KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in the Field of Technology Electrical Engineering and the Main Field of Study Machine Learning

Second cycle, 30 credits

Cross-attention Modelling for Neonatal Adverse Events Detection from Multi-modal Time-series Data

Implementing a multi-head cross attention model for neonatal sepsis and necrotizing enterocolitis detection using several in-hospital timeseries data

SARAH REICHHUBER

Stockholm, Sweden, 2024

Cross-attention Modelling for Neonatal Adverse Events Detection from Multi-modal Time-series Data

Implementing a multi-head cross attention model for neonatal sepsis and necrotizing enterocolitis detection using several in-hospital timeseries data

SARAH REICHHUBER

Degree Programme in Electrical Engineering
Date: January 2, 2024

Supervisor: Antoine Honoré
Examiner: Saikat Chatterjee
School of Electrical Engineering and Computer Science
Swedish title: Korsuppmärksamhetsmodellering för detektion av allvarliga neonatala tillstånd från multimodal tidssekvensdata
Swedish subtitle: Implementering av en multihead cross attention-modell för detektion av neonatal sepsis och nekrotiserande enterokolit med hjälp av flera tidssekvensdata från sjukhusvistelsen

Abstract

Premature neonates are at risk for several life-threatening events during their first weeks of life, such as sepsis and necrotizing enterocolitis. These events are hard to detect in neonates, due to varying symptoms and unspecific clinical signs. This project aims to detect late-onset sepsis and necrotizing enterocolitis in patients up to 24 hours before clinical suspicion. It is well-established that these adverse conditions impact neonatal vital signs, which motivates the exploration of vital signs-based models for late-onset sepsis and necrotizing enterocolitis detection. A Multi-Head Cross-Attention model was implemented and trained using vital signs data from 30 neonatal intensive care unit patients, of whom 21 experienced any or both of the severe events, along with other relevant in-hospital longitudinal variables. The additional variables included respiration support, medication, mechanical ventilation pressure, weight measurements, fraction of inspired oxygen and laboratory results. Additionally, the postnatal age, birth weight, and sex of the patients were incorporated as input data for the model. When detecting whether and at what times patients experienced late-onset sepsis and/or necrotizing enterocolitis during their hospitalisation, the model with optimised parameters achieved an AUROC score of 0.66 after 100 epochs of training. Overall, Multi-Head Cross-Attention models show promise for early detection of neonatal adverse events, but their complexity necessitates further exploration and development. Additionally, a larger dataset is required to comprehensively evaluate the model's performance.

Keywords

Attention mechanism, Deep learning, Neonatal sepsis, Necrotizing enterocolitis

Sammanfattning

För tidigt födda nyfödda löper risk för flera livshotande händelser under sina första levnadsveckor, exempelvis sepsis och Nekrotiserande enterokolit. Dessa händelser är svåra att upptäcka hos nyfödda på grund av varierande symptom och ospecifika kliniska tecken. Syftet med detta projekt är att upptäcka sen neonatal sepsis och nekrotiserande enterokolit hos patienter upp till 24 timmar innan klinisk misstanke. Det är väl etablerat att dessa tillstånd påverkar neonatala vitala tecken, vilket motiverar att utforska modeller baserade på vitala tecken för att upptäcka sen neonatal sepsis och nekrotiserande enterokolit. En Multi-Head Cross-Attention-modell implementerades och tränades med data från vitala tecken från 30 patienter som vistats på neonatala intensivvårdsavdelningen, varav 21 upplevde en eller båda av de allvarliga händelserna, tillsammans med andra relevanta medicinska longitudinella variabler. De ytterligare variablerna inkluderade respiratorstöd, medicinering, tryck vid mekanisk ventilation, viktmätningar, syremättnad i blodet och laboratorieresultat. Dessutom inkluderades postnatal ålder, födelsevikt och patienternas kön som indata i modellen. När modellen förutsåg om och när patienterna upplevde sen neonatal sepsis och/eller nekrotiserande enterokolit under sin sjukhusvistelse uppnådde den med optimerade parametrar ett AUROC-värde på $0,66$ efter 100 träningsiterationer. Sammantaget verkar Multi-Head Cross-Attention-modeller lovande för tidig upptäckt av allvarliga neonatala händelser, men deras komplexitet gör att det krävs ytterligare utforskning och utveckling. Dessutom behövs en större datamängd för att fullständigt utvärdera modellens prestanda.

Nyckelord

Uppmärksamhetsmekanism, Djupinlärning, Neonatal sepsis, Nekrotiserande enterokolit

Acknowledgments

I would like to thank Antoine Honoré for his enduring support and guidance throughout this thesis - your hard work and dedication in helping others is truly inspiring. I would also like to thank Saikat Chatterjee and Eric Herlenius for contributing with inspiration and enthusiasm into the project. Lastly, I would like to thank my boyfriend, Johan Hallberg, for his constant encouragement and for always being willing to dive into conversations about my work.

Stockholm, January 2024
Sarah Reichhuber

Contents

List of Figures

List of Tables

Medical Terms

FiO_2	Fraction of inspired oxygen
O_2	Oxygen
P_aO_2	Arterial oxygen partial pressure
SpO_2	Peripheral oxygen saturation
anemia	Condition in which the body has a lack of healthy red blood cells
apnea	Temporary cessation (end of) of breathing, especially during sleep *(Opposite: tachypnea)*
bradycardia	Slow heart rate *(Opposite: tachycardia)*
CRASH	Cultures, Resuscitation, and Antibiotics Started Here - Moment when clinical symptoms of infection of LOS for neonates start, leading to that clinicians start antibiotic therapy and order blood cultures
fever	High body temperature *(Opposite: hypothermia)*
hypotension	Low blood pressure
hypothermia	A medical emergency that occurs when the body loses heat faster than it can produce heat, causing a dangerously low body temperature *(Opposite: fever)*
necrotizing	Undergoing necrosis, which is a form of cell injury resulting in the premature death of cells in living tissue by self-destruction
respiratory	Affecting breathing organs
tachycardia	The medical term for a high heart rate, over 100 beats a minute *(Opposite: bradycardia)*
tachypnea	Rapid and shallow breathing, over 60 breaths per minute *(Opposite: apnea)*

List of acronyms and abbreviations

AUROC	Area Under the Receiver Operating Characteristics
CMM	Center for Molecular Medicine
CNS	Culture Negative Sepsis
CNS	Central Nervous System
CPS	Culture Positive Sepsis
CRASH	Cultures, Resuscitation, and Antibiotics Started Here
DL	Deep Learning
ECG	Electrocardiography
EHR	Electronic Health Record
EOS	Early-Onset Sepsis
FNR	False Negative Rate
FPR	False Positive Rate
GBD	Global Burden of Disease
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HeRO	Heart Rate Observation
HRV	Heart Rate Variability
IBI	Inter-Beat-Interval
IV	Intravenous
IVH	Intraventricular Hemorrhage
KI	Karolinska Institutet
KTH	Kungliga Tekniska Högskolan (The Royal Institute of Technology)
LOS	Late-Onset Sepsis

MHCA	Multi-Head Cross Attention
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NEC	Necrotising Enterocolitis
NeoNEEDs	Neonatal Necrotizing Enterocolitis Early Detection Score
NICU	Neonatal Intensive Care Unit
PEEP	Positive End-Expiratory Pressure
PNA	Postnatal Age
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
RR	Respiratory Rate
SIRS	Systemic Inflammatory Response Syndrome
SOFA	Sequential Organ Failure Assessment
TNR	True Negative Rate
TPR	True Positive Rate
VLBW	Very Low Birth Weight
WHO	World Health Organization

Chapter 1

Introduction

This master thesis project is carried out in a collaboration between Kungliga Tekniska Högskolan (The Royal Institute of Technology) (KTH) and Karolinska Institutet (KI), with the intention to improve the outcome for neonates at the Neonatal Intensive Care Unit (NICU) by investigating potential new clinical decision support algorithms. As the project is part of the machine learning master programme at KTH, the focus will be on Machine Learning (ML)-based models, and attention-based models in particular. This chapter describes the specific problem that the thesis addresses, the context of the problem, the goals of this thesis project, and outlines the structure of the thesis.

1.1 Background

Modern hospitals in Europe, and in particular Karolinska University hospital in Stockholm, Sweden, create enormous amounts of data through patient journals, laboratory results and vital signs measurements. To improve patient care, patient treatments and diagnosis, it is crucial for the data to be used for research in a collaboration between medical doctors/researchers at Karolinska Hospital and information management experts. A collaboration between KI and KTH has yielded a database regrouping the medical information sources for neonates hospitalised in the neonatal intensive care unit at Karolinska University hospital. Neonatal Intensive Care Units, NICUs, are specialised hospital departments dedicated to providing intensive medical care to newborn babies. The aim of this project is to enhance the well-being and outcomes of patients receiving care in NICUs, by investigating novel ML -algorithms utilising a retrospective dataset of neonates previously hospitalised at the

Karolinska NICU.

