatal seizures," *Seminars in Fetal and Neonatal Medicine*, vol. 18, no. 4. Elsevier, 2013, pp. 185–191.

[4] B. Litt and J. Echauz, "Prediction of epileptic seizures," *The Lancet Neurology*, vol. 1, no. 1, pp. 22–30, 2002.

[5] L. Webb, M. Kauppila, J. A. Roberts, S. Vanhatalo, and N. J. Stevenson, "Automated detection of artefacts in neonatal EEG with residual neural networks," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106194, 2021.

[6] R. A. Hrachovy and E. M. Mizrahi, *Atlas of neonatal electroencephalography*. Springer Publishing Company, 2015.

[7] S. Noachtar and J. Rémi, "The role of EEG in epilepsy: a critical review," *Epilepsy & Behavior*, vol. 15, no. 1, pp. 22–33, 2009.

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

[9] B. Olmi, L. Frassineti, A. Lanata, and C. Manfredi, "Automatic Detection of Epileptic Seizures in Neonatal Intensive Care Units Through EEG, ECG and Video Recordings: A Survey," *IEEE Access*, vol. 9, pp. 138 174–138 191, 2021.

[10] S. Saminu, G. Xu, Z. Shuai, I. Abd El Kader, A. H. Jabire, Y. K. Ahmed, I. A. Karaye, and I. S. Ahmad, "A recent investigation on detection and classification of epileptic seizure techniques using EEG signal," *Brain Sciences*, vol. 11, no. 5, p. 668, 2021.

[11] A. Dereymaeker, A. H. Ansari, K. Jansen, P. J. Cherian, J. Vervisch, P. Govaert, L. De Wispelaere, C. Dielman, V. Matic, A. C. Dorado *et al.*, "Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: the neoguard eeg database," *Clinical Neurophysiology*, vol. 128, no. 9, pp. 1737–1745, 2017.

[12] J. Halford, D. Shiau, J. Desrochers, B. Kolls, B. Dean, C. Waters, N. Azar, K. Haas, E. Kutluay, G. Martz *et al.*, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings," *Clinical Neurophysiology*, vol. 126, no. 9, pp. 1661–1669, 2015.

[13] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[14] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: communicating uncertainty in medical machine learning," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–6, 2021.

[15] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 464–473, 2011.

[16] T. Becker, K. Vandecasteele, C. Chatzichristos, W. Van Paesschen, D. Valkenborg, S. Van Huffel, and M. De Vos, "Classification with a deferral option and low-trust filtering for automated seizure detection," *Sensors*, vol. 21, no. 4, p. 1046, 2021.

[17] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[18] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," *Advances in neural information processing systems*, vol. 31, 2018.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[20] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 41–50.

[21] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.

[22] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[23] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 694–699.

[24] G. Shafer and V. Vovk, "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

[25] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[27] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[28] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[29] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizure annotations," *Scientific data*, vol. 6, p. 190039, 2019.

[30] A. Borovac, S. Guðmundsson, G. Thorvardsson, and T. P. Runarsson, "Influence of human-expert labels on a neonatal

seizure detector based on a convolutional neural network," *The NeurIPS 2021 Data-Centric AI Workshop*, 2021.

[31] N. J. Stevenson, K. Tapani, and S. Vanhatalo, "Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 5991–5994.

[32] J. J. Alix, A. Ponnusamy, E. Pilling, and A. R. Hart, "An introduction to neonatal EEG," *Paediatrics and Child Health*, vol. 27, no. 3, pp. 135–142, 2017.

[33] M. Kitayama, H. Otsubo, S. Parvez, A. Lodha, E. Ying, B. Parvez, R. Ishii, Y. Mizuno-Matsumoto, R. A. Zoroofi, and O. C. Snead III, "Wavelet analysis for neonatal electroencephalographic seizures," *Pediatric neurology*, vol. 29, no. 4, pp. 326–333, 2003.

[34] A. M. Husain, "Review of neonatal EEG," *American journal of electroneurodiagnostic technology*, vol. 45, no. 1, pp. 12–35, 2005.

[35] G. Xu, T. Ren, Y. Chen, and W. Che, "A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis," *Frontiers in Neuroscience*, vol. 14, p. 578126, 2020.

[36] H. Mukhtar, S. M. Qaisar, and A. Zaguia, "Deep convolutional neural network regularization for alcoholism detection using EEG signals," *Sensors*, vol. 21, no. 16, p. 5456, 2021.

[37] A. Shoeibi, D. Sadeghi, P. Moridian, N. Ghassemi, J. Heras, R. Alizadehsani, A. Khadem, Y. Kong, S. Nahavandi, Y.-D. Zhang *et al.*, "Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models," *Frontiers in Neuroinformatics*, vol. 15, 2021.

[38] A. Harati, S. Lopez, I. Obeid, J. Picone, M. Jacobson, and S. Tobochnik, "The TUH EEG CORPUS: A big data resource for automated EEG interpretation," *2014 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE, 2014, pp. 1–5.

[39] J. Gotman, J. Ives, and P. Gloor, "Frequency content of EEG and EMG at seizure onset: possibility of removal of EMG artefact by digital filtering," *Electroencephalography and clinical neurophysiology*, vol. 52, no. 6, pp. 626–639, 1981.

[40] A. Borovac, S. Gudmundsson, G. Thorvardsson, S. M. Moghadam, P. Nevalainen, N. Stevenson, S. Vanhatalo, and T. P. Runarsson, "Ensemble learning using individual neonatal data for seizure detection," *arXiv preprint arXiv:2204.07043*, 2022.

[41] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Investigating the impact of CNN depth on neonatal seizure detection performance," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5862–5865.

[42] D. Y. Isaev, D. Tchapyjnikov, C. M. Cotten, D. Tanaka, N. Martinez, M. Bertran, G. Sapiro, and D. Carlson, "Attention-based network for weak labels in neonatal seizure detection," *Proceedings of machine learning research*, vol. 126, p. 479, 2020.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.

[45] K. Lee, H. Jeong, S. Kim, D. Yang, H.-C. Kang, and E. Choi, "Real-Time Seizure Detection using EEG: A Comprehensive Comparison of Recent Approaches under a Realistic Setting," *arXiv preprint arXiv:2201.08780*, 2022.

[46] R. Krishnan and O. Tickoo, "Improving model calibration with accuracy versus uncertainty optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 237–18 248, 2020.

[47] J. Thagaard, S. Hauberg, B. v. d. Vegt, T. Ebstrup, J. D. Hansen, and A. B. Dahl, "Can you trust predictive uncertainty under real dataset shifts in digital pathology?" *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 824–833.

[48] Z. Zhang, A. V. Dalca, and M. R. Sabuncu, "Confidence calibration for convolutional neural networks using structured dropout," *arXiv preprint arXiv:1906.09551*, 2019.

Paper V

Calibration methods for automatic seizure detection algorithms

Ana Borovac, David H. Agustsson, Thomas P. Runarsson, and Steinn Gudmundsson

Accepted for publication

Calibration Methods for Automatic Seizure Detection Algorithms

Ana Borovac[1,2], David Hringur Agustsson[1], Tomas Philip Runarsson[1], and Steinn Gudmundsson[1]

[1] Faculty of Ind. Eng., Mech. Eng. and Comput. Sci., University of Iceland, Reykjavik, Iceland
[2] Kvikna Medical ehf., Reykjavik, Iceland
{anb48, dha4, tpr, steinng}@hi.is

Abstract.
Background: Automatic seizure detection algorithms have been in development for years with the aim of making the analysis of long EEG recordings more efficient. To train such detectors, a large amount of EEG data with precise seizure annotations is required. However, due to privacy concerns, and the inherent complexity of EEG signals, obtaining data sets diverse enough to capture all relevant EEG patterns is difficult. The current state-of-the-art seizure classification algorithms are far from perfect and routinely misclassify EEG segments as seizure where there is no seizure activity and vice versa. A seizure detection algorithm that can indicate where its predictions are of low confidence, thereby requiring verification by a human expert, carries substantial real-world value. Modern seizure detectors based on deep neural networks can output probability/confidence estimates alongside seizure/non-seizure classification, but little attention has been given to how accurate these estimates are, in other words, how well the detector is calibrated.
Methods: In this study, we analyzed the calibration of seizure detectors based on a convolutional neural network, that were trained on adult and neonatal EEG data, respectively. Four calibration methods from the literature, temperature scaling, ensemble, dropout, and mixup were evaluated.
Results: We found that the uncalibrated detectors make the vast majority of the predictions with confidence close to one, i.e., they are overconfident and, therefore, the detectors with higher overall accuracy are better calibrated. Our results indicate that all the calibration methods studied here make the detectors less confident in incorrect predictions, a desirable trait, but to a lesser extent, they also result in detectors less confident in correct predictions. The best calibration was obtained with the ensemble and dropout methods. When class labels in the seizure data are highly imbalanced, it is recommended that confidence estimates for individual classes are analyzed separately.

Keywords: calibration, uncertainty, deep neural networks, automatic seizure detection, electroencephalogram

1 Introduction

Seizures are a common neurological emergency, with an estimated prevalence of 1 % [30], that can cause permanent brain damage, and even death, if untreated [45]. Almost half of the neonates affected by seizures face long-term neurodevelopmental disorders [51]. Adults experiencing seizures are at higher risk for psychiatric disorders such as depression and are about 10 times more likely to commit suicide than the general population [22]. Heart dysfunction provoked by seizures can cause sudden death [44]. To improve the lives of people with seizures, prompt detection and appropriate treatment is crucial. Treatment options include anti-epileptic drugs, brain surgery and electrical brain stimulation [27, 28, 38, 41].

The current gold standard for neonatal and adult seizure detection is an electroencephalogram (EEG), a recording of the electrical activity of the brain. EEG signal acquisition is typically done by placing electrodes on the scalp and the voltage difference between pairs of electrodes is recorded. The recordings can span from minutes to days, depending on the clinical indication for EEG monitoring. Due to signal complexity and high variability [20, 34, 53], analysis of EEG recordings requires time and special expertise that is not always available [7].

The goal of automated seizure detection algorithms (SDAs) [35, 43] is to accelerate EEG analysis significantly while preserving the current level of diagnostic accuracy. Such SDAs could enable the widespread use of EEGs, e.g. in intensive care units, without the need for experts to monitor each recording. The development of SDAs that perform as well as human experts faces two main challenges. First, due to patient privacy issues, there is often a limited amount of data available for algorithm training [10]. Second, obtaining precise annotations of seizure onset and offset times is challenging, as human experts may disagree on the presence of seizure events [9, 17]. SDAs trained on relatively small data sets are expected to have difficulties classifying unseen EEG segments accurately. Compounding the problem is the presence of label noise in the data because of ambiguity in the human annotations of the EEG [5]. Combining seizure/non-seizure predictions with confidence estimates would make the detectors more useful in a clinical setting [3, 24]. By doing so, EEG intervals with low-confidence predictions can be flagged for review by a human expert. The end result would be faster analysis without compromising the accuracy of the annotations. For example, a study using an SDA based on a support vector machine (SVM) suggests that by passing 40 % of the data with the least confident predictions to a human expert, an accuracy of 99 % could be achieved [2].

Many modern SDAs are based on deep neural networks (DNNs) [35, 43]. For a given EEG segment, a neural network classifier outputs a value between zero and one. This value can be interpreted as an estimate of the probability that the segment contains a seizure. A value close to zero indicates that the segment is unlikely to contain a seizure and a value close to one indicates that the segment most likely contains a seizure. The value can thus be regarded as the *confidence* the classifier has in the prediction. By thresholding at, say, 0.5, the segment can be classified as a seizure or non-seizure segment and labelled accordingly in the EEG recording. However, the accuracy of these confidence estimates has received limited attention in the context of SDAs [6]. In case the estimates are accurate, a classifier is considered to be well-*calibrated*. In other words, if a classifier outputs a confidence estimate of, e.g., 0.7 for some EEG segments, and it is correct in its prediction for 70 % of these segments, the classifier is well-calibrated. The same should also hold for other confidence levels.

Guo et al. [16] have reported that DNNs trained on image and document classification tasks tend to be overconfident in their predictions, despite achieving high classification accuracy. Based on their empirical results, they suggest several potential causes that result in poorly calibrated DNNs, including increased model capacity, batch normalization, training with small weight decay and using the cross-entropy loss function [54]. Thulasidasan et al. [50] suggest that training with 0/1 annotations negatively influences calibration and is improved by utilizing mixup [56]. Hein et al. [19] showed that DNNs employing the ReLU activation function can be particularly overconfident in their predictions for data far away from the training data. On the other hand, Minderer et al. [31] found that state-of-the-art DNNs for image classification tend to be well-calibrated and suggest that improvements in model accuracy benefit calibration. It should be noted that the image classification data sets employed in the above studies typically feature hundreds of classes whereas seizure detection is normally formulated as a binary classification task. Researchers have proposed various approaches to improve the calibration of DNNs [1, 12]. These methods include post-processing techniques such as isotonic regression [55] and Platt scaling [39], which adjust the output probabilities of the network in order to improve calibration. Methods such as mixup [50, 56] and dropout [11] modify the training process, and in the case of dropout, also the prediction process.

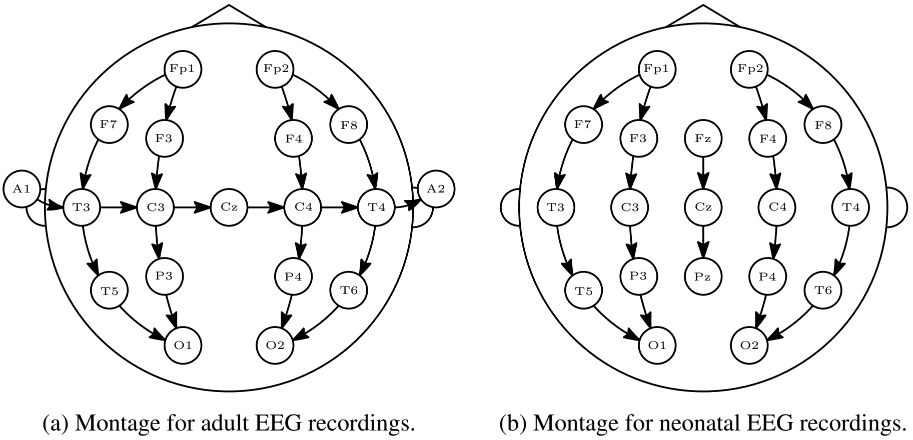
In this work, we extend our previous analysis of SDA calibration [6] by analyzing four different calibration methods that have been found to work well with DNNs, albeit in different settings. We show that neonatal and adult SDAs based on a convolution neural network are overconfident in their predictions and that detectors with higher overall accuracy are better calibrated. All the calibration methods evaluated here, temperature scaling, ensemble, dropout and mixup, make the detector less confident for incorrect predictions. A comparison of the methods is done on two publicly available data sets; one adult and one neonatal data set.

2 Methods

2.1 Data

The adult EEG data set was obtained from version 2.0.0 of the TUH EEG seizure corpus [18], which consists of recordings with diverse recording set-ups and seizure types. The most frequent seizure type is focal non-specific seizures, but other types, such as generalized non-specific seizures and complex partial seizures are also present. In this study, we utilized a subset of recordings recorded with averaged reference, i.e. average potential of all the electrodes was used as a reference. The acquisition of the signals was done with a version of a NicoletOne EEG system (Natus, USA) and the sampling frequency was between 250 Hz and 1000 Hz. Human experts annotated the recordings using a bipolar temporal central parasagittal montage with 22 channels (Fig. 1), and the same montage was employed in this study. Specifically, channels Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, T3-C3, C3-Cz, Cz-C4, C4-T4, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, A1-T3 and T4-A2 were derived from the recorded signals. The data set contains predefined training, validation and test sets.

The neonatal data set used in this study consists of 79 recordings [47]. Acquisition of the EEG signals was done with NicoletOne EEG system (Natus, USA), using 19 electrodes with the reference electrode located at the midline and the sampling frequency was 256 Hz. To annotate the recordings, three human experts utilized a bipolar longitudinal (double banana) montage with 18 channels. Schematic representation of the montage is given in Fig. 1. For this study, the same set of channels was used, including Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz and Cz-Pz, which were derived from the recorded EEG signals. The data set does not come with predefined training, validation and test set splits. Table 1 shows summary statistics for the two data sets. We note that the data sets are imbalanced, i.e., less than 10 % of the total recording duration corresponds to seizure segments.



(a) Montage for adult EEG recordings. (b) Montage for neonatal EEG recordings.

Fig. 1: Montages for adult (a) and neonatal (b) EEG recordings. In both cases, the electrodes are positioned according to the 10-20 system. Each arrow denotes an EEG channel that is used as input to an SDA.

Since most adult seizures are present in the frequency range between 3 and 30 Hz [15] and neonatal seizures can be as slow as 0.5 Hz [13], the EEG signals were filtered with a Butterworth band-pass filter with cut-off frequencies 0.5 Hz and 30 Hz. Before filtering, the EEG signals were downsampled to 250 Hz and further downsampled to 62 Hz after filtering, reducing the input size and subsequently model size by approximately factor 4. After filtering and downsampling, each recording was cut into 16 seconds segments.

Table 1: Summary statistics for the adult [18] and neonatal [47] data sets used in this study. A patient "with seizures" has at least one 16 seconds seizure segment. Standard deviations are shown in parentheses.

	Adult data set			Neonatal data set
	Training	Validation	Test	
Number of patients	297	41	41	79
Total duration of recordings [hours]	603.08	372.21	119.98	111.90
Total duration of seizures [hours]	26.41	10.91	7.57	10.91
Fraction of seizure activity [%]	4.38	2.93	6.31	9.75
Average duration of recordings per patient [hours]	2.03 (3.29)	9.08 (17.85)	2.93 (2.07)	1.42 (0.56)
Average duration of seizures per patient with seizures [hours]	0.24 (0.41)	0.34 (0.42)	0.22 (0.31)	0.28 (0.38)
Number of seizure segments	19148	8066	5197	8563
Number of non-seizure segments	127220	79759	24547	20233

To increase the amount of seizure data available for training, an overlap of 12 seconds was used for the seizure segments.

2.2 Seizure Detection Algorithm

The detector takes multi-channel EEG as input and outputs seizure/non-seizure probability estimates. This is accomplished by extracting features from each EEG channel via 11 convolutional layers with 32 filters of size 3×1, followed by batch normalization layers and ReLU activation functions [36]. Average pooling is applied before the fourth, seventh, and tenth convolutional layers. An attention layer is used to combine feature vectors extracted from individual EEG channels into one feature vector [21]. The classification part of the network is a fully connected layer that maps feature vectors of dimension 58 to two outputs (seizure/non-seizure) utilizing a softmax activation function to obtain values in $(0, 1)$ and can be interpreted as class probabilities. With only $29,964$ learnable parameters, the SDA is practically tiny, compared to state-of-the-art networks used in natural language processing and computer vision. A benefit of using such a small network is that it can be deployed on devices with limited computation resources. A detailed description of the SDA is given in [4].

The adult and neonatal detectors were trained by optimizing the negative log-likelihood loss function with the Adam optimizer and a mini-batch size of 256. To address the imbalance between the number of available seizure and non-seizure segments in the training sets, each mini-batch contained 128 seizure and 128 non-seizure segments. One epoch corresponds to a single pass through all available seizure segments and an equal number of randomly selected non-seizure segments. The SDAs were trained for 50 epochs where the initial learning rate of 0.001 was halved every 20 epochs. The number of epochs and the learning rate decay were chosen so that the area under the curve computed on the adult validation set was maximized. The hyper-parameter values used in this experiment are similar to those of previous experiments [5] conducted on the neonatal data set. We observed that the performance of the SDA is insensitive to small changes in hyper-parameter values.

2.3 Calibration methods

Temperature scaling

Platt scaling [39] is a generic method to transform classifier outputs to a probability distribution over classes. It was originally proposed for use with SVM classifiers and has previously been used with an SVM-based neonatal seizure detector to smooth classifier outputs and to aggregate predictions over multiple channels [48]. The method fits a parameterized sigmoid or softmax function to (unscaled) classifier outputs. In [16] a simplified version with one learnable parameter called *temperature scaling* was used to improve the calibration of neural networks trained on image and document data. In case the non-seizure class is denoted with 0 and the seizure class with 1, the calibrated seizure/non-seizure probability estimates are obtained as follows,

$$\hat{p}_j^{(i)} = \frac{\exp\left(z_j^{(i)}/\tau\right)}{\exp\left(z_0^{(i)}/\tau\right) + \exp\left(z_1^{(i)}/\tau\right)}; \quad i = 1,2,\ldots,N, \; j = 0,1, \tag{1}$$

where $z_c^{(i)}$ $(c = 0,1)$ are the unscaled outputs of instance i and $\hat{p}_j^{(i)}$ is the calibrated probability of instance i belonging to class j. The value of the τ parameter is chosen based on a held-out validation set after the network has been trained to avoid overfitting [39]. In case $\tau = 1$ the calibrated probabilities are equal to the softmax outputs and when τ is large the probabilities approach $1/2$. Applying temperature scaling to the unscaled outputs of the network does not change the predicted seizure/non-seizure labels, the only difference is in the probability (confidence) estimates.

Dropout

Dropout is a simple and widely used regularization technique for improving the generalization of DNNs [46]. The idea behind dropout is to randomly drop nodes from the network with a fixed probability, p. This forces the network to learn more robust features that are not dependent on any single node and reduce overfitting. Dropout is usually only used during training, i.e., the full network is used to obtain predictions. Dropout can also be employed in the prediction phase (Monte Carlo dropout). In this setting, the final seizure/non-seizure prediction is obtained by averaging T softmax outputs. It has been shown empirically that Monte Carlo improves the calibration of DNNs [29] and can be interpreted as approximate Bayesian inference [11]. Dropout with $p = 0.1$ was used for the convolutional and attention layers and $p = 0.5$ for the fully connected layer [11, 46]. The average of $T = 10$ softmax predictions was used for final probability estimates (averaging over a larger number of predictions gave similar results, data not shown).

Deep ensembles

An ensemble of multiple DNNs, referred to as *deep ensemble* in the following, has been shown to give small improvements in classification performance compared to the best individual model in the ensemble [23]. An added benefit of using an ensemble of DNNs is improved calibration. Lakshminarayanan et al. [26] found that an ensemble with only five models trained with the same setup can lead to a noticeable improvement in calibration. Here we used an ensemble of 10 SDAs. Each individual SDA was trained with the same training parameters and the same data. The resulting SDAs were nevertheless not identical since network weights were randomly initialized for each network prior to training. Additionally, the order in which the data was presented to the network was different since the data was randomly shuffled for every epoch. Once the detectors were trained, the final prediction was obtained by averaging the softmax outputs.

Mixup

Mixup is a data-agnostic augmentation method that has been found to improve the generalization of many neural network architectures [56]. It has also been found to improve the model calibration of classifiers for both images and text [50]. Mixup creates augmented training examples by forming linear combinations of feature-target pairs. A new feature-target $(\tilde{x}, \tilde{y})$ is generated as follows,

$$\tilde{x} = \lambda x^{(i)} + (1 - \lambda) x^{(j)}, \tag{2}$$

$$\tilde{y} = \lambda y^{(i)} + (1 - \lambda) y^{(j)}, \tag{3}$$

where $x^{(i)}$ and $x^{(j)}$ are two randomly selected training EEG segments, $y^{(i)}$ and $y^{(j)}$ are corresponding 0/1 (non-seizure/seizure) labels and $\lambda \in [0, 1]$ is a random variable drawn from a Beta distribution with hyper-parameter α. It is important to select an appropriate α to achieve good results, in this study, $\alpha = 0.3$ was used after testing several different values on the adult validation set.

2.4 Evaluation

The adult SDA was assessed using a dedicated test set, while the evaluation of the neonatal SDA was done using leave-one-subject-out cross-validation since the data set lacks a distinct test set.