1.2 Problem

Neonates, primarily premature, are at major risk for sepsis and other severe infections. The Global Burden of Disease (GBD) estimate that each year, there are globally around 1.3 million incident cases of neonatal sepsis and other infections, corresponding to approximately 937 cases per 100 000 live births [1]. Out of these cases, 203 000 are estimated to result in sepsis-attributable neonatal death. However, these numbers are not certain as the definition of sepsis and the way to collect medical data varies globally, and so the number of global sepsis cases could be up to four times higher than estimated by the GBD. Besides sepsis, neonates are at risk of Necrotising Enterocolitis (NEC). Studies suggest the incidence of NEC globally to be between 5% and 10% of preterm Very Low Birth Weight (VLBW) infants. While the mortality rate of NEC varies across ages of neonates, total rates between 15% and 30% have been reported [2][3].

1.2.1 Original problem and definition

Adverse conditions, such as neonatal sepsis and NEC, are often hard to identify in premature infants due to nonspecific clinical signs and ineffective test methods. The gold standard for sepsis for instance is a blood sample culture analysis, which often returns falsely negative results, due to the too small amount of blood that the medical staff can draw from neonates. This leads to severe morbidity and mortality amongst preterm neonates [4][5]. Diagnosing NEC is exceptionally challenging due to the lack of specific signs or accurate tests for diagnosis [6]. Furthermore, as antibiotics are an effective treatment for sepsis and NEC, the uncertain diagnosing of the events has the side-effect of overuse of antibiotics, contributing to antibiotics resistance in bacteria. In the long run, this undermines the effectiveness of antibiotic treatments in neonates [7]. Early, non-culture dependent, detection of sepsis and NEC could lead to improved neonatal outcomes in NICUs, reducing morbidity and mortality as well as medical costs [8][9].

1.2.2 Scientific and engineering issues

From a scientific point of view, the issue is how to detect and predict adverse events for neonates, using only readily available in-hospital longitudinal co-

variates. Since sepsis and NEC are difficult to diagnose for neonates, a well performing early detection model could aid medical staff to provide correct treatment. Moreover, the understanding of the specific data sources which was primarily utilised by the model, along with the time localisation of the moment of interest upon which its predictions are founded, would provide valuable insights into the fundamental processes governing the progression of diseases in infants. Several candidate ML-based models could achieve these desired properties. Attention-based models have seldom been investigated for neonatal time series data. This types of models have shown to be appropriate when dealing with irregularly sampled time-series, such as medical data [10]. In this project we propose to evaluate a Multi-Head Cross Attention (MHCA) model, where cross-attention models are used independently on each data source, and merged linearly to produce a disease prediction. The model is discussed in details in Chapter 2. The issue this thesis addresses is whether such MHCA models based on vital signs data together with other in-hospital longitudal co-variates, such as medication and respiration support, could help to detect the adverse events Late-Onset Sepsis (LOS) and NEC for neonates within 24 hours earlier than the diagnosis recorded in the patient's Electronic Health Record (EHR) journal.

1.3 Purpose

From the medical point of view, the results of this thesis could be an aid in clinical detection and more accurate treatment of adverse events for neonates. This could potentially lead to decreased morbidity and mortality in neonates, as well as more resource effective use of treatments such as antibiotics. The aim is also for both doctors and researchers to learn about causes and causation's of the adverse events by evaluating what parts of the data the model focuses on to produce a prediction. From a more technical stand point, the purpose of the thesis is to add knowledge on how cross-attention models can be used when medical data are multimodal and irregularly sampled time-series.

1.4 Goals

The goal of this project is to create a MHCA model, built on neonatal medical data, and analyse how well it detects severe events or might be of other use to research on the topic. This has been divided into the following three sub-goals:

1. Implement a functioning MHCA model.

2. Evaluate the model performance.

3. Use the attention maps of the model to interpret what the model focuses on when performing the adverse event detection.

1.5 Research Methodology

The methodology of this project will be based on research and papers where similar models have been built and implemented. With these models as inspiration, a MHCA model suited for the medical data available will be built. The model will be evaluated, adjusting hyper parameters and optimising computational speed and load, until an optimal model under these circumstances is found. The model versions will be evaluated in a retrospective cohort study, using a pre-defined set of evaluation measures.

1.6 Delimitations

The project is limited to include only data available from the NICU at Karolinska University Hospital, Stockholm. Furthermore, only the data of preterm neonates was used. A set number of modalities, *i.e.,* data sources, were used to train the model, see Chapter 3. The number of patients included in the final dataset were 30. It was also decided to only consider the severe events LOS and NEC when training the model, since these conditions have an acknowledged correlation with changes in the vital signs [4] [11]. Furthermore, this project was conducted during a limited amount of time, namely nine months, after which the best achieved results were reported.

1.7 Sustainability

There are several sustainable factors considered in this thesis project. Firstly, one project aim is to reduce the use of preventive antibiotics, both from an antibiotics resistance perspective and to improve outcomes of patients. Furthermore, the data used in the project is currently recorded and stored but not used. Therefore, we here aim to make use of the mentioned data by using it to train and design computationally and memory effective models.

1.8 Structure of the Thesis

Chapter 2 presents relevant background information about the modalities, the adverse events LOS and NEC and attention-based models. Chapter 3 introduces the methodology and method used to solve the problem of the project. Chapter 4 describes how the project was implemented and what design choices were made. In Chapter 5, results are presented, followed by a discussion of the results in Chapter 6. Lastly, conclusions and future works are described in Chapter 7.

Chapter 2

Background

This chapter begins by providing background information about the adverse events LOS and NEC for neonates. The different medical data modalities are then described. Then, this chapter describes a selection of previously implemented vital signs-based clinical decision support systems and continues by introducing attention-based models. The chapter also presents related works, consisting of one conducted by the same organisation as behind this project and a short introduction to previously implemented attention mechanisms, amongst them the transformer network, having a huge influence on the model created in this project.

2.1 Adverse Events for Neonates

Preterm neonates are neonates born before 37 completed weeks of pregnancy. The World Health Organization (WHO) estimated that in the year 2020, approximately 13.4 million babies were born preterm, corresponding to between 4% and 16% of all births across countries [12]. Karolinska University Hospital estimate that about 7000 babies, meaning roughly 6% of all born babies, in Sweden are born prematurely every year [13]. The possible causes behind preterm births are several, such as multiple pregnancies, high blood pressure and diabetes. They could also be due to medical reasons such as infections and pregnancy complications. There is possibly also a genetic influence [12].

Preterm neonates have immature immune systems, compared to term neonates, making them more susceptible to bacteria and infections [14] [15]. There are several adverse events that neonates can experience during the first hours and days of life, many of them life-threatening. In this project, neonatal

LOS and NEC are considered. An introduction to each of them is presented in the following subsections.

2.1.1 Sepsis

Sepsis is a life-threatening condition occurring when the body damages its own organs and tissue as a response to an infection. The definition of sepsis has varied over time. For adults, the International Sepsis Definitions conference specified respiratory rate, heart rate, temperature, and blood pressure criteria for sepsis, severe sepsis, septic shock and Systemic Inflammatory Response Syndrome (SIRS) in 2001. In 2002, similar criteria for children and term neonates were defined. However, these have proven to not hold for preterm neonates [4]. In the third international consensus in year 2016 [16], the criteria for sepsis were redefined, as the previous ones were not specific or precise enough. Sepsis was defined to be a ”life-threatening organ dysfunction caused by a dysregulated host response to infection”. The committee agreed that ”organ dysfunction” should be represented by an increase in the invented Sequential Organ Failure Assessment (SOFA) score by at least 2 points, where a higher SOFA score indicates a higher probability of mortality. This threshold score corresponds to a mortality risk of approximately 10%. This definition of sepsis also holds for adults, and not specifically neonates.

Neonatal sepsis can be divided into Early-Onset Sepsis (EOS) and Late-Onset Sepsis (LOS), where the term EOS is used to describe sepsis when experienced within the first 72 hours after birth and LOS is used for sepsis experienced more than 72 hours after birth [8]. Both are frequently detected for preterm infants, but LOS is even more common than EOS because of long hospital stays and immature immune systems of preterms. In this project, only LOS is examined.

Premature neonates have an increased risk for overwhelming sepsis, shock, and death compared to mature neonates, due to their low gestational age and their immature immune systems. For preterm neonates in the NICUs, neonatal sepsis can reach a mortality rate of up to 11.3%, making it one of the main causes of death in the group. Generally, sepsis severity is correlated with age, sex and genetic factors.

The neonatal response to sepsis includes changes in body temperature, respiratory drive and oxygenation, heart rate characteristics, and blood pressure. The severity of the response is usually lower in the earlier stages of sepsis, and higher in the advanced stages. For adults and children, there exists specific thresholds of the vital signs based conditions tachycardia,

hypotension, and fever or hypothermia for sepsis. These thresholds do, however, not hold for neonates, especially premature, since changes in vital signs related to sepsis may present as either hypothermia or fever, apnea or tachypnea, and bradycardia or tachycardia due to their immature immune system. Furthermore, these symptoms can occur due to other causes than sepsis [16][17].