The SDAs were assessed for their classification performance using the area under the curve (AUC), sensitivity (SE), and specificity (SP). Sensitivity refers to the fraction of correctly classified seizure segments, while specificity refers to the fraction of correctly classified non-seizure segments. The *confidence* of a prediction is the softmax output of the predicted seizure/non-seizure class, i.e., the class with the higher probability estimate. Since the threshold for seizure/non-seizure prediction is set at 0.5, the confidence estimates range between 0.5 and 1.0.

A *reliability diagram* [32] is a visual representation of a classifier's calibration, as shown in figure 2. The diagram shows the fraction of accurately predicted segments, providing an empirical estimation of the true underlying accuracy, against confidence levels. A well-calibrated classifier is indicated by empirical frequencies that align closely with the line of average confidence within a given bin. If there is sufficient data, the average confidence line should approximate the identity line.

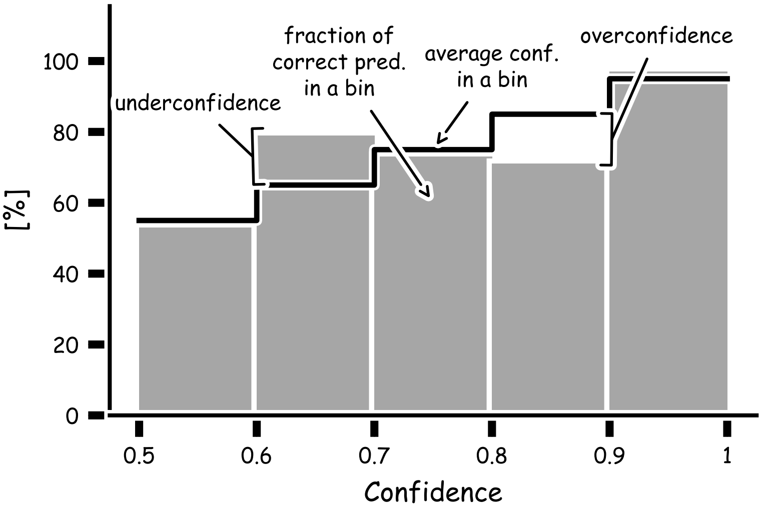


Fig. 2: Reliability diagram. The interval between the lowest (0.5) and highest (1.0) possible confidence values is split into five equally sized bins. All EEG segments are allocated to a bin based on the confidence of their predictions. Grey bars represent the fraction of the correctly predicted segments in a bin. The black curve represents the average confidence in each bin. Differences between the bars and the curve indicate miscalibration, i.e. the SDA is either underconfident or overconfident for the predictions in the bin.

To evaluate the calibration metrics, all N available seizure and non-seizure segments were split into $K = 5$ bins based on the confidence estimate made by the SDA. Bin edges were set such that the interval between the lowest (0.5) and highest (1.0) possible confidence is partitioned into equally sized intervals. The set of segments in bin k is denoted with B_k and $|B_k|$ is the number of segments in bin k. The fraction of correct predictions (*empirical frequency*) in a bin k is denoted with $\mathrm{acc}(B_k)$ and the average confidence estimate in bin k with $\mathrm{conf}(B_k)$.

The *expected calibration error* (ECE) [16],

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \left|\mathrm{acc}(B_k) - \mathrm{conf}(B_k)\right|, \tag{4}$$

measures the difference between predicted confidence and empirical frequency. Bins with more segments weigh more than bins with fewer segments. The closer the metric is to zero, the better calibrated the model is.

In medical applications, classifiers that are not overconfident in the predictions are preferred. Therefore, we include the *overconfidence error* (OE) [50] for calibration evaluation,

$$\mathrm{OE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \mathrm{conf}(B_k) \cdot \max(\mathrm{conf}(B_k) - \mathrm{acc}(B_k), 0), \tag{5}$$

A modification of the static calibration error (SCE) [33] is proposed to capture calibration of individual classes (seizure and non-seizure) when the class frequencies differ widely,

$$\mathrm{SCE} = \frac{1}{KC} \sum_{k=1}^{K} \sum_{c=1}^{C} \frac{|B_{c_k}|}{N_c} \left|\mathrm{acc}(B_{c_k}) - \mathrm{conf}(B_{c_k})\right|. \tag{6}$$

In comparison with the original definition in [33], this definition differs in the weighting factor $|B_{c_k}|/N_c$, where N_c is the number of segments of class c (seizure or non-seizure). Here the weights are proportional to the number of segments in each class and not the total number of segments. As a result, all the classes have the same weight in the overall sum and the imbalanced data issue is addressed. In other words, the static calibration error is the average expected calibration error using segments of just one class.

We also include two calibration metrics which measure the distance of the estimated probability to the target label. Specifically, the Brier score (BS) [8],

$$\mathrm{BS} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}^{(i)} - y^{(i)}\right)^2, \tag{7}$$

and negative log-likelihood (NLL) [40],

$$\mathrm{NLL} = -\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right), \tag{8}$$

where $\hat{y}^{(i)}$ is the softmax output of instance i for class 1 (seizure class) and $y^{(i)}$ is the target label of instance i.

2.5 Implementation

All code was written in Python 3.9. EEG recordings in EDF format were read with the MNE library [14] (version 0.24.1) and pre-processed with SciPy [52] (version 1.8.0). The detectors were developed with PyTorch [37] (version 1.11.0) and an NVIDIA GeForce GTX 1080 Ti graphics card. The code is available at a GitHub repository (github.com/anaborovac/Calibrated-SDA).

3 Results and Discussion

In the following, we refer to an SDA which does not utilize any specific calibration method as *uncalibrated*. We show that the calibration of detectors is highly correlated with overall classification accuracy as most correct and incorrect predictions appear to have high confidence. Using calibration methods does not result in perfectly calibrated SDAs, however, lower confidence in incorrect predictions is obtained with all the methods, temperature scaling, ensemble, dropout and mixup.

3.1 Tuning hyper-parameters

The number of training epochs, learning decay schedule and the α parameter in mixup were optimized by maximizing the AUC on the adult validation set. The neonatal data set is relatively small and it is therefore costly to set aside separate data for validation. Instead of reducing the amount of neonatal data available for training, we decided to simply train the neonatal detector using the same hyper-parameters that we obtained for the adult detector. Fig. 3 shows the negative log-likelihood loss and AUC during training on the adult data sets. The validation loss fluctuates significantly but the AUC is relatively stable after approx. 30 epochs. Longer training and different weight decay schedules gave similar results (data not shown). The fluctuations in the validation loss may be due to a small and imbalanced data set, overfitting or ambiguity in annotations of seizure/non-seizure segments in the data sets.

3.2 Uncalibrated SDAs

To construct the deep ensembles, 10 sets of uncalibrated adult and neonatal SDAs were obtained by starting from random initial weights. We begin by analyzing these detectors individually to gain insight into the variability in the classification and calibration performance of individual classifiers.

Fig. 4 shows that the performance of the adult SDAs is on average slightly better in comparison with the neonatal SDAs. This is not unexpected since the adult training set is significantly larger. The figure also shows that neonatal SDAs have less variability than adult SDAs. This may simply be a consequence of the use of leave-one-subject-out cross-validation on the neonatal data set since averaging over 38 detectors has a smoothing effect. Fig. 4 also shows the expected trade-off between sensitivity and specificity. Detectors with high seizure detection rates incorrectly classify more non-seizure segments as seizures and vice versa. However, since the AUC values are similar for

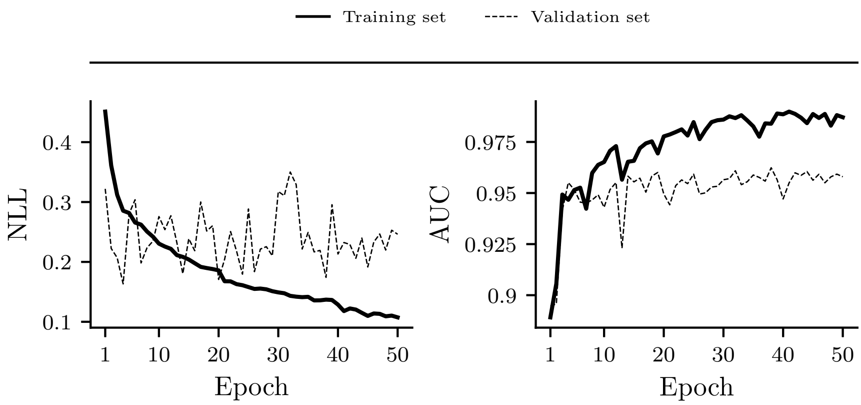


Fig. 3: Left: Negative log-likelihood (NLL) loss on the adult data set during training of an uncalibrated SDA. Right: Corresponding area under the curve (AUC) values.

all adult and neonatal detectors, respectively, the threshold for seizure/non-seizure prediction may be adjusted to achieve the desired classification performance. Adult SDAs exhibit considerable variance in sensitivity. A possible explanation is that the number of seizure segments available for training is much lower than the number of non-seizure segments. This may result in detectors that are not able to accurately capture the relevant

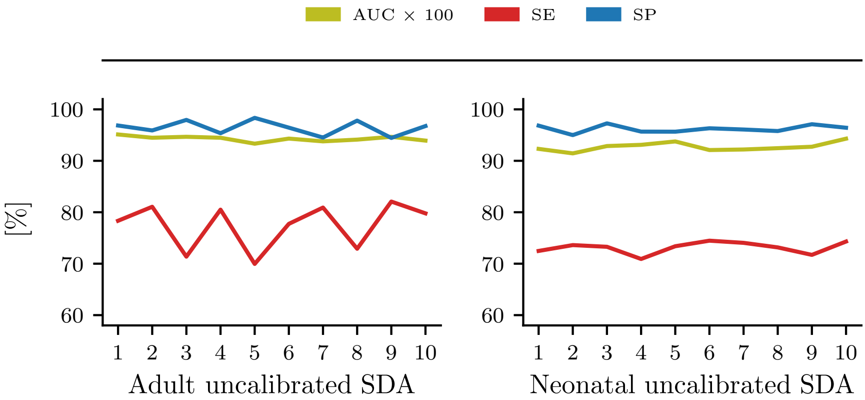


Fig. 4: Area under the curve (AUC), sensitivity (SE) and specificity (SP) for individual uncalibrated adult (left) and neonatal (right) SDAs from the deep ensembles (arbitrary order). Metrics are averaged across all patients which have at least one seizure segment. A separate test set is used to compute metrics for the adult data set while leave-one-subject-out cross-validation is used to compute the metrics for the neonatal data set.

features that differentiate seizures from non-seizure segments. Other factors that could contribute to higher variance are the heterogeneity of the different seizure types and incorrect annotations due to ambiguity in the scoring of EEGs by human experts [9, 17].

Fig. 5 shows the expected and static calibration errors for individual adult and neonatal detectors. In both cases, the expected calibration and overconfidence errors were practically identical (data not shown). This means that the SDAs are overconfident in their predictions, i.e. they are incorrect more frequently than the probabilities returned by the SDAs indicate. This can partly be explained by the use of ReLU activation functions [19] and batch normalization layers [16] in the detectors. The calibration may have been further compromised due to the training of the detectors using binary labels (non-seizure/seizure) [50] and cross-entropy loss function [54].

For the neonatal data set the expected and static calibration errors were very similar since the fraction of seizure segments in left-out neonatal patients is 49 % whereas for the adult test set only 17 % of the segments are seizure segments. Since static calibration error is an average of expected calibration errors calculated separately for seizure and non-seizure segments, it is more suitable for imbalanced data sets (e.g., the adult data set) than the expected calibration error.

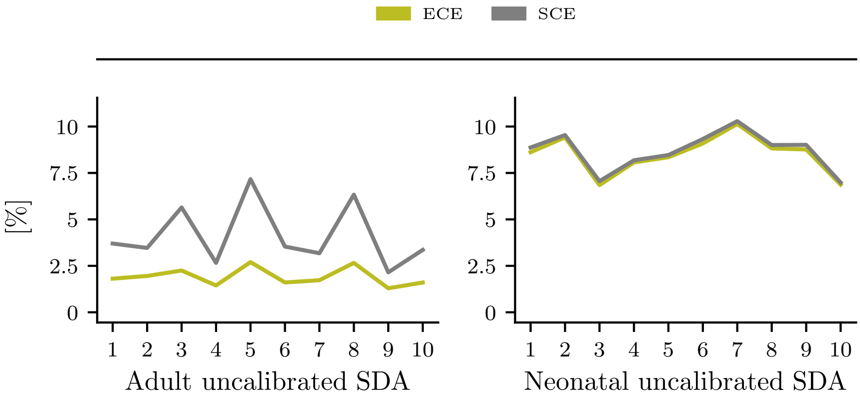


Fig. 5: Expected calibration error (ECE) and static calibration error (SCE) for individual uncalibrated adult (left) and neonatal (right) SDAs from the deep ensembles. The metrics are calculated based on all available segments in the test set for adult SDAs and left-out patients for neonatal SDAs.

Due to the imbalance in the data sets we decided to investigate the calibration of seizure and non-seizure classes individually. Fig. 6 shows that lower sensitivity/specificity results in higher expected calibration error, i.e. in worse calibration. This is a consequence of most (above 83 %) segments being predicted with confidence greater than 0.9. From equation (4) it follows that the bin with segments predicted with the highest confidence affects the expected calibration error the most. Therefore, if the overall accuracy is very high and also close to the overall confidence, the detector is well-calibrated. The figure shows that the expected calibration error is higher for the seizure segments than for the non-seizure segments. However, there are more non-seizure segments in the adult test set and therefore calibration on the non-seizure segments weighs more in the computation of the expected calibration error.

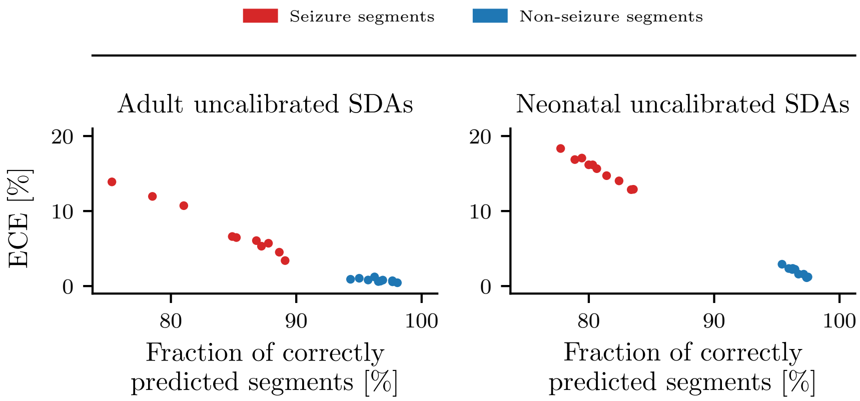


Fig. 6: Expected calibration error (ECE) calculated using seizure (red) and non-seizure (blue) segments. Each red and blue point represents one uncalibrated adult (left) and neonatal (right) SDA in the deep ensembles. The metrics are calculated from all available segments in the test set for the adult SDAs and from the left-out patients for neonatal SDAs.

3.3 Calibrated SDAs

The AUC was averaged over all SDAs in an ensemble and the individual classifier with AUC closest to the ensemble average was selected as a representative uncalibrated SDA in the following. Table 2 shows how different calibration methods affect the performance of the adult and neonatal SDAs. The classification performance of the adult SDAs using temperature scaling is identical to the uncalibrated classifier since the scaling procedure only affects confidence estimates and not the predicted class. The method is therefore not listed separately in the table. Temperature scaling was not applied to the neonatal SDAs since there was no dedicated validation set available for tuning the temperature parameter τ.

The performance metrics in Table 2 have fairly large variance across patients for all the SDAs, the sensitivity metric in particular. A likely explanation is that there are far fewer seizure segments (13 %) in the training set, in comparison to non-seizure segments (Table 1). Although each mini-batch is balanced during training, the seizure class has fewer

Table 2: Patient-based classification metrics for uncalibrated and calibrated adult and neonatal SDAs. Metrics are averaged across patients with at least one 16 seconds long seizure segment. For uncalibrated detectors, the range of values for all detectors in the ensemble is reported. Standard deviations are shown in parentheses.

	Uncalibrated	Calibrated		
		Ensemble	Dropout	Mixup
Adult SDA				
Area under the curve	0.94 (0.11)	0.95 (0.11)	0.95 (0.11)	0.96 (0.09)
Sensitivity [%]	77.74 (26.51)	79.72 (25.94)	79.57 (26.26)	77.77 (23.62)
Specificity [%]	96.44 (5.44)	97.51 (4.07)	97.0 (4.14)	96.45 (5.03)
Neonatal SDA				
Area under the curve	0.93 (0.12)	0.95 (0.10)	0.92 (0.14)	0.92 (0.13)
Sensitivity [%]	71.71 (30.77)	74.11 (27.84)	71.67 (28.68)	68.94 (31.48)
Specificity [%]	97.11 (7.04)	97.28 (7.95)	94.07 (12.77)	96.91 (5.98)

examples defining it. Furthermore, the metrics are only based on 31 adult and 38 neonatal patients, respectively, and differences in performance for 2 – 3 patients can end up having a significant effect on the overall mean and standard deviation. In both data sets, there are approx. three patients that the SDAs consistently had problems classifying and result in sensitivity values below 50 % and/or specificity below 90 %. The heterogeneity of the different seizure types and ambiguity in EEG signals are likely to contribute to the variability as well. From Table 2 we conclude that calibration methods do not outperform uncalibrated detectors, but they also do not noticeably degrade classification performance in terms of the area under the curve, sensitivity and specificity. This is in line with previous studies which applied ensembling, dropout and mixup for image classification [23, 25, 49, 50, 57].

The calibration metrics in Table 3 were computed by averaging over all available test segments, instead of first computing the corresponding metrics over the patients and then averaging. The reason is that the number of segments behind each patient varies widely. For some patients, there are fewer than 100 segments and this causes difficulties when computing metrics based on confidence bins. Overall, large improvements in calibration were not observed for either data set. However, we observe that all the calibration methods reduce the overconfidence error, a highly desired feature in medical applications. An overconfidence error close to zero for the adult data set implies that the calibrated SDAs are mainly underconfident in their predictions since the expected calibration errors are non-zero. The neonatal SDAs are overconfident, after employing the calibration methods, but the level of overconfidence error has decreased.

Table 3: Segment-based calibration metrics for uncalibrated and calibrated adult and neonatal SDAs. The metrics were calculated on all available segments in a test set for the adult SDAs and on the left-out patients for the neonatal SDAs.

	Uncalibrated	Calibrated			
		Temp. scaling	Ensemble	Dropout	Mixup
Adult SDA					
Expected cal. error [%]	1.61	2.58	0.62	1.58	4.05
Overconfidence error [%]	1.55	0.0	0.03	0.0	0.0
Static calibration error [%]	3.54	3.21	2.02	2.31	3.65
Brier score	0.04	0.04	0.03	0.04	0.04
Negative log-likelihood	0.17	0.17	0.13	0.14	0.16
Neonatal SDA					
Expected cal. error [%]	8.76	-	4.72	4.80	1.76
Overconfidence error [%]	8.76	-	4.72	4.80	1.51
Static calibration error [%]	9.01	-	5.40	5.27	7.62
Brier score	0.10	-	0.08	0.09	0.11
Negative log-likelihood	0.59	-	0.29	0.32	0.35

For the neonatal SDAs, a large difference between expected and static calibration errors was not expected since the data set is balanced. For mixup, however, the two metrics differed considerably (Table 3). This indicates that the detector is overconfident for segments of one class and underconfident for the other. When the seizure and non-seizure components of the metrics are analyzed separately (Fig. 8), it appears that the SDA with mixup is overconfident in predicting seizure segments and not confident enough when predicting non-seizure segments.

The predicted confidence values are analysed in more detail in Fig. 9 where the confidence estimates of correct and incorrect predictions are analyzed separately. Diagrams with similar patterns are obtained when only seizure or non-seizure segments are used (data not shown). The uncalibrated SDAs are confident in the predictions, both correct and incorrect, and most of them have confidence close to one. As much as it is desired that the SDAs are confident in their correct predictions, it is also important that incorrect predictions are made with lower confidence, making it possible to inform the user that some parts of EEG are difficult to classify. In this case, binary seizure/non-seizure predictions can be accompanied by confidence values as illustrated in Fig. 7. For all the calibration methods, the number of incorrect predictions in the most confident bins is

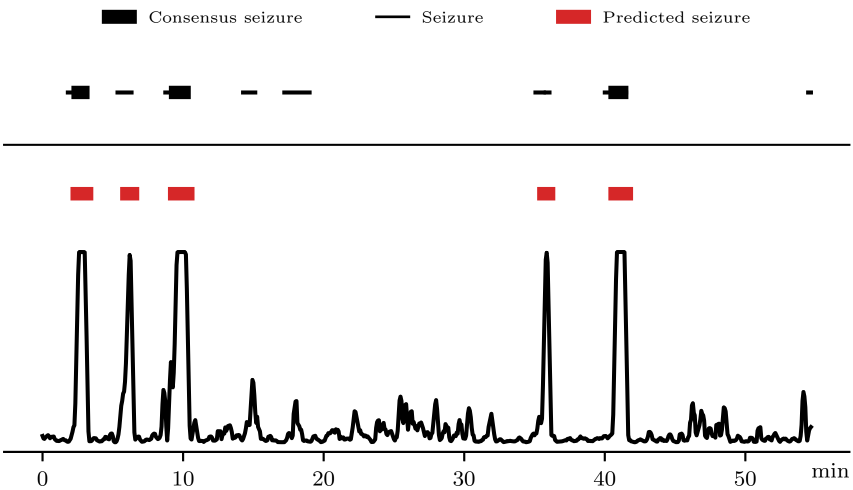


Fig. 7: Predictions for a short neonatal EEG recording (< 1 hour), obtained with an SDA employing dropout. Black blocks represent consensus seizures where the three human experts scoring the recording were in agreement. Black lines represent seizures annotated by at least one of the three experts. Red blocks represent seizure predictions and the corresponding probability estimates (confidence values) are denoted with a black curve.

clearly reduced, which is what we want, but the SDAs are less confident in their correct predictions compared to uncalibrated SDAs. However, in the latter case, the reduction is fairly small and mainly the bin with the second confidence increases in size.

Mixup results in an SDA with the lowest average confidence among the calibration methods studied here and with the lowest number of segments in the most confident bin. This also means that the largest number of incorrect predictions are, as preferred, predicted with confidence close to 0.5. In the clinical setting, this would imply that more segments would be passed for an additional review done by a human expert, but only a few incorrectly classified would be missed. This is especially noticeable for neonatal SDA. Note that the hyper-parameters were tuned on the adult validation set and different results could be obtained if they were to be tuned on neonatal data.

4 Conclusion

In line with a previous study [6], we find that uncalibrated SDAs tend to be overconfident in their predictions and the probability corresponding to an incorrect prediction gives little indication that the prediction is wrong. Since most predictions are made with confidence close to one, a more accurate detector is also better calibrated.