Currently, neonatal sepsis is detected through positive blood cultures, which are ordered by doctors when they discover clinical signs of illness. Blood cultures are, however, both time-consuming and sensitive to contamination [8]. The results of blood cultures normally take between 24-48 hours, and even if negative, sepsis cannot be ruled out [18]. As it is not feasible to take the optimal 6 ml of blood from neonates, the test sensitivity is not high. Both Culture Negative Sepsis (CNS) and Culture Positive Sepsis (CPS) are examined during this project, since the aim is to detect sepsis regardless of the results of the blood cultures. The lack of reliable diagnostics for LOS lead to overuse of antibiotics treatment for neonates. Prolonged antibiotic exposure in preterm infants does, however, have severe side effects, such as increased antibiotics resistance, infections and potential NEC and death [5]. Other effects of sepsis include the increased risk of neurodevelopmental impairment (for extremely low birthweight neonates) as well as enlonged hospitalisations, resulting in increased medical costs [9].

2.1.2 Necrotising enterocolitis

NEC causes severe, necrotizing damage to the intestine of neonates. It is one of the main complications of preterm infants, experienced by 5-10% of VLBW neonates and with a mortality of 15-30% [2]. Some of the risk factors for neonatal NEC that have been found include low birth weight, being small for ones gestational age and premature birth [19]. Furthermore, studies suggest that breast feeding could decrease the risk of NEC. The symptoms include abdominal swelling, vomiting, tiredness, apnea and bradycardia. Small gas bubbles on the intestinal wall of the infant can be found through X-ray or ultrasound. These might lead to the intestine of the baby bursting [20]. The more severe signs of NEC, such as abdominal distention and bloody stool, often show accompanied by sepsis [5]. The pathogenesis of NEC is not completely defined [21], and due to its complexity and multifactoriality, it is often hard to predict in individuals [2] [19] [5]. The classic clinical approach to diagnosing NEC is based on finding general signs of inflammation in the body, stomach-related symptoms, and using X-rays to confirm if there is

inflammation in the digestive system [6]. The complexity of the condition might lead to overtreatment, *i.e.,* treating patients that do not experience NEC. As treatment, the infant gets only Intravenous (IV) nutrition along with antibiotics and in severe cases, surgery can be necessary [20]. Furthermore, blood tests might be necessary. However, these methods can have negative consequences on the infant such as increased risk of anemia, growth hinders and possibly event increased risk of NEC development [6].

2.2 Medical Data

Seven modalities of longitudinal, medical data, proven or believed to be correlated with sepsis and/or NEC for neonates, or affecting their vital signs, were used in this project. The seven modalities are:

1. Vital signs
2. Respiration support
3. Medication
4. Mechanical ventilation pressure
5. Weight measurements
6. Fraction of inspired oxygen
7. Laboratory results

LOS and NEC are known to affect the vital signs of preterm infants, which are mostly recorded regularly and at high sampling frequency. Thus vital signs can be seen as the main modality for the model to detect the severe events. The other modalities have either also shown a correlation to the severe events, or have an impact on the vital signs. We now endeavour to describe each of the modalities with the relevant details to motivate their use in the dataset to train the severe events detection model. Besides the modalities, the impact of the demographic variables age, sex and birth weight on the severe events are described.

2.2.1 Vital signs

When experiencing sepsis or NEC, changes in the vital signs occur for neonates [4] [11], as presented in Chapter 2.1. Consequently, analysing changes in vital signs might yield insight in the risk and process of the infant sickening. Previous studies have shown an increased model performance when including the three vital signs heart rate, oxygen saturation and respiration rate compared to only including heart rate characteristics [22]. Thus, the vital signs used in this project are the heart rate, oxygen saturation and respiration rate, each shortly presented below.

Heart Rate

Changes in heart rate, or Inter-Beat-Intervals (IBIs), might occur for neonates due to sepsis or NEC, but could be caused by other conditions, such as lung inflammations, brain injuries or in correlation with surgeries, or simply because of prematurity [4].

Oxygen Saturation

The oxygen saturation of the patients is estimated from pulse oxymetry and is denoted SpO_2. This measurement describes the level of oxygen available to the organs and is a proxy for respiratory function of a patient [23].

For preterm neonates, one of the most common signs of sepsis is an increase in the frequency and severity of central apnea. Furthermore, a study on the correlation between vital signs and LOS and NEC found that the best predictor for both severe events was the cross-correlation of heart rate and SpO_2 [11].

Respiratory Rate

The Respiratory Rate (RR) of newborns is normally between 30 and 60 breaths per minute [24]. Neonates with sepsis may experience tachypnea, defined as a respiratory rate over 60 breaths per minute, or respiratory distress. For preterm neonates, an increase in central apnea is one of the most prominent signs of sepsis. Additionally, it is common for preterms to have apnea of prematurity as a consequence of undeveloped control of their breathing [4].

2.2.2 Respiration support

When neonates experience breathing difficulties or are prematurely born, they could be in need of respiration support for a period of time. The respiration support is given through a plastic tube in the patient's nose or mouth, connected to a respirator. Doctors can adjust the respirator to control the number of breaths, pressure and oxygen level the infant receives [25]. Thus, if the neonate receives respiratory support, this will impact the vital signs, mainly the respiration rate.

2.2.3 Medication

Neonates, primarily premature, may receive a number of drugs due to various complications during their first days of life. These medications might have an impact on the vital signs of the neonates, and can thereby affect the use of vital signs as physiomarkers for sepsis or NEC. For instance, caffeine can lead to increased heart rate, anticholinergic medications can cause tachycardia and glucocorticoids can lead to increased heart rate variability, fever or hypotension [4].

2.2.4 Mechanical ventilation pressure

Mechanical ventilation is used for severely ill patients as a support-system to maintain sufficient lung functioning [26]. The mechanical ventilation pressure can be measured as Positive End-Expiratory Pressure (PEEP) and the pressure peak. PEEP is the positive pressure, greater than the atmospheric pressure, remaining in the airways after exhalation of a patient that is mechanically ventilated. Mechanical ventilation is a known confounder for the vital signs data, for instance, an artificially high airway pressure changes the SpO_2 drastically. Taking into account this information is thus of primary importance to interpret the vital signs modality correctly.

2.2.5 Weight measurements

Repetitive weight measurements of neonates are typically performed daily in NICUs. As low birth weight has a strong correlation with both types of severe events in this project, it is of interest to keep track of the weight of the infant during the hospitalisation. It has also been shown that a risk factor for NEC is being small for one's gestational age [19]. Overall, weight measurements are used as proxies to estimate the developmental stages of premature infants.

2.2.6 Fraction of inspired oxygen

The fraction of inspired oxygen, FiO_2, is an estimation of the oxygen content a patient inhales. FiO_2 adjustments are made to maintain the target SpO_2 range of the neonate through oxygen blenders [27]. This affects the vital signs similarly to, and is often used in combination with, mechanical ventilation.

2.2.7 Laboratory results

Primarily, laboratory results refer to blood gas analysis. The measurements might include analytes such as pH value (indicating if a person is acidemic or alkalemic), P_aO_2 (arterial oxygen partial pressure, which if low shows an abnormal oxygenation of blood) and O_2 content (the sum of oxygen of a patient dissolved in plasma and chemically bound to hemoglobin). These are routinely measured in NICUs and often provide crucial information to the medical staff.

2.2.8 Demographic data

Besides the longitudinal sampling for the modalities presented above, the dataset also includes demographic information about the patients, consisting of gestational age (in weeks), sex and birth weight. Thus, as there are three types of demographic data, $d_{demo} = 3$.

Gestational Age

The gestational age of neonates has a critical impact on the risk of sepsis and death, with low gestational age leading to higher risk [4]. Premature infants run the highest risk of LOS [9]. For NEC, premature birth has also been identified as a risk factor [19]. A graph of the mortality rate due to NEC versus different gestational ages for neonates can be seen in Clark *et al.,* [3].

Sex

While sex is reported to have an impact on the mortality of sepsis in adult patients [28], men running a higher risk of mortal outcome, the impact of sex in infant patients is not as prominent [4]. For NEC, one study found a higher mortality risk for male patients [3]. Generally, male patients appear to be at higher risk for experiencing LOS and NEC than female patients, but also to be at higher risk for severe morbidity and mortality in general when born before 30 weeks of gestation [29].

Birth Weight

Low birth weight is a known risk factor for both LOS and NEC, VLBW patients being at higher risk [9] [19] [3].

2.3 Vital Signs-Based Clinical Decision Support Systems

Several clinical decision support systems, providing risk predictions of mainly LOS and sometimes NEC, have been designed and tested in practice. An example of a commercially accessible early warning system for sepsis detection in neonates, which relies on vital signs data, is the Heart Rate Observation (HeRO) monitor. The HeRO score is based on abnormal heart rate characteristics, since sepsis is known to be associated with decreased Heart Rate Variability (HRV) and transient heart rate decelerations. The score is a comparison of the predicted risk of sepsis for the infant in the next 24 hours, compared to the average risk of sepsis for all VLBW infants at any time. When this score was displayed to clinicians in a study, the total mortality showed to decrease by 22% and the mortality due to sepsis by 40% [4] [30]. However, this system still suffers from a high false alarm rate and is often not used in practice.