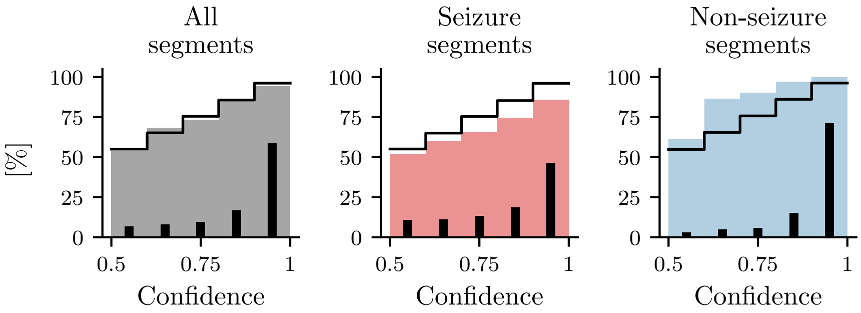


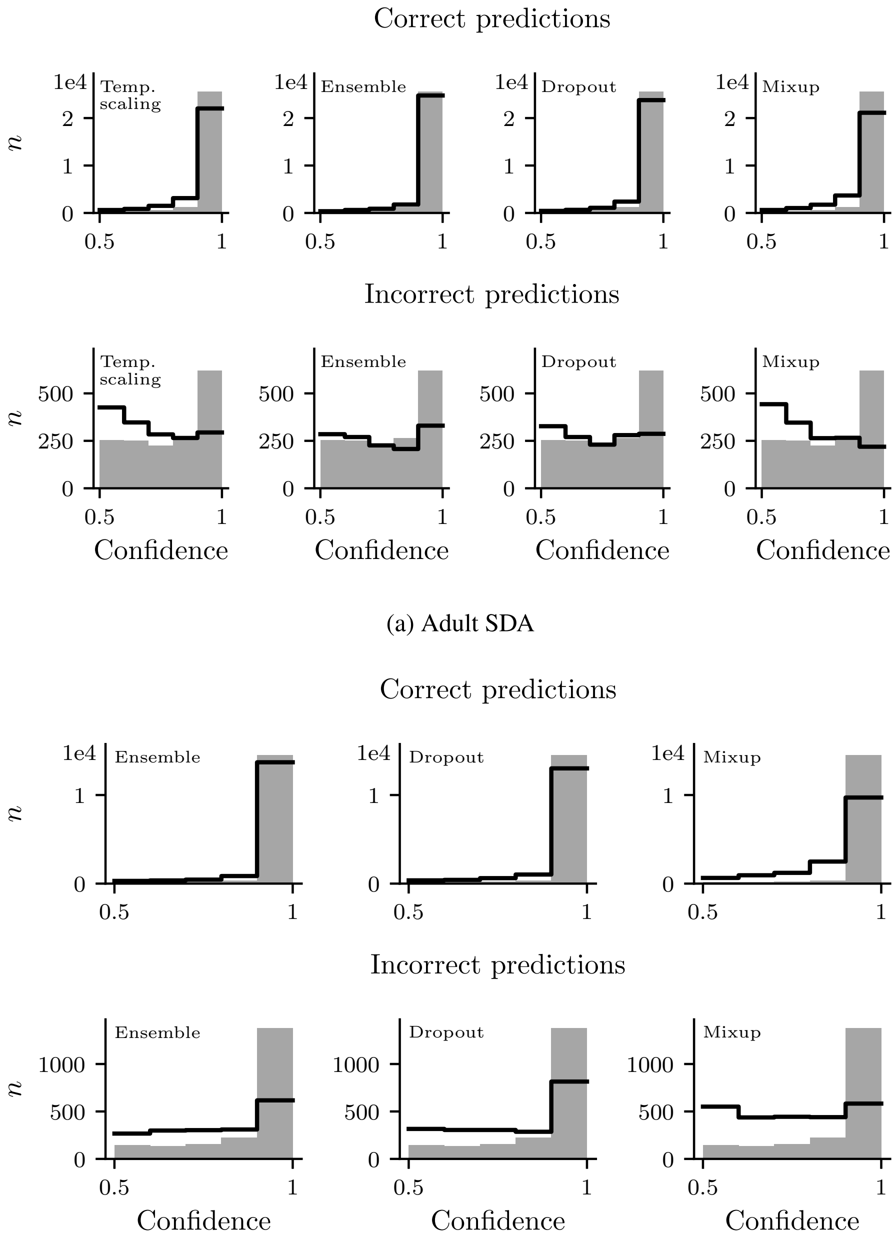
Fig. 8: Reliability diagrams for a neonatal SDA trained with mixup. The black step function indicates the average confidence of segments in each bin. Coloured bars indicate the fraction of correctly predicted segments in a bin. Black bars indicate the fraction of segments in a bin. The detector is overconfident for seizure segments and underconfident for non-seizure segments leading to an expected calibration error of 1.76 and a static calibration error of 7.62.

The four calibration methods included in this study did not degrade classification performance, i.e. their sensitivity, specificity and AUC values were similar to the uncalibrated SDAs. A slight improvement among the classification metrics for the adult SDA utilizing ensembling, dropout or mixup, was observed. These methods can be regarded as regularisation techniques that reduce model overfitting and improve generalization which in turn can explain increased detector accuracy.

With some additional computational costs we found a modest improvement in calibration, with the ensemble method giving the largest improvement, followed by dropout. Mixup gave mixed results for the adult SDA but did better on the neonatal data. Temperature scaling gave little improvement.

In [6] we found that dropout gave a noticeable improvement in expected calibration error for both adult and neonatal, SDAs. A possible explanation for this discrepancy is that here we are starting from a more accurate uncalibrated classifier than in our earlier work which then tends to be better calibrated [31]. The pre-processing and evaluation steps used here are slightly different from previous studies which also contributes to the difference. In this study, the EEG data was not normalized prior to feeding it to the network, and the mini-batch size was larger. This resulted in slightly more accurate uncalibrated SDAs than before. Additionally, here the non-seizure segments do not overlap and consequently, seizure segments represent a bigger portion of the test data. Since calibration on these segments tends to be worse than on the non-seizure segments, the overall expected calibration error is higher. Analyzing the calibration of each class is therefore advised in case of class imbalance.

All the calibrated detectors were noticeably less confident in incorrect predictions compared to uncalibrated detectors. EEG segments with low confidence values can then be



(a) Adult SDA

(b) Neonatal SDA

Fig. 9: Gray histograms represent the number of correctly and incorrectly classified segments of an arbitrarily chosen uncalibrated adult and neonatal SDA. The step functions represent the number of correctly and incorrectly classified segments.

passed to a human expert for manual review and eventual correction. This is a desirable property of an SDA if the main objective is to develop a detector that is as accurate as possible. However, in order for the SDA to be useful in clinical practice and make reviewing more time-efficient, the expert should not be given the majority of the acquired data for review. To limit the amount of data that requires human expertise, the confidence in correct predictions should be close to one and these EEG segments would therefore not be passed on to the expert. This pattern was observed for the ensemble and dropout, for temperature scaling and mixup the drop in confidence for correct predictions was larger and unfavourable.

Further work is needed to evaluate if such a setup makes EEG scoring more efficient in the clinical environment. Introducing confidence estimates alongside binary seizure/non-seizure predictions to the EEG monitors could however confound the interpretation of the user. A study on how to best present the confidence estimates is therefore needed. Another possibility would be to present to the user only a few selected EEG segments from which the detector would learn and subsequently improve. The segments could e.g., be chosen with an active learning approach [42].

Bibliography

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (2021)

[2] Becker, T., Vandecasteele, K., Chatzichristos, C., Van Paesschen, W., Valkenborg, D., Van Huffel, S., De Vos, M.: Classification with a deferral option and low-trust filtering for automated seizure detection. Sensors **21**(4), 1046 (2021)

[3] Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence **1**(1), 20–23 (2019)

[4] Borovac, A., Gudmundsson, S., Thorvardsson, G., Moghadam, S.M., Nevalainen, P., Stevenson, N., Vanhatalo, S., Runarsson, T.P.: Ensemble learning using individual neonatal data for seizure detection. IEEE journal of translational engineering in health and medicine **10**, 1–11 (2022)

[5] Borovac, A., Guðmundsson, S., Thorvardsson, G., Runarsson, T.P.: Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network. In: The NeurIPS 2021 Data-Centric AI Workshop (2021)

[6] Borovac, A., Runarsson, T.P., Thorvardsson, G., Gudmundsson, S.: Calibration of Automatic Seizure Detection Algorithms. In: 2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). pp. 1–6. IEEE (2022)

[7] Boylan, G., Burgoyne, L., Moore, C., O'Flaherty, B., Rennie, J.: An international survey of EEG use in the neonatal intensive care unit. Acta paediatrica **99**(8), 1150–1155 (2010)

[8] Brier, G.W., et al.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)

[9] Dereymaeker, A., Ansari, A.H., Jansen, K., Cherian, P.J., Vervisch, J., Govaert, P., De Wispelaere, L., Dielman, C., Matic, V., Dorado, A.C., et al.: Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: the neoguard eeg database. Clinical Neurophysiology **128**(9), 1737–1745 (2017)

[10] Eicher, J., Bild, R., Spengler, H., Kuhn, K.A., Prasser, F.: A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. BMC Medical Informatics and Decision Making **20**(1), 1–14 (2020)

[11] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)

[12] Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342 (2021)

[13] Gotman, J.: Automatic detection of epileptic seizures. Handbook of clinical neurophysiology **3**, 155–165 (2003)

[14] Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al.: MEG and EEG data analysis with MNE-Python. Frontiers in neuroscience p. 267 (2013)

[15] Grewal, S., Gotman, J.: An automatic warning system for epileptic seizures recorded on intracerebral EEGs. Clinical neurophysiology **116**(10), 2460–2472 (2005)

[16] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

[17] Halford, J., Shiau, D., Desrochers, J., Kolls, B., Dean, B., Waters, C., Azar, N., Haas, K., Kutluay, E., Martz, G., et al.: Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. Clinical Neurophysiology **126**(9), 1661–1669 (2015)

[18] Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M., Tobochnik, S.: The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In: 2014 IEEE signal processing in medicine and biology symposium (SPMB). pp. 1–5. IEEE (2014)

[19] Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41–50 (2019)

[20] Hrachovy, R.A., Mizrahi, E.M.: Atlas of neonatal electroencephalography. Springer Publishing Company (2015)

[21] Isaev, D.Y., Tchapyjnikov, D., Cotten, C.M., Tanaka, D., Martinez, N., Bertran, M., Sapiro, G., Carlson, D.: Attention-based network for weak labels in neonatal seizure detection. Proceedings of machine learning research **126**, 479 (2020)

[22] Jones, J.E., Hermann, B.P., Barry, J.J., Gilliam, F.G., Kanner, A.M., Meador, K.J.: Rates and risk factors for suicide, suicidal ideation, and suicide attempts in chronic epilepsy. Epilepsy & Behavior **4**, 31–38 (2003)

[23] Ju, C., Bibaut, A., van der Laan, M.: The relative performance of ensemble methods with deep convolutional neural networks for image classification. Journal of Applied Statistics **45**(15), 2800–2818 (2018)

[24] Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine **4**(1), 1–6 (2021)

[25] Krishnan, R., Tickoo, O.: Improving model calibration with accuracy versus uncertainty optimization. Advances in Neural Information Processing Systems **33**, 18237–18248 (2020)

[26] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)

[27] Lamberink, H.J., Otte, W.M., Bluemcke, I., Braun, K.P., Aichholzer, M., Amorim, I., Aparicio, J., Aronica, E., Arzimanoglou, A., Barba, C., et al.: Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. The Lancet Neurology **19**(9), 748–757 (2020)

[28] Le, V.T., Abdi, H.H., Sánchez, P.J., Yossef, L., Reagan, P.B., Slaughter, L.A., Firestine, A., Slaughter, J.L.: Neonatal antiepileptic medication treatment patterns: a decade of change. American journal of perinatology **38**(05), 469–476 (2021)

[29] Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports **7**(1), 1–14 (2017)

[30] Litt, B., Echauz, J.: Prediction of epileptic seizures. The Lancet Neurology **1**(1), 22–30 (2002)

[31] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems **34**, 15682–15694 (2021)

[32] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. pp. 625–632 (2005)

[33] Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring Calibration in Deep Learning. In: CVPR Workshops. vol. 2 (2019)

[34] Noachtar, S., Rémi, J.: The role of EEG in epilepsy: a critical review. Epilepsy & Behavior **15**(1), 22–33 (2009)

[35] Olmi, B., Frassineti, L., Lanata, A., Manfredi, C.: Automatic Detection of Epileptic Seizures in Neonatal Intensive Care Units Through EEG, ECG and Video Recordings: A Survey. IEEE Access **9**, 138174–138191 (2021)

[36] O'Shea, A., Lightbody, G., Boylan, G., Temko, A.: Investigating the impact of CNN depth on neonatal seizure detection performance. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 5862–5865. IEEE (2018)

[37] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[38] Perucca, E., Brodie, M.J., Kwan, P., Tomson, T.: 30 years of second-generation antiseizure medications: impact and future perspectives. The Lancet Neurology **19**(6), 544–556 (2020)

[39] Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

[40] Quinonero-Candela, J., Rasmussen, C.E., Sinz, F., Bousquet, O., Schölkopf, B.: Evaluating predictive uncertainty challenge. In: Machine Learning Challenges Workshop. pp. 1–27. Springer (2006)

[41] Razavi, B., Rao, V.R., Lin, C., Bujarski, K.A., Patra, S.E., Burdette, D.E., Geller, E.B., Brown, M.G.M., Johnson, E.A., Drees, C., et al.: Real-world experience with direct brain-responsive neurostimulation for focal onset seizures. Epilepsia **61**(8), 1749–1757 (2020)

[42] Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)

[43] Saminu, S., Xu, G., Shuai, Z., Abd El Kader, I., Jabire, A.H., Ahmed, Y.K., Karaye, I.A., Ahmad, I.S.: A recent investigation on detection and classification of epileptic seizure techniques using EEG signal. Brain Sciences **11**(5), 668 (2021)

[44] Schuele, S.U.: Effects of seizures on cardiac function. Journal of clinical neurophysiology **26**(5), 302–308 (2009)

[45] Scott, R.C.: What are the effects of prolonged seizures in the brain? Epileptic Disorders **16**(s1), S6–S11 (2014)

[46] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)

[47] Stevenson, N.J., Tapani, K., Lauronen, L., Vanhatalo, S.: A dataset of neonatal EEG recordings with seizure annotations. Scientific data **6**, 190039 (2019)

[48] Temko, A., Thomas, E., Marnane, W., Lightbody, G., Boylan, G.: EEG-based neonatal seizure detection with support vector machines. Clinical Neurophysiology **122**(3), 464–473 (2011)

[49] Thagaard, J., Hauberg, S., Vegt, B.v.d., Ebstrup, T., Hansen, J.D., Dahl, A.B.: Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 824–833. Springer (2020)

[50] Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems **32** (2019)

[51] Uria-Avellanal, C., Marlow, N., Rennie, J.M.: Outcome following neonatal seizures. In: Seminars in Fetal and Neonatal Medicine. vol. 18, pp. 224–232. Elsevier (2013)

[52] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods **17**(3), 261–272 (2020)

[53] Webb, L., Kauppila, M., Roberts, J.A., Vanhatalo, S., Stevenson, N.J.: Automated detection of artefacts in neonatal EEG with residual neural networks. Computer Methods and Programs in Biomedicine **208**, 106194 (2021)

[54] Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning. pp. 23631–23644. PMLR (2022)

[55] Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)

[56] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

[57] Zhang, Z., Dalca, A.V., Sabuncu, M.R.: Confidence calibration for convolutional neural networks using structured dropout. arXiv preprint arXiv:1906.09551 (2019)

Paper VI

Nurses' experiences and perspectives on (a)EEG monitoring in neonatal care: A qualitative study

Xiaowan Wang, Ana Borovac, Agnes van den Hoogena, Maria L. Tataranno, Manon J. N. L. Benders, and Jeroen Dudink

Journal of Neonatal Nursing, 2023

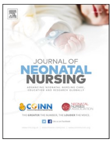
ARTICLE IN PRESS

Journal of Neonatal Nursing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Neonatal Nursing

journal homepage: www.elsevier.com/locate/jnn



Nurses' experiences and perspectives on aEEG monitoring in neonatal care: A qualitative study

Xiaowan Wang [a,1], Ana Borovac [b,c,1], Agnes van den Hoogen [a], Maria Luisa Tataranno [a,d], Manon J.N.L. Benders [a,d], Jeroen Dudink [a,d,*]

[a] Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, the Netherlands
[b] Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, Reykjavík, Iceland
[c] Kvikna Medical ehf., Reykjavík, Iceland
[d] Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Handling Editor: Ms L Altimier

Keywords:
aEEG
NICU nurses
Qualitative research
Seizure detection

ABSTRACT

Purpose: This study aimed to gather nurses' experiences and perspectives regarding the amplitude-integrated electroencephalogram (aEEG) monitoring system in neonatal intensive care units (NICUs) and to explore potential avenues for future improvements.

Design and Methods: This study employed a descriptive qualitative design. Semi-structured interviews were conducted with 20 nurses from the level-III NICU of a Dutch medical center. The collected interview data were analyzed using thematic analysis.

Results: Seven main themes emerged: training in aEEG monitoring, proficiency in aEEG electrode placement and pattern interpretation, usual practices of using aEEG, neonatologist-nurse cooperation on aEEG, the performance of the automated seizure detection software, the usefulness of aEEG monitoring in the NICU, and feedback about the current aEEG monitoring system.

Conclusions: Nurses confirmed that aEEG is a valuable tool for cerebral function monitoring in the NICU; however, improvements are necessary. For better utilization of aEEG in the NICU, it is recommended to enhance nurses' aEEG knowledge and skills and apply state-of-art techniques to improve the monitoring system.

Practice implications: To enhance the aEEG knowledge of NICU nurses, we suggest introducing structured training programs, conducting routine case-centered discussions, and creating readily available reference resources. To optimize the aEEG monitoring system, it is essential to incorporate innovative electrodes, provide remote accessibility, integrate advanced algorithms, and develop an intuitive graphical user interface.

1. Introduction

Infants admitted to the neonatal intensive care unit (NICU) often have a substantial risk for brain injury or dysfunction. Hence, there is a growing interest in the continuous monitoring of brain function for these vulnerable neonates, which can enable targeted brain-focused care and outcome prediction (Bonifacio and Van Meurs, 2019). One valuable tool for achieving this purpose is the amplitude-integrated electroencephalogram (aEEG), a simplified form of EEG using only two or four electrodes (Bonifacio and Van Meurs, 2019; El-Dib et al., 2023a; El-Dib et al., 2023b). The aEEG provides a time-compressed display of raw EEG signals on a semi-logarithmic scale. When used alongside its corresponding raw EEG traces, the aEEG can aid healthcare professionals in the early detection of brain dysfunctions such as seizures, thereby enabling timely and appropriate interventions (Rakshasbhuvankar et al., 2015; Variane et al., 2017). For these reasons, the aEEG has become increasingly popular in NICUs worldwide (Bruns et al., 2017; Tao and Mathur, 2010; Wang et al., 2021).

Nurses provide around-the-clock care at the infant's bedside and are the primary users of aEEG monitors. Their experiences and perspectives on aEEG usage in the NICU can, therefore, provide unique and valuable insights into improving the current monitoring systems. Despite this,

[*] Corresponding author. Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Lundlaan 6, 3584 EA, Utrecht, the Netherlands.

E-mail address: j.dudink@umcutrecht.nl (J. Dudink).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jnn.2023.08.003
Received 14 June 2023; Received in revised form 9 August 2023; Accepted 13 August 2023

Please cite this article as: Xiaowan Wang et al., *Journal of Neonatal Nursing*, https://doi.org/10.1016/j.jnn.2023.08.003

existing aEEG research has largely neglected this group and their viewpoints.

To address this research gap, we conducted semi-structured interviews with NICU nurses from a hospital in the Netherlands. Our primary objective was to gather the nurses' perspectives on the usefulness of aEEG monitoring in the NICU and to identify any concerns they might have regarding the current monitoring system. Based on the feedback received, we explored potential avenues for improving the aEEG monitoring system in the future. In addition, we assessed the nurses' proficiency in using aEEG monitors, including their ability to independently place electrodes and interpret background patterns, as well as their cooperation with neonatologists on aEEG usage. For example, can nurses perform the first interpretation and know when to ask for help and support from the clinicians? In doing so, we are able to determine whether additional training or support was needed for the nurses, ultimately improving the overall care in the NICU setting.

2. Methods

2.1. Study design

This study employed a descriptive qualitative approach with semi-structured individual interviews to explore nurses' experiences and perspectives on using aEEG monitoring in the NICU (Moser and Korstjens, 2018; Sandelowski, 2000; 2010). The preparation of this study was carried out in accordance with the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist (Tong et al., 2007).

2.2. Setting and participants

This study was conducted at the Neonatology department of the Wilhelmina Children's Hospital (WKZ), University Medical Center (UMC) Utrecht, The Netherlands. The WKZ is one of the ten locations that provide intensive care (IC) in the Netherlands. The IC consists of three units, each with a capacity of eight beds. These IC units are used to treat preterm infants with intensive care needs, such as extremely (<28 weeks of gestation) and very preterm (28–32 weeks of gestation) infants, very small infants with birth weight <1000 g, and infants born term with a need for cardio-respiratory support and/or intensive monitoring.

In addition, the WKZ has a temporary stopover area between the IC and medium care divisions, which is called high care (HC, also known as post-IC in some hospitals). The HC is a single unit of eight beds that provides treatments for infants that need less acute medical care and have higher weights than IC patients. However, as infants admitted to the HC are still young (e.g., very preterm infants), the HC is merged into IC in some other hospitals. Hence, this study focused on both IC and HC units.

The inclusion criteria of this study were nurses and physician assistants (PAs) from IC and HC units, as both are often present at the bedside and responsible for visually interpreting aEEG recordings. In each IC or HC unit, three or four nurses are scheduled to provide daily care for infants. In each IC unit, one or two PAs are scheduled to monitor the infants' progress and ensure continuity and quality of care. For simplicity, nurses and PAs are collectively hereafter referred to as "nurses", except when PAs need to be mentioned separately.

At the time of this study, two-channel aEEG monitoring was performed as standard care for extremely preterm infants in their first three days after birth and for any infant at risk for seizures and perioperative complications in the WKZ. The aEEG usage demands in IC units are typically much higher than in the HC unit, indicating that IC nurses usually have more experience in both aEEG application and interpretation than HC nurses.

The aEEG monitoring is performed using the BrainZ monitor (Natus Medical Inc., Seattle, WA) in the WKZ. Subcutaneous needle electrodes are used and placed over the frontal and parietal lobes (F3, F4, P3, P4). The reference electrode is placed at Cz. The monitor displays one-(P3–P4) or two-channel (left: F3–P3, right: F4–P4) aEEG and the corresponding raw EEG traces. Automated alert software for seizure detection is built into the BrainZ monitor for clinical screening purposes. A laminated card providing typical neonatal aEEG patterns (including background activity, sleep-wake cycling, and seizure activity) is attached to each aEEG monitor for nurses as reference.

2.3. Data collection

We performed face-to-face semi-structured interviews with IC and HC nurses in May 2022 at the hospital. Based on discussions with the whole research team, an interview guide consisting of 11 points was designed to gain nurses' sociodemographic characteristics and responses to aEEG-related questions (Supplementary Table S1). After obtaining verbal informed consent from nurses, AB and XW carried out individual interviews based on the guide. Each interview lasted approximately 15–20 min and was recorded in written notes.

2.4. Data analysis

The interview data were digitally transcribed and inductively analyzed by XW and AB according to Braun and Clark's six-step thematic analysis approach, including (1) familiarising with the data, (2) initial coding, (3) identifying broader themes, (4) reviewing themes, (5) defining themes, and (6) producing the report (Braun and Clarke, 2006). To ensure trustworthiness, these steps were rigorously followed throughout the analysis. Any uncertainty or disagreements were resolved via discussion between the two researchers or via consultation with a researcher (AvdH) with expertise in qualitative analysis. All interview data were anonymized before analysis, and each nurse was allocated a participant number (e.g., N1, N2, etc.).

3. Results

Twenty participants completed the interviews (nineteen females and one male, allocated numbers: N1–N20), comprising twelve IC nurses, four HC nurses, three student nurses, and one PA. They reported a mean of 16.8 years (SD = 12.5, range: 0.3–40.0) of working as a nurse or PA, and a mean of 7.8 years (SD = 6.1, range: 0.2–20.0) of experience with aEEG.

After detailed analysis and interpretation of the interview data, we identified seven main themes. The first two themes explored the main resources that nurses could utilize to develop their competency in aEEG and how they evaluated their proficiency in this area. The next three themes discussed nurses' experiences with aEEG and the practical challenges they faced in their daily practice. Finally, the last two themes delved into nurses' perspectives and feedback regarding the use of the aEEG monitoring system in the NICU. Each theme is described in detail below, underpinned by quotes from the participants.