An early clinical signs and symptoms warning score for NEC is the Neonatal Necrotizing Enterocolitis Early Detection Score (NeoNEEDs), where five clinical categories were scored: behaviour, cardiac, respiratory, gastrointestinal, and feeding tolerance. A prospective clinical trial study showed that the use of the warning score decreased NEC severity rates [31]. The study also showed that the most frequently encountered early presenting sign of NEC was cardiorespiratory instability.

The first predictive algorithm utilising multiple vital signs monitoring for detecting events for neonates such as LOS and NEC was RALIS. The vital signs data used included heart rate, respiratory rate, temperature, desaturations and bradycardias. If the RALIS generated score becomes higher than five, an alert is triggered [5].

Several ML-based methods using vital signs data also exist today. For instance, one study used Naïve Bayes algorithm for sepsis detection using the vital signs HRV, SpO_2 and RR, yielding an Area Under the Receiver Operating Characteristics (AUROC) score of 0.82 [22]. One project created a model called DeepLOS for LOS prediction in preterm neonates, based on residual convolutional networks and feature map attention, using IBIs data [8]. This

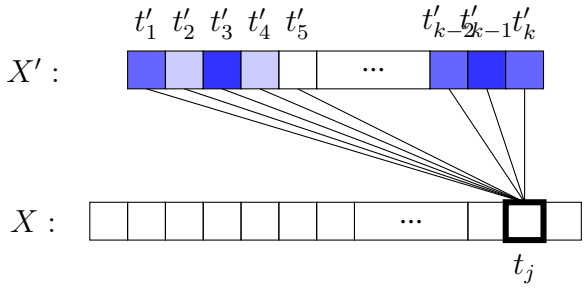


Figure 2.1: Example of attention weights. The darker the colour of each X' (key) time stamp, the more relevant the value at that time stamp is to the value at the highlighted X (query) time stamp t_j.

model achieved an AUROC score of 0.85 when performing predictions 6 hours before CRASH.

2.4 Multi-Head Cross-Attention

Medical health data often consists of irregularly sampled time series, such as medication data or laboratory results. The data may have missing values, e.g. not all laboratory values sampled at the same time, and be sampled at modality specific irregular or regular time. Since machine learning models usually assume the data to be fully observable and with a fixed size, the irregularity in the sampling of medical data can cause problems when modelling [10] [32]. Another challenge with time series data is that the dependency between time samples is usually assumed to be in proportion to the time distance between them, although this might not always be the case. Attention mechanisms are a way to model dependencies in time not regarding to the distance between input and output space, but dependent on their similarity. The intuition behind attention is illustrated in Figure 2.1, where one time series X' is matched to another time series X. The time series are called the key and the query, which will be described further in Section 2.4.1. One network architecture based on attention mechanism is the Transformer, introduced by Vaswani *et al.,* [33].

2.4.1 The scaled dot-product attention

A standard attention layer has two input sequences, $X \in \mathbb{R}^{T \times D}$ and $X' \in \mathbb{R}^{T' \times D'}$, for instance two time series, of lengths T and T' and dimensions D and D', respectively. If $X = X'$, the mechanism is referred to as self-attention,

otherwise it is referred to as cross-attention. These sequences are used to build the query matrix $Q \in \mathbb{R}^{T \times D}$, the key matrix $K \in \mathbb{R}^{T' \times D}$ and the value matrix $V \in \mathbb{R}^{T' \times D'}$, such that

$$Q = f_Q(X), \tag{2.1}$$

$$K = f_K(X'), \tag{2.2}$$

and

$$V = f_V(X'), \tag{2.3}$$

where f_Q, f_K and f_V are trainable, linear row-wise transforms. These quantities function as the inputs to the model. The attention matrix $A \in \mathbb{R}^{T \times T'}$ is yielded by matching the key sequences, the rows of the K matrix, to the query sequences, the rows of the Q matrix The matching is done by a row-wise dot-product, between all query and key sequences as:

$$A = \text{softmax}_{row}\left(\frac{QK^T}{\sqrt{D}}\right). \tag{2.4}$$

The softmax function is applied to create weights between 0 and 1. The scaling factor $\sqrt{D}$ is added to counteract large dot-products for large values of D, which could potentially lead to the softmax function yielding extremely small gradients [33] [34]. The value matrix V is then weighted with the attention matrix, resulting in the output sequence $Z \in \mathbb{R}^{T \times D'}$ [34]. Mathematically, this is described as

$$Z = AV. \tag{2.5}$$

The intuition behind attention can be described as "an averaging of values associated to keys matching a query".

2.4.2 Causal attention

To ensure causality, *i.e.,* that no information of future time stamps is used when estimating the output at each respective time stamp, a mask matrix $B \in \{0, 1\}^{T \times T'}$ with the same dimensions as the attention matrix A, is typically used and added to the model as follows:

$$\forall (j,k) \in [T] \times [T'],\ B(j,k) = \begin{cases} 1 & \text{if } t_j \geq t'_k \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

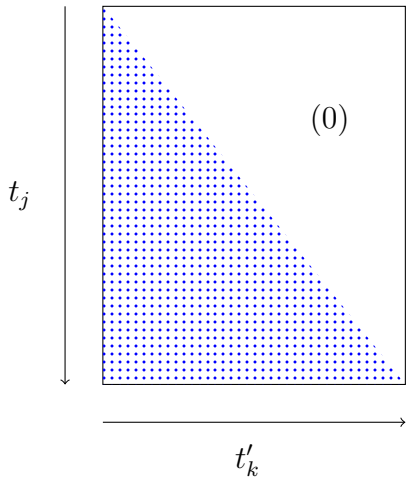


Figure 2.2: The mask matrix B. Inspired by teaching slides by Fleuret [35].

The mask matrix is illustrated in Figure 2.2, where it is seen that the matrix entries are set to zero whenever $t_j \geq t'_k$, *i.e.,* when the keys are in the future of the query time point [34]. The outputs Z are then yielded as

$$Z = (A \odot B)\, V \tag{2.7}$$

where $\odot$ is the element-wise multiplication of A and B, known as the Hadamard product.

2.4.3 Multi-head attention

The above mentioned method to yield the output Z from the attention function can be done in parallel for different combinations of modalities X:s and X':s, or "heads", leading to multi-head attention. For M different partial attention outputs, $Z_1 \in \mathbb{R}^{T \times D'_1}$, ..., $Z_M \in \mathbb{R}^{T \times D'_M}$, computed as described above, the final output can then be the result of a trainable linear operation with weight matrix $W \in \mathbb{R}^{H \times C}$, where $H = M \times D'$ and C is the dimension of the output sequence, as

$$\hat{Y} = [Z_1, ..., Z_M]\, W \in \mathbb{R}^{T \times C}. \tag{2.8}$$

As seen, the models prediction of the output will have dimension T, as the length of the sequence X used to obtain the query. The dimension C represents the output dimension, i.e. the dimension of the sequence to be predicted. This is further explained in Chapter 3.

2.5 Related Work

In this section, we first introduce a project made by the same research collaboration parties as this project. Following, a short introduction to previous studies using attention mechanism is presented.

2.5.1 Vital-signs based detection of sepsis using machine learning

The problem of neonatal sepsis detection has been addressed before in the KI/KTH research group, for instance by Honoré *et al.,* [22]. The authors have presented a machine learning based algorithm, Naïve Bayes, trained on the same vital signs and demographic data as used in this project. One difference is that the entire NICU population from the KI data was included and here our population is more restricted. The authors concluded that compared to their previous study, where the heart rate characteristics was the only vital sign used for sepsis detection, adding respiratory- related vital signs improved the performance. They also inferred that prospective studies are needed to further fine-tune machine learning algorithms for the same purposes.

2.5.2 Attention mechanisms

Models relying on attention mechanisms, like the Transformer, were initially introduced as sequence-to-sequence models with the primary purpose of machine translation. Transformer-based pre-trained models perform very well in natural language processing, and have thus turned to be the go-to models in the field [34]. Today, they have also shown great potential in numerous other Deep Learning (DL) fields, such as computer vision and audio processing. With the growing popularity of attention mechanisms, they have also been applied in further disciplines, amongst them chemistry and medicine. Some previous studies utilising attention models on patient data include Shukla and Marlin (2021) [10], Chauhan *et al.,* (2022) [32] and Ge *et al.,* (2022) [36].

Chapter 3

Method

The purpose of this chapter is to provide an overview of the research method used in this thesis. Section 3.1 describes the research process. Section 3.2 focuses on the data collection techniques used for this research. Section 3.3 explains the limits that have been decided on what data to use in the project. In Section 3.4, the approach for labelling the data is presented. Section 3.5 explains how the data was preprocessed before training, and is followed by a summary of the project data in Section 3.6. Section 3.7 introduces the data representation. The model implementation and construction is described in Section 3.8, and then the experimental design can be found in Section 3.9. Finally, the evaluation framework of the model is introduced in Section 3.10.

3.1 Research Process

The research process is built on the following steps:

Step 1 Do background research and explore data,

Step 2 Plan structure of the MHCA model,

Step 3 Evaluate the constructed model

Step 4 Generate results from the model

First, background research about attention mechanisms, vital signs-based prediction models and neonatal adverse events is made in parallel with data exploration. Next, the model is planned and built. Based on the model evaluation, the model is adjusted to improve performance. Lastly, after an unspecific number of model upgrade iterations, the best found model is used to generate results, to be presented in Chapter 5. The process is illustrated in Figure 3.1.

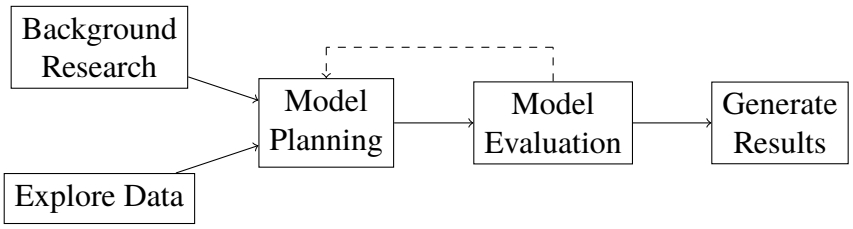


Figure 3.1: The research process: Background research and data exploration are made in parallel. Then the model is built and evaluated and then, based on the evaluation, re-built. Finally, the model results are generated.

3.2 Data Collection

The data used in this project is limited to data collected from the NICU at Karolinska University Hospital in Solna and Huddinge, Stockholm, Sweden. The dataset is collected and made available for research using a computer architecture depicted on Figure 3.2. The Karolinska university hospital stores patient journals in the Takecare system which is later annotated by medical doctors. Laboratory results and data from mechanical respiratory support, handled in the Clinisoft system, and data from vital signs monitoring are produced by Philips. The resulting database is then used for research in collaboration between KI and KTH. Furthermore, the Center for Molecular Medicine (CMM) provides computational servers to be used for research purposes .

The neonates in the dataset were hospitalised between July 2017 and May 2020. They were born in a variety of different hospitals in Sweden, including Danderyd, Solna, Linköping and Södersjukhuset. It was decided to only include data from premature neonates, *i.e.*, infants born before 37 weeks of gestation, in the project since that group of neonates are at a much higher risk of experiencing the adverse events considered, as described in the Background section 2. The resulting dataset included 30 patients, of whom 21 experienced either LOS or NEC during their hospitalisation. For each patient, the data consists of the seven modalities presented in the Background, Section 2.2.

3.2.1 Vital signs data collection

The vital signs used in this project are heart rate, oxygen saturation and respiration rate. The heart rate data was received as IBIs derived from electrocardiography, the oxygen saturation was measured from pulse oximetry and the respiration rate was derived from chest impedance. The vital signs

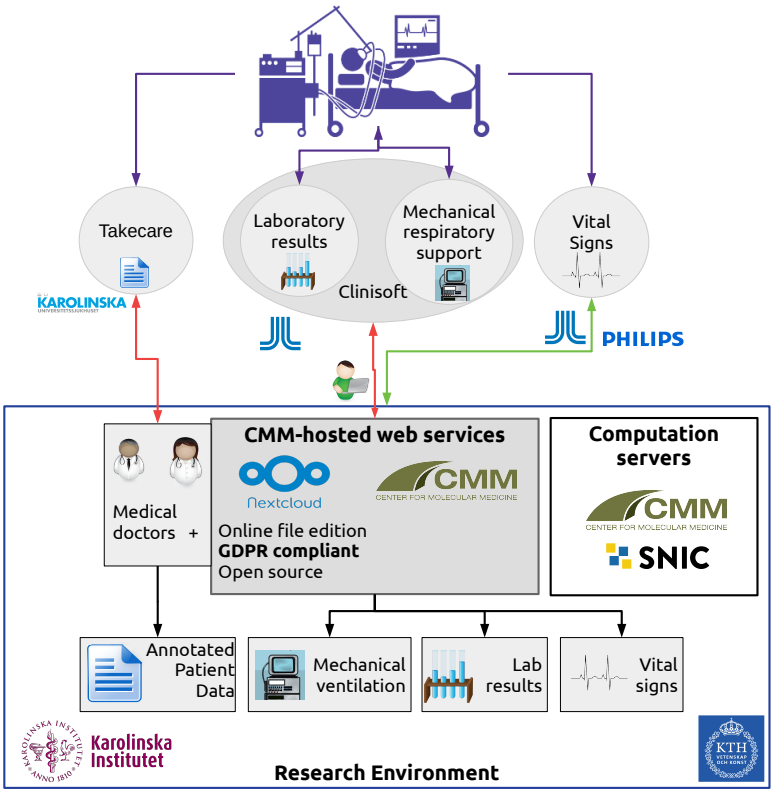


Figure 3.2: Computer architecture of research environment between Karolinska university hospital, KI and KTH.

data was obtained from standard patient monitors used in the daily clinical care at the Karolinska University Hospital (Philips IntelliVue MX800 Patient Monitor, Philips Healthcare, Amsterdam, the Netherlands) and stored on the CMM storage servers. Each of the three vital signs are derived from high-frequency waveform data, and resampled at 1 Hz (one measurement per second). Although the vital signs data for each patient consists of three sequences of measurements for the three vital signs, with potential gaps.

3.2.2 Other modalities data collection

The data from the other modalities, *i.e.,* laboratory results, medications, pressure, weight measurements, FiO_2 levels and respiratory support, were denoted by trained medical staff relying on clinical notes or documentation provided by doctors in the NICU at the time of the event. As an example, when a physician administers medication to a patient, the administration time

is recorded automatically together with the dosage amount. These recorded time stamps and notes subsequently served as the foundation for the patient database, created by the medical personnel.

3.2.3 Ethical aspect

The main ethical issue concerned during this project is the handling of personal data. The patient data is stored according to General Data Protection Regulation (GDPR). To ensure protection of the sensitive data, the personal ID:s of the patients was processed with a salted-hash algorithm. There are no use of personal ID:s during this project, and so the medical data remains pseudo-anonymous.

From a sustainability perspective, the project aims at reducing the use of preventive antibiotics, to improve outcome in hospitalised patients. In the project development, we endeavour to make use of data that is currently recorded but unused at the patient's bedside. We use a compute infrastructure located in Sweden, and endeavour to design models with a small architecture to not waste memory or computational power in the final model, but to make it effective.

3.3 Data Limiting

The original dataset consists of the data from 222 patients hospitalised at the NICU. The occurrence of several other severe events, besides sepsis and NEC, are also documented, such as pneumonia, lung bleeding, Central Nervous System (CNS) infections and Intraventricular Hemorrhage (IVH). These events were initially considered to be included for detection by the model developed during this project, but were later excluded due to (1) limited computational capacity, thus limiting the number of patients we could use, and (2) the relationship between the clinical condition and the vital signs were not as documented as for sepsis and NEC. Since only 21 out of the documented 222 NICU patients experienced LOS and/or NEC during their hospitalisation, it was decided to only include 30 patients in the final dataset. Nine patients which did not suffer from any severe events were randomly selected and were included along with the 21 positive patients in the final dataset. This ensured that the symptoms from other events were not confused with signs of sepsis or NEC. The limitation of the number of patients was done mainly due to the limited computational power and the time limit of the project. The priority was set on including the ill neonates in the dataset.

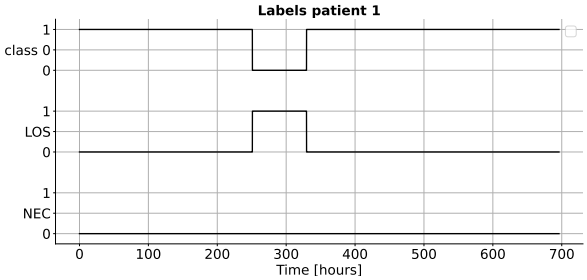
To further limit the computational complexity, it was also decided to use a maximum of 29 days of data for each patient. The 29 days correspond to 14 days of data before, and 14 days after, the first severe event is encountered in the patient medical data, where the positive labels are defined to last for one day, as described in Section 3.4. If the first severe event of the patient started less than two weeks after the start of the vital signs recording, or later than two weeks before the vital signs end, the 29 days included more data from before or after the event to make sure the total amount of vital signs data used would still be 29 days. For patients with less than 29 days of vital signs data available, all of the available data was used. For negative patients, the first 29 days of vital signs data were used.

The main part of the patient data are the vital signs, which were measured with a frequency of 1 Hz, corresponding to one sample per second, for each patient for different intervals of time during their hospitalisation with possible gaps in time. However, it was decided to down-sample the frequency by a factor of 10 due to computational limits. The final data was thereby sampled six times per minute, every 10:th second.

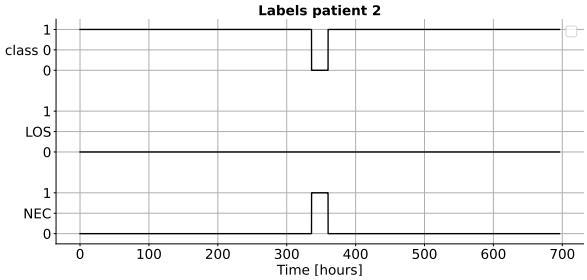
3.4 Data Labelling

The model was designed to identify the critical neonatal events LOS and NEC. The time of onset of these events was defined as the time the first blood culture was drawn, as an effect of clinical suspicion by a NICU doctor. If another LOS/NEC episode was detected within 14 days, it was marked as the same LOS/NEC episode when labelling the data. The data labels for each patient are multi-dimensional, with one dimension for each of the two types of adverse events and with the same length as the vital signs timeline for the patient. A third class was added to the labels, to also be to detected by the model, called *class0*. This class corresponds to none of the severe events being present. Thus, the output dimension of the model will be $C = 3$.

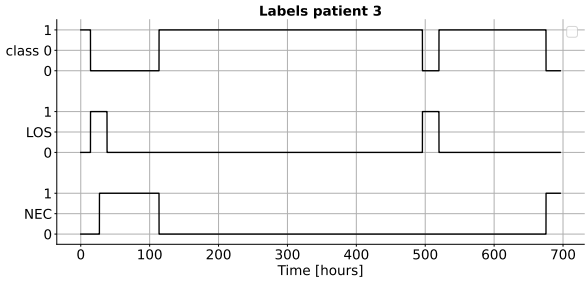
For the severe events classes, the label is set to be 0 (negative) for timestamps when the adverse event is absent and 1 (positive) for the 24 hours before the first suspicion of illness by the doctors, or a positive test, was registered in the patient journal. This allows for multi-labelling, *i.e.*, a patient can be positive for both severe adverse events at the same time. *class0* is equal to 1 (positive) at timestamps where both LOS and NEC are equal to 0 (negative), and otherwise it is 0. In Figure 3.3, three examples of labels are presented: one of a patient experiencing LOS, one of a patient experiencing NEC and one of a patient experiencing two LOS and two NEC episodes.



(a) Example of labels of a patient experiencing LOS



(b) Example of labels of a patient experiencing NEC



(c) Example of labels of a patient experiencing LOS and NEC

Figure 3.3: Three examples of patient labels

The number of patients experiencing each of the events, along with the total number of positively labelled time samples for each of the classes in the entire dataset, can be seen in Table 3.1. The total number of patients experiencing any severe event, referred to as the positive patients, is 21. Out of those 21 patients, nine experienced only LOS, eight experienced only NEC and four experienced both conditions within the maximal 29 days of vital signs data during their hospitalisation. The number of negative patients, *i.e.*, the patients never experiencing any of the severe events, was nine. All patients had the label *class0* at some time point during their hospitalisation.

Table 3.1: Number of patients experiencing each of the adverse events. Down-sampled by 10.

Class	Number of patients	# positive time stamps
LOS	13	150332
NEC	12	239051
class0	30	6394492

3.5 Data Preprocessing

The data preprocessing consisted of several steps. Firstly, the format of the data was adjusted, as described in the following subsection. Then, the number of patients and the length of the time series were reduced. Two new co-variates were added to replace the timeline in days and time format. The construction of the target data is described after that, and following that some statistics about missing values and the complete dataset is presented. Lastly, the ethical aspects are discussed.

3.5.1 Data restructuring

Primarily, the structure of the data was changed to create a timeline for each modality and patient. For instance, the medication data of each patient was constructed as a timeline with the quantity of each medication the patient received at each time. The medications data was represented as a N_m-dimensional vector. For a given patient-time, the medication that were not injected were assigned NaN (”not a number”) value. An example of the structure of the medication data for a patient is seen in Figure 3.4, where each row represents a timestamp of medication and each column a type of medicine. The entries which are not NaN, show the quantity of the medicines

Timeline	Med1	···	MedN_m
2018-04-03 9:44:00	1.7	···	NaN
2018-04-03 9:53:00	NaN	···	0.5
⋮	⋮	⋮	⋮

Figure 3.4: Example of the structure of the medical data, where at the first time stamp the patient received 1.7 units of medicine 1 and at the second time stamp, they receive 0.5 units of medicine N_m.

Timeline	fio2
2018-04-03 12:08:00	31.0
2018-04-03 12:14:00	29.0
⋮	⋮

Figure 3.5: Example of the structure of the FiO_2 data, as a timeline.

the patient was given at that time stamp. Note that all medicines might not be given to the patient at any time point, but they could also be given several types of medicine at the same time. A similar structure was used for the laboratory results. The weight measurement, pressure and FiO_2 data were each represented by a timeline with their respective values of the patient for each timestamp measurements were made, see example in Figure 3.5.

The respiratory data was processed as a timeline with a binary value for each type of respirator - for each timestamp, the column of the respirator is 1 if the respirator was active at this timestamp and NaN if it was not active. Several types of respiratory support can be given at the same time, but it is usually not the case. An example of the respirator data structure is seen in Figure 3.6.

All of the events for the above mentioned modalities have timestamps documented with minute precision. The time stamps are irregularly spaced, *i.e.,* time intervals between observations are irregular and dependent upon clinical decisions that are difficult to model.

A summary of the number of time points of vital signs data available for the set of neonates is presented in Table 3.2. There is a spread in the length of the vital signs data available across patients. As seen, one patient has 44366

Timeline	Resp1	···	RespN_r
2018-04-03 11:02:00	1	···	NaN
2018-04-03 11:45:00	1	···	NaN
⋮	⋮	⋮	⋮

Figure 3.6: Example of the structure of the respirator data, where the respirator column si 1 if that respirator is active at the given time stamp, and NaN otherwise.

samples of vital signs data available (corresponding to around five days of data) while as the patients with the longest sequences of vital signs data 250560 samples (29 days) of data available. The length of the data is depending on the duration and severity of the patients hospitalisation. The average number of vital signs measurements available for the patients is 26.2 days.

Table 3.2: Vital signs lengths for distinct patients - minimum (min), mean, maximum (max) and standard deviation (std)

	min	mean	max	std
# vital signs measurements	44366	225985.1	250560	54567.5

3.5.2 Timeline adjustment

We chose to deal with the irregularly spaced time series by augmenting the vector representing the data with two additional dimensions. The first dimension corresponds to a time-delta column indicating the time difference (in hours) to the previous measurement. The second dimension correspond to the Postnatal Age (PNA) of the neonate.

3.5.3 Normalisation

Given the variation in data units and ranges across different modalities, normalisation was applied to standardise the data for each co-variate in each modality. We ensured that each co-variate of each modality had zero mean and unit variance across the training dataset. The validation set was normalised using the scaling and translation coefficients calculated on the training set (the splitting into training and validation set will be more discussed in Section 4.1).

3.5.4 Missing values

As the model is supposed to build on the vital signs data primarily, an important aspect of the model is the extent to which the vital signs data is available, *i.e.*, not missing, when the severe events occur. For the positive timestamps of each of the severe events, the percentage of those that data of each of the vital signs is available is summarised in Table 3.3. As seen, no vital signs data is completely available. Mainly, the respiratory rate during NEC events are missing, and just 38.7% of the positive time stamps for NEC has respiratory rate data available.

Table 3.3: Percentage of timestamps where the vital signs heart rate (IBI), oxygen saturation (SpO_2) and respiratory rate (RR) data are available for positive LOS and NEC timestamps, respectively.

Event	**IBI**	SpO_2	**RR**
LOS	75.8%	92.5%	75.5%
NEC	69.8%	90.9%	38.7%

Some examples of patient vital signs data together with the patient labels are shown in Figures 3.7, 3.8 and 3.9. As seen, the time intervals of missing vital signs data is varying amongst the vital signs and across patients.

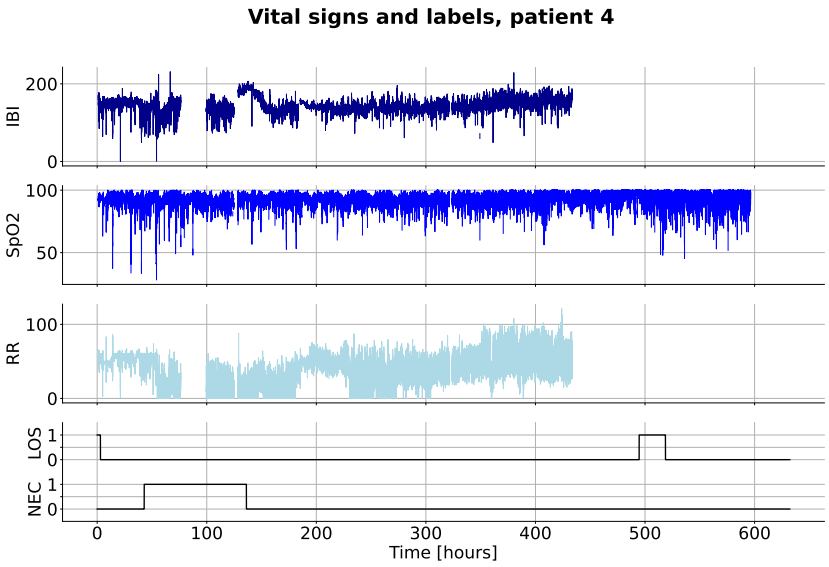


Figure 3.7: Vital signs data and labels for patient 4

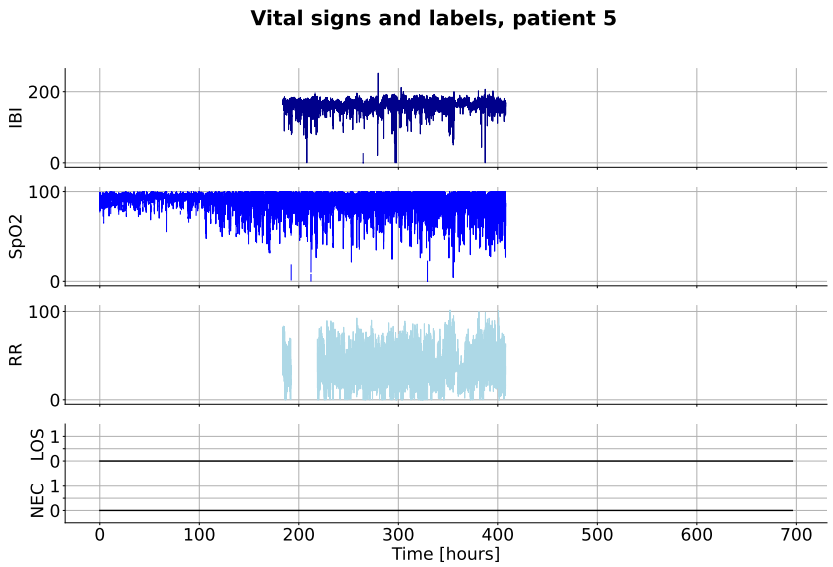


Figure 3.8: Vital signs data and labels for patient 5

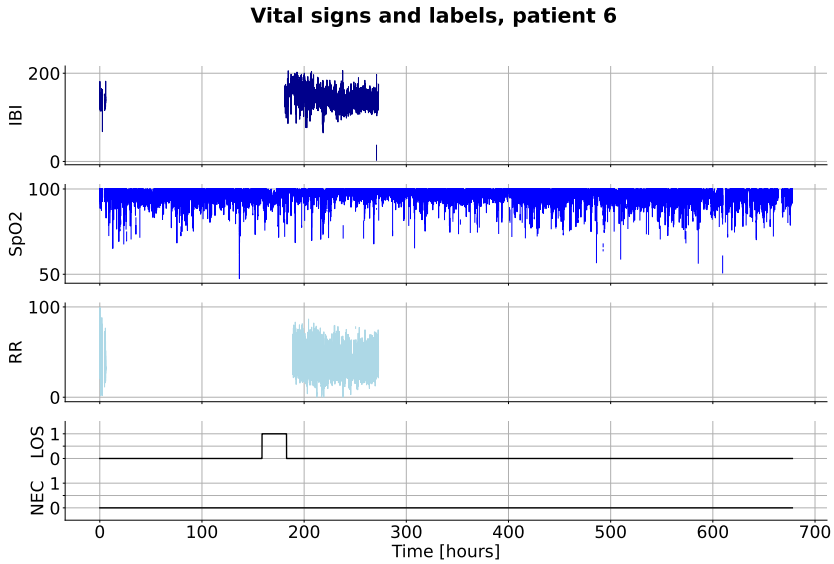


Figure 3.9: Vital signs data and labels for patient 6

We chose to let the model to learn to use the information of missing data. When vital signs data was missing, the value at that timestamp was set to -999, a value chosen arbitrarily outside the range of the corresponding vital sign.

The potential reason for one or more of the vital signs missing are many. For instance, the purpose for recording medical data is not primarily for research, but for guiding the medical staff and for medical requirements [32]. Table 3.3 shows that SpO_2 is more often used in practice than the Electrocardiography (ECG)-dependent IBI and RR.

3.6 Summary of the Dataset

The dataset consists of both demographic and longitudal variables. The demographic variables are sex, birth weight and the gestational age when hospitalised. Out of the 30 patients included in the dataset, 15 were male and 15 were female. The birth weights ranged from 400g to 2512g, with 24 patients with a VLBW, *i.e.,* a birth weight of less than 1500g. The gestational age of the patients varied between 22.6 weeks to 34.0 weeks, with the mean age being 26.6 weeks. A summary of the demographic data for all patients is seen in Table 3.6. A summary of the same quantities for only the positive and only the negative patients are seen in Tables 3.5 and 3.4, respectively.

Table 3.4: Summary of the demographic data of all patients - minimum, mean, maximum and standard deviation

All Patients	**min**	**mean**	**max**	**std**
Birth weight [g]	400	991.2	2512	513.3
Gestational age [weeks]	22.6	26.6	34.0	3.0

Table 3.5: Summary of the demographic data of the positive patients patients - minimum, mean, maximum and standard deviation

Positive Patients	**min**	**mean**	**max**	**std**
Birth weight [g]	400	892.6	2512	500.1
Gestational age [weeks]	22.6	26.1	34.0	2.9

Table 3.6: Summary of the demographic data of the negative patients - minimum, mean, maximum and standard deviation

Negative Patients	**min**	**mean**	**max**	**std**
Birth weight [g]	568	1221.3	1860	494.6
Gestational age [weeks]	23.4	27.7	31.4	3.2

Table 3.7: Summary of the sex and number of VLBW patients amongst all patients, positive patients and negative patients

	# Patients	Male	Female	VLBW
All	30	15	15	24
Positive	21	10	11	18
Negative	9	5	4	6

3.7 Data Representation

In the following section, the notation of the patient data and the target data are described.

3.7.1 Patient data

The dataset used in the project consists of $N = 30$ patients, whereas the number of modalities are $M = 7$. For the i:th patient, the time series data of modality m is represented as a matrix $\bar{X}_m^{(i)} \in \mathbb{R}^{T_m^{(i)} \times d_m}$, where $T_m^{(i)} \in \mathbb{N}$ is the length of the time series and $d_m \in \mathbb{N}$ is the dimension of the modality, which for instance for the modality *medicine* would be the different sorts of medicines the patient can receive. The data normalisation was performed for each modality $m = 1, 2, ..., 7$ and patient $i = 1, 2, ..., 30$ as

$$\tilde{X}_m^{(i)} = \frac{\bar{X}_m^{(i)} - \mu_m}{\sigma_m}, \tag{3.1}$$

where $\bar{X}_m^{(i)}$ is the un-normalised version of the data of modality m and patient i, μ_m is the mean value of modality m across all training patients and σ_m is the standard deviation of modality m across all training patients. For each modality time series, the set of sorted sampling time points are represented as $\mathcal{T}_m^{(i)} = [t_{m,1}^{(i)}, \ldots, t_{m,T_m^{(i)}}^{(i)}] \in \mathbb{R}^{T_m^{(i)}}$. The time unit was decided to be the post-natal age of the patient measured in hours. So at the time of birth of patient i, $t^{(i)} = 0$. The timestamps of measurements do not need to be the same for different modalities, but the timestamps are on the same scale. Furthermore, for each modality, the time-delta between two consecutive timestamps are defined as

$$\Delta_m^{(i)} = [0, t_{m,2}^{(i)} - t_{m,1}^{(i)}, \ldots, t_{m,T_m^{(i)}}^{(i)} - t_{m,T_m^{(i)}-1}^{(i)}]^\top \in \mathbb{R}^{T_m^{(i)}} \tag{3.2}$$

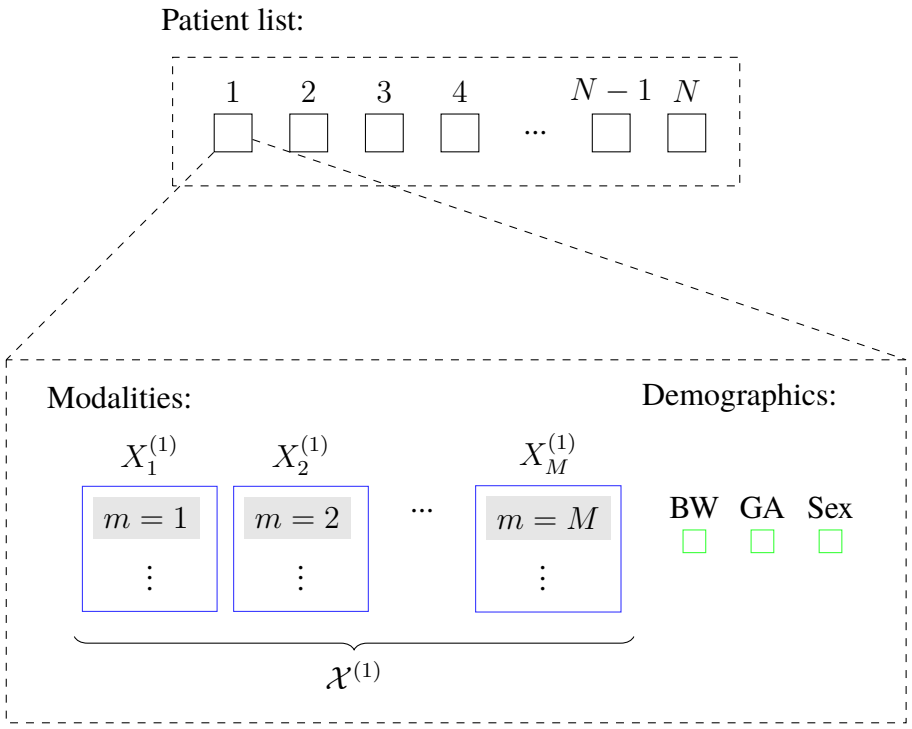


Figure 3.10: For each patient, the patient data consists of data of each modality $m = 1, 2, ..., 7$ and demographic data birth weight (BW), gestational age (GA) and sex.

Given the time of birth $T^{(i)}_{birth}$, the PNA is represented by

$$\alpha^{(i)}_m = [t^{(i)}_{m,1} - T^{(i)}_{birth}, t^{(i)}_{m,2} - T^{(i)}_{birth}, ..., t^{(i)}_{m,T^{(i)}_m} - T^{(i)}_{birth}]^\top \in \mathbb{R}^{T^{(i)}_m}. \tag{3.3}$$

For the i:th patient, the complete data of the m:th modality is represented as the combination of the data, time-deltas and PNAs as

$$X^{(i)}_m = [\tilde{X}^{(i)}_m, \Delta^{(i)}_m, \alpha^{(i)}_m] \in \mathbb{R}^{T^{(i)}_m \times (d_m+2)}. \tag{3.4}$$

Without loss of generality, we set $X^{(i)}_1$, where $m = 1$, to be the vital signs data. The collected data of all M modalities for patient i is then denoted $\mathcal{X}^{(i)} = \{X^{(i)}_m\}^M_{m=1}$. The complete dataset is thus stored per patient, with the structure seen in Figure 3.10, for the example patient $i = 1$.

3.7.2 Target data

The targets of the i:th patient are represented by $Y^{(i)} \in \mathbb{R}^{T^{(i)} \times C}$, where $T^{(i)} = \max\{T_m^{(i)}\}_{m=1}$ is the length of the data, *i.e.,* the length of the longest modality, and $C \in \mathbb{N}$ is the dimension of the target data. Here, the dimension of the target data equals the number of classes the targets are divided into, which are $class0$, LOS and NEC. Thus, $C = 3$. In the target data, all time samples within 24 hours before the patient gets diagnosed with severe event j are labelled as 1 (positive) for the column j. The rest of the time samples are labelled as 0. Since a neonate can experience several severe events at the same time, the positive classes correspond to a multi-label set, meaning several columns corresponding to different severe events can be equal to 1 at the same time sample. The same patient can also experience the same severe event at separate times. The $class0$ is positive, *i.e.*, equal to 1, at the time samples where all of the other classes are negative.

3.8 Model Implementation

The model built during the project was mainly based on the attention-based Transformer network presented by Vaswani *et al.,* [33] as well as varying methods for making attention more memory efficient.

3.8.1 Model construction

The vital signs data were used to compute the queries and the seven modalities (including the vital signs) were used to compute the key-value pairs in the attention mechanism formulation, see (2.4). The process from patient data to estimated outputs is shown in Figures 3.11 and 3.12. For each patient i, each modality is once used to generate the key in the attention modelling to be compared with the queries, *i.e.,* the vital signs. The process is illustrated in Figure 3.11, where $X_1^{(i)}$ represents the vital signs data, and $X_m^{(i)}$ for $m = 1, ..., 7$ represents the data of each of the seven modalities. The linear transforms f_{Q_m}, f_{K_m} and f_{V_m} are specific for each modality m, and create the patient specific Q, K and V matrices. The attention of each modality and patient, $A_m^{(i)}$, is calculated as in (2.4), and the causality matrix as in (2.6). For each modality, the partial output $Z_m^{(i)}$ is generated as in (2.7). The partial outputs of each modality are then concatenated into $Z^{(i)}$, as seen in Figure 3.12. $Z^{(i)}$ is then linearly transformed by the function f_W of size $(CM) \times C$, and added by the linearly transformed version of the demographic data. The

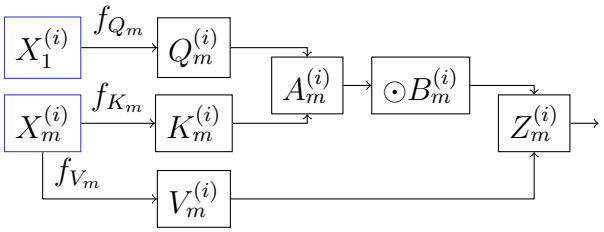


Figure 3.11: Cross-attention layer implementing $Z_m^{(i)} = attn_m^c(X_1^{(i)}, X_m^{(i)})$ for patient i and modality m.

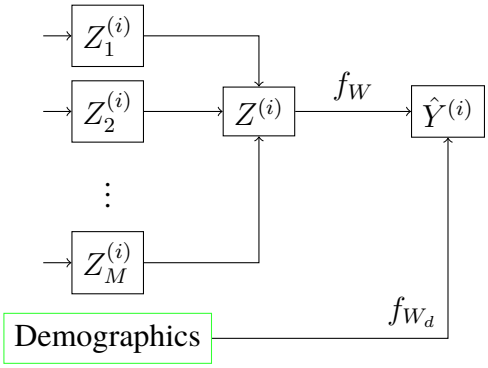


Figure 3.12: The $Z_m^{(i)}$:s are concatenated to $Z^{(i)}$, which through the linear transform f_W and added with the demographic data after the linear transform f_{W_d} yields the estimated output $\hat{Y}^{(i)}$.

demographic data transform, f_{W_d}, is of size $d_{demo} \times C$. The transformed $Z^{(i)}$ and the transformed demographic data are added and yield the predicted outputs for the i:th patient, $\hat{Y}^{(i)}$.

3.8.2 Transforms to Q, K and V

The respective trainable transforms f_{Q_m}, f_{K_m} and f_{V_m} are identical across patients but distinct for each modality $m = 1, 2, ..., 7$. Each transform consists of n_l convolutional layers, all but the last layer followed by a ReLU activation function to avoid vanishing gradients. Out of the n_l layers, the first one is referred to as the input layer and the last one is referred to as the output layer. The input layer is of size $d_1 \times h$ for f_{Q_m} and of size $d_m \times h$ for f_{K_m} and f_{V_m}, where h is the size of the hidden dimensions. The output layer is of size $h \times C$ for all three transforms. The layers between the input and output layers, referred to as the hidden layers, are all of size $h \times h$. The layers were all set to be with bias adjustment and for the same kernel size k. Dropout was lastly

added to all transforms, with a dropout probability p_{do} that was varied, to avoid overfitting.

3.9 Experimental Design

In this section, the environment and tools used during the project are introduced.

3.9.1 Test environment

The test environment was built with a Singularity container. Singularity is a container platform, and the containers enable packaging of software so that it can be run on different systems and operating systems without installing the software [37].

3.9.2 Hardware and software used

Most computations during the project were made on the CMM computational server. The server has four CUDA devices, *i.e.,* the NVIDIA developed platform utilising Graphics Processing Units (GPUs) for parallel computing, each with a maximum capacity of 16,384 MiB (around 17,180 MB). The data pre-processing and handling was done on one device and the model training on another. The data, results, plots, models *etc.,* were stored on the CMM storage server.

All code in the project was written in the Python programming language, version 3.7. The model was created using several Python packages and extensions, amongst others, PyTorch, Pandas, Matplotlib and Numpy.

3.10 Evaluation Metrics

The model evaluation was based on several methods: the loss plot, confusion matrix and AUROC. These methods are briefly described below.

3.10.1 Loss plot

The loss plot is representing the rate at which the model learns for each epoch. A low loss is optimal when searching for the best performing model. The model training can be considered complete when the loss is converging. Here,

both the loss of the training and validation sets are to be presented. The loss function utilised is further described in Section 4.3.2.

3.10.2 Confusion matrix

Confusion matrices illustrate the number of samples correctly classified against the number of samples mis-classified for each class [38]. It also shows what classes the incorrectly classified samples got classified as. If the confusion matrix includes only two classes