3.1. Training in aEEG monitoring

Nurses gained foundational knowledge on aEEG during their formal mandatory education and developed practical skills through working experience and interactions with more experienced colleagues. Furthermore, the UMC Utrecht provides e-learning courses and clinical guidelines to help nurses reinforce and update their existing knowledge regarding aEEG.

> N14: "I learned about aEEG from education, e-learning, on the job, and from colleagues."

3.2. Proficiency in aEEG electrode placement and interpretation

A few nurses encountered challenges with aEEG electrode placement. This was largely due to their concerns over the use of subcutaneous needle electrodes which could cause pain and skin infection. The implementation of aEEG for very small infants, infants with thick hair, or infants with excessive movement also posed challenges.

> N12: "Application of needle electrodes is hard for very small infants (<800 g) as their heads are small or when an infant has a lot of hair as the tape does not stick."

Regarding aEEG interpretation, almost all nurses reported difficulties, with less experienced nurses finding it more challenging. Several nurses found it hard to distinguish between artifacts and seizures, and some perceived aEEG interpretation as subjective and ambiguous. Additionally, a few nurses struggled with interpreting aEEG patterns of very small infants, such as extremely preterm infants.

3.3. *Usual practices of using aEEG*

Most nurses used both aEEG and corresponding raw EEG traces during their daily care work. Some of them mentioned that they checked aEEG first and then verified the results using raw EEG. Nonetheless, a few less-experienced nurses relied solely on aEEG because of its simpler interpretation compared to raw EEG.

The majority of nurses checked aEEG monitors once or twice per hour. A few nurses indicated that they always gave a glance at the aEEG monitor remotely from the nurses' station, besides performing a careful check. Nearly half of the nurses chose to adjust their checking frequency according to the infant's medical condition, with more frequent checks for those with severe illness or seizures.

> N1: "I check the aEEG monitor at least twice per hour and more often if the infant is very ill."

3.4. *Neonatologist-nurse cooperation on aEEG*

Neonatologists checked the aEEG monitors regularly during their shifts. Still, they would be called by the nurses for an additional consultation in case of suspected seizures or any other abnormal activities or when the nurses were uncertain about the interpretation. If seizures were diagnosed, the nurses would seek guidance from neonatologists to determine the next course of action, which might include treatment or other options. One senior nurse, who gained more experience than some neonatologists, mentioned that she only requested well-experienced neonatologists' help.

> N10: "I ask for a neonatologist's support when there is automated seizure detection, when I am in doubt about the interpretation of the traces, or when there is something worrying."

3.5. *Performance of the automated seizure detection software*

The most experienced nurses used the automated seizure detection software as a guide to select points of interest while examining the entire aEEG recordings. Only a few nurses immediately called a neonatologist for consultations once an alert was triggered by the seizure detection software. The rest of the nurses (together with colleagues if needed) confirmed or rejected the detections by inspecting the raw EEG traces. The seizure detection software might miss seizures or, more frequently, falsely detect seizures (i.e., non-seizure activity was alerted).

> N20: "Most problems are related to falsely detected seizures rather than missed ones."

According to nurses, false alarms were mainly associated with movement artifacts and rhythmic artifacts related to mechanical ventilation, respiratory artifacts, electrocardiograms, hiccups, and cooling. Furthermore, false detections could also be caused by incorrectly placed electrodes, poor electrode connection due to hair, or when an infant was lying on one or both parietal electrodes (P3 and/or P4). Since there were only four electrodes used for the acquisition, some seizure activity could be missed if their origin was far away from the electrodes.

3.6. *The usefulness of aEEG monitoring in the NICU*

While aEEG monitoring, at the time of the study, was part of the standard care for extremely preterm infants in their first three days after birth in the WKZ, only two less experienced nurses—one HC nurse and one student nurse—believed that this was necessary. In contrast, most nurses deemed that aEEG should only be used when there are indications, given that needle electrodes could have adverse effects on the skin of vulnerable infants.

One of the most commonly mentioned indications for aEEG monitoring was seizures or suspicion thereof. According to the nurses, seizures could be caused by cerebral bleeding (e.g., intraventricular hemorrhage), perinatal asphyxia, or elevated serum bilirubin levels, and were suspected in the presence of abnormal movements or apnea (only in term infants). The aEEG monitoring was also useful during and after surgeries to detect potential hypotension, hypoxia, fluctuations in cerebral blood flow, and hyper- or hypocapnia. When brain injury was observed through another diagnostic technique, such as ultrasound, aEEG monitoring was used for observation and outcome prediction purposes.

> N6: "The monitoring is especially useful for seizure detection or other clinical problems, but not necessary for extremely preterm infants in their first three days."

3.7. *Feedback about the current aEEG monitoring system*

Almost all nurses expressed their dissatisfaction with the use of needle electrodes in aEEG monitoring (particularly for very small infants) and made suggestions for optimal electrode design based on their own clinical practice. Most importantly, ideal EEG electrodes should be non-invasive and made of soft, comfortable, and skin-friendly materials that are suitable to use over long periods of time.

The nurses emphasized another critical consideration: an ideal electrode should produce EEG signals of high quality and good stability. These signals should be robust against heartbeat- and respiration-induced artifacts, and the electrodes should stay in place even when the infant is highly active.

Furthermore, the optimal EEG electrodes should be easy to apply and allow accurate and consistent placement. For instance, the electrode surface could be color-coded and/or labeled with the electrode positions (e.g., F3) to ease the application process. To further simplify the application procedure, better electrode-skin contact without the need for hair shaving is also desired. Moreover, a wireless or mobile aEEG monitoring system enabling flexible recordings would be especially valuable in the busy NICU setting where the patients are relocated frequently, e.g., for a magnetic resonance imaging (MRI) examination.

> N11: "Ideal electrodes should enable good connection and be non-invasive. Also, we do not want to sacrifice (the infant's) hair."

Additionally, from several nurses' perspectives, a better aEEG monitoring system should also be compatible with other monitoring techniques such as near-infrared spectroscopy (NIRS) and existing devices (e.g., continuous positive airway pressure). The aEEG monitor should not only allow nurses to quickly record common events (e.g., seizures) in real time but also enable them to retrospectively add event markers to previous time points (especially when they are too busy to record events in real-time) or add detailed comments to a recorded event.

4. Discussion

Nurses are primary users of the aEEG monitoring system in the NICU and are commonly present at the bedside much longer than neonatologists. With this qualitative study, we investigated nurses' experiences and opinions of aEEG monitoring in neonatal care. Here we will discuss current major problems and future directions for improvement.

4.1. Summary of current problems

Neonatal nurses' knowledge and skills in aEEG monitoring vary, and their competency in this area is mainly improved through working experience. While there are some courses and talks available, these resources often provide only fragmented information, which is insufficient for nurses to form a comprehensive understanding of aEEG knowledge.

The inadequate understanding of aEEG among nurses has resulted in a notable discrepancy between their perception of its usefulness and its actual value in certain cases. For instance, utilising aEEG monitoring as part of the standard care for extremely preterm infants during their first three days after birth in the WKZ is backed by mounting evidence that supports the important role of aEEG in predicting subsequent brain growth and long-term outcomes for this population (Benders et al., 2015; Klebermass et al., 2011; Song et al., 2015; van 't Westende et al., 2022; Wikström et al., 2012). However, most nurses are unaware of these findings in their daily work, and many even question the necessity of this practice, citing the minimal brain activity observed in these infants. This knowledge gap can ultimately affect the quality of care provided.

Furthermore, the interpretation of aEEG recordings is challenging for most nurses, especially when it comes to distinguishing between artifacts and seizure activity, as they can appear very similar (Rakshasbhuvankar et al., 2015). Although the automated seizure detection software is a valuable addition to the aEEG monitoring system currently used in the WKZ, it can sometimes incorrectly identify artifacts as seizures. Moreover, the problem of electrode misplacement further complicates the interpretation process. Additionally, aEEG interpretation is a subjective and ambiguous process that requires special expertise, which is why nurses often seek consultation from neonatologists. It is also worth noting that several less-experienced nurses rely solely on aEEG traces, which may lead to missing short seizures and misinterpreting cortical activity due to artifacts (Rakshasbhuvankar et al., 2015).

To improve the current aEEG monitoring system, nurses gave suggestions from a user experience perspective. They stated that a good aEEG monitor should be safe, easy to use, compatible with existing devices, and should provide reliable and stable long-term recordings.

4.2. Future directions

4.2.1. Building knowledge and expertise

Systematic training programs. At the time of the study, the aEEG training provided by the WKZ was experienced as infrequent and inadequate, which made it difficult for nurses to develop a comprehensive understanding of aEEG and the underlying motivations of some aEEG monitoring practices. Thus, we recommend the development of a more systematic and structured training procedure.

To cater to the varying needs of different nurses, this training procedure should consist of different types of programs targeting nurses at different stages and ages. For instance, an orientation program should be designed for beginner nurses, providing them with basic aEEG implementation skills and essential knowledge, such as identifying typical aEEG background patterns. Furthermore, ongoing training through workshops, seminars, and refresher courses is necessary to keep nurses updated with the latest research findings on aEEG, such as its role in long-term outcome prediction and application to other neonatal groups.

Regular case-based discussions. Given the subjective nature of aEEG interpretation, it is highly recommended to conduct frequent meetings to discuss several typical and ambiguous cases. These meetings provide a platform for neonatologists and nurses to visually review the newest aEEG recordings and make decisions collaboratively. Regular discussion sessions also offer more experienced nurses and neonatologists the opportunity to share their aEEG knowledge, facilitating effective communication and cooperation among team members, ultimately leading to improved quality of care. Furthermore, such discussions help define more detailed and clearer aEEG classification criteria, which in turn enhances interpretation accuracy.

Pocketbook and in-house knowledge engine. The laminated reference card that comes with each aEEG monitor has limited utility in certain cases. It only provides typical aEEG pattern information, which may not be sufficient for nurses who require more detailed explanation and assistance when aEEG experts are absent in the NICU. To address this issue, we propose developing a pocketbook that includes detailed examples and answers to frequently asked questions. For example, a sample illustration of the aEEG background activity for an infant with hemorrhage would aid in timely diagnosis and treatment (Benavente-Fernández et al., 2015). To enhance portability, the pocketbook can be designed in an accordion fold pattern.

Furthermore, it would be beneficial to create a neonatal aEEG knowledge engine, similar to a wiki, that contains a wealth of information and resources such as electrode placement tips, sample cases, and essential literature tailored to nurses' needs. Such an engine would enable them to quickly locate the information they require, thereby improving the efficiency of care.

4.2.2. Applying state-of-art techniques

Novel electrodes. Needle electrodes are commonly used for neonatal aEEG monitoring in the WKZ as well as other hospitals, but they can cause discomfort and harm to an infant's vulnerable skin. Thus, there is a growing need for new aEEG electrodes made from comfortable materials and suitable for long-term monitoring. While gel electrodes are a popular alternative due to their non-invasive nature, they have their drawbacks such as the potential for allergic reactions and difficulty in washing off. Additionally, the signal quality decreases when the gel dries out, making them less suitable for long-term monitoring.

Dry electrodes have emerged as an attractive option for neonatal aEEG recording in the near future. This novel technology is non-invasive and does not require the application of conductive gel, thus allowing faster electrode placement, removal, and washing. Studies on adults have shown that dry electrodes are more comfortable and produce signals of sufficient quality, making them a preferred option over needle and gel electrodes (Hinrichs et al., 2020; Ng et al., 2022).

Automated analysis of aEEG traces. Automated seizure detection is an essential aspect of aEEG monitoring in the NICU, while the current seizure detection software used in the WKZ is prone to inaccuracies. To enhance its accuracy and reliability, there is a need to incorporate state-of-the-art techniques, such as deep neural networks, which have demonstrated significant improvements in seizure detection performance compared to traditional methods (Olmi et al., 2021). To ensure trust and transparency in the detection process (Kompa et al., 2021), it is crucial to provide confidence levels (Borovac et al., 2022) and specify the channels used for detection (Isaev et al., 2020). This will enable nurses (and other clinicians) to focus their attention to the parts of the recording that require more scrutiny, thereby increasing efficiency in everyday clinical practice.

To minimize false alarms caused by artifacts, an improved artifact detection software should be incorporated alongside the seizure detection software. Furthermore, to help nurses interpret aEEG recordings, the monitoring system should include other quantitative features such as inter-burst intervals, spectral power, and multiscale entropy (Wang et al., 2023).

Improved monitoring system. In addition to automated analysis, an improved aEEG monitoring system should also offer remote access to address the shortage of aEEG expertise in NICUs (Dilena et al., 2021). This would enable experienced clinicians to monitor patients remotely, which is especially helpful when they cannot be physically present at the NICU. Remote access also allows for a more efficient organization of the

NICU, where nurses can check the aEEG monitoring system from a central station without walking to the monitors.

To accommodate the busy environment of the NICU, the monitoring system should allow users to record past events, such as feeding, in addition to real-time recordings. This is important for later review and interpretation of the aEEG traces. For example, if the infant is moved to another bed, the recording may contain artifacts that do not reflect the electrocortical activity.

To simplify the use of the overall monitoring system in the NICU, it would be beneficial if aEEG monitors could be integrated with other monitoring devices, such as NIRS (Bonifacio and Van Meurs, 2019). This would reduce the number of displays in the crowded NICU, prevent duplication of patient information, and decrease the likelihood of human error.

There are several monitoring systems available on the market that include some of these desired features. For instance, nëo (eemagine Medical Imaging Solutions GmbH, Germany) was specifically designed for use in a neonatal hospital environment. This system has automated aEEG pattern analysis (e.g., bursting activity and seizures) and allows the recording of an eight-channel EEG. In addition, there are also more generic EEG monitoring systems that offer neonatal functionalities, e.g., StratusEEG (Kvikna Medical ehf., Iceland), Neurofax EEG-1200 (Nihon Kohden Corporation, Tokyo), and NEUROWERK (Micromed S.p.A., Italy).

4.3. Limitations

The present work has several limitations. As the data was collected from a single center with a specific focus on neonatal neurology, our findings might not be generalized to nurses from other places or nurses with different backgrounds. Future research should include more nurses from different hospital settings. Response bias might have existed because the interviews were conducted in English, however, the mother tongue of the interviewed nurses was Dutch. Therefore, the answers might have been more exhaustive if the interviews had been conducted in Dutch. Additionally, several nurses who were unable to speak English did not join the interviews, probably causing sample bias. Finally, due to the nature of the qualitative analysis, a few answers that were hard to merged into a theme were excluded (as *outliers*), which might cause information loss.

5. Conclusion

Nurses exhibit a significant disparity in aEEG proficiency. Despite this, they have affirmed the usefulness of aEEG in several clinical indications in the NICU, particularly in detecting seizures. Nonetheless, the accuracy of the current automated seizure detection software requires improvement. Furthermore, the (minimally) invasive nature of needle electrodes sometimes deters nurses from advocating for prolonged aEEG monitoring of very small infants.

Nurses are key players in aEEG monitoring. Equipping them with sufficient aEEG knowledge and efficient monitors can substantially improve brain function monitoring for infants and, consequently, the overall care in the NICU. To build nurses' expertise, we recommend implementing systematic training, holding regular case-based discussions, and developing portable reference tools. To further improve the monitoring system, novel electrodes, remote access, advanced automated analysis algorithms, and a user-friendly graphical interface are necessary.

CRediT authorship contribution statement

Xiaowan Wang: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - Original draft, Writing- Reviewing and Editing.

Ana Borovac: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing- Original draft, Writing - Reviewing and Editing.

Agnes van den Hoogen: Methodology, Resources, Writing - Reviewing and Editing, Supervision.

Maria Luisa Tataranno: Resources, Writing - Reviewing and Editing, Supervision.

Manon J. N. L. Benders: Resources, Writing - Reviewing and Editing, Supervision.

Jeroen Dudink: Conceptualization, Methodology, Resources, Writing - Reviewing and Editing, Supervision, Project administration.

Ethical statement

The current study did not fall within the scope of the Medical Research Involving Human Subjects Act (abbreviation in Dutch: WMO) since participants involved in this study were not exposed to procedures or treatment and did not follow a certain behavioral strategy. This study was carried out following the Good Clinical Practice principles. All participants were verbally informed about the objectives of the current study and its voluntary and anonymous nature.

Funding source

This work was supported by the European Commission [Grant agreement number: EU H2020 MSCA-ITN-2018-#813483, INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS)].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jnn.2023.08.003.

References

Benavente-Fernández, I., Lubián-López, S.P., Jiménez-Gómez, G., Lechuga-Sancho, A.M., Garcia-Alloza, M., 2015. Low-voltage pattern and absence of sleep-wake cycles are associated with severe hemorrhage and death in very preterm infants. Eur. J. Pediatr. 174, 85–90.

Benders, M.J., Palmu, K., Menache, C., Borradori-Tolsa, C., Lazeyras, F., Sizonenko, S., Dubois, J., Vanhatalo, S., Hüppi, P.S., 2015. Early brain activity relates to subsequent brain growth in premature infants. Cerebr. Cortex 25, 3014–3024.

Bonifacio, S.L., Van Meurs, K., 2019. Neonatal neurocritical care: providing brain-focused care for all at risk neonates. Semin. Pediatr. Neurol. 32, 100774.

Borovac, A., Runarsson, T.P., Thorvardsson, G., Gudmundsson, S., 2022. Calibration of automatic seizure detection algorithms. In: 2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–6.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3, 77–101.

Bruns, N., Blumenthal, S., Meyer, I., Klose-Verschuur, S., Felderhoff-Müser, U., Müller, H., 2017. Application of an amplitude-integrated EEG monitor (cerebral function monitor) to neonates. JoVE, e55985.

Dilena, R., Raviglione, F., Cantalupo, G., Cordelli, D.M., De Liso, P., Di Capua, M., Falsaperla, R., Ferrari, F., Fumagalli, M., Lori, S., Suppiej, A., Tadini, L., Dalla Bernardina, B., Mastrangelo, M., Pisani, F., 2021. Consensus protocol for EEG and amplitude-integrated EEG assessment and monitoring in neonates. Clin. Neurophysiol. 132, 886–903.

El-Dib, M., Abend, N.S., Austin, T., Boylan, G., Chock, V., Cilio, M.R., Greisen, G., Hellström-Westas, L., Lemmers, P., Pellicer, A., Pressler, R.M., Sansevere, A., Szakmar, E., Tsuchida, T., Vanhatalo, S., Wusthoff, C.J., Bonifacio, S., Wintermark, P., Aly, H., Chang, T., Chau, V., Glass, H., Lemmon, M., Massaro, A., Wusthoff, C., deVeber, G., Pardo, A., McCaul, M.C., On behalf of the Newborn Brain Society, G., Publications, C, 2023a. Neuromonitoring in neonatal critical care part II: extremely premature infants and critically ill neonates. Pediatr. Res. 94, 55–63.

El-Dib, M., Abend, N.S., Austin, T., Boylan, G., Chock, V., Cilio, M.R., Greisen, G., Hellström-Westas, L., Lemmers, P., Pellicer, A., Pressler, R.M., Sansevere, A., Tsuchida, T., Vanhatalo, S., Wusthoff, C.J., Bonifacio, S., Wintermark, P., Aly, H., Chang, T., Chau, V., Glass, H., Lemmon, M., Massaro, A., Wusthoff, C., deVeber, G., Pardo, A., McCaul, M.C., on behalf of the Newborn Brain Society, G., Publications, C, 2023b. Neuromonitoring in Neonatal Critical Care Part I: Neonatal Encephalopathy and Neonates with Possible Seizures. Pediatr. Res. 94, 64–73.

Hinrichs, H., Scholz, M., Baum, A.K., Kam, J.W.Y., Knight, R.T., Heinze, H.-J., 2020. Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG system for clinical applications. Sci. Rep. 10, 5218.

Isaev, D.Y., Tchapyjnikov, D., Cotten, C.M., Tanaka, D., Martinez, N., Bertran, M., Sapiro, G., Carlson, D., 2020. Attention-based network for weak labels in neonatal seizure detection. In: Finale, D.-V., Jim, F., Ken, J., David, K., Rajesh, R., Byron, W., Jenna, W. (Eds.), Proceedings of the 5th Machine Learning for Healthcare Conference. PMLR, Proceedings of Machine Learning Research, pp. 479–507.

Klebermass, K., Olischar, M., Waldhoer, T., Fuiko, R., Pollak, A., Weninger, M., 2011. Amplitude-integrated EEG pattern predicts further outcome in preterm infants. Pediatr. Res. 70, 102–108.

Kompa, B., Snoek, J., Beam, A.L., 2021. Second opinion needed: communicating uncertainty in medical machine learning. npj Digit. Med. 4, 4.

Moser, A., Korstjens, I., 2018. Series: practical guidance to qualitative research. Part 3: sampling, data collection and analysis. Eur. J. Gen. Pract. 24, 9–18.

Ng, C.R., Fiedler, P., Kuhlmann, L., Liley, D., Vasconcelos, B., Fonseca, C., Tamburro, G., Comani, S., Lui, T.K., Tse, C.-Y., Warsito, I.F., Supriyanto, E., Haueisen, J., 2022. Multi-center evaluation of gel-based and dry multipin EEG caps. Sensors.

Olmi, B., Frassineti, L., Lanata, A., Manfredi, C., 2021. Automatic detection of epileptic seizures in neonatal intensive care units through EEG, ECG and video recordings: a survey. IEEE Access 9, 138174–138191.

Rakshasbhuvankar, A., Paul, S., Nagarajan, L., Ghosh, S., Rao, S., 2015. Amplitude-integrated EEG for detection of neonatal seizures: a systematic review. Seizure 33, 90–98.

Sandelowski, M., 2000. Whatever happened to qualitative description? Res. Nurs. Health 23, 334–340.

Sandelowski, M., 2010. What's in a name? Qualitative description revisited. Res. Nurs. Health 33, 77–84.

Song, J., Xu, F., Wang, L., Gao, L., Guo, J., Xia, L., Zhang, Y., Zhou, W., Wang, X., Zhu, C., 2015. Early amplitude-integrated electroencephalography predicts brain injury and neurological outcome in very preterm infants. Sci. Rep. 5, 13810.

Tao, J.D., Mathur, A.M., 2010. Using amplitude-integrated EEG in neonatal intensive care. J. Perinatol. 30, S73–S81.

Tong, A., Sainsbury, P., Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int. J. Qual. Health Care 19, 349–357.

van 't Westende, C., Geraedts, V.J., van Ramesdonk, T., Dudink, J., Schoonmade, L.J., van der Knaap, M.S., Stam, C.J., van de Pol, L.A., 2022. Neonatal quantitative electroencephalography and long-term outcomes: a systematic review. Dev. Med. Child Neurol. 64, 413–420.

Variane, G.F.T., Magalhães, M., Gasperine, R., Alves, H.C.B.R., Scoppetta, T.L.P.D., Figueredo, R.d.J.G., Rodrigues, F.P.M., Netto, A., Mimica, M.J., Gallacci, C.B., 2017. Early amplitude-integrated electroencephalography for monitoring neonates at high risk for brain injury. J. Pediatr. 93, 460–466.

Wang, X., Bik, A., de Groot, E.R., Tataranno, M.L., Benders, M.J.N.L., Dudink, J., 2023. Feasibility of automated early postnatal sleep staging in extremely and very preterm neonates using dual-channel EEG. Clin. Neurophysiol. 146, 55–64.

Wang, Z., Zhang, P., Zhou, W., Zhou, X., Shi, Y., Cheng, X., Lin, Z., Xia, S., Zhou, W., Cheng, G., 2021. Electroencephalography monitoring in the neonatal intensive care unit: a Chinese perspective. Transl. Pediatr. 10. No 3 (March 25, 2021): Translational Pediatrics.

Wikström, S., Pupp, I.H., Rosén, I., Norman, E., Fellman, V., Ley, D., Hellström-Westas, L., 2012. Early single-channel aEEG/EEG predicts outcome in very preterm infants. Acta Paediatr. 101, 719–726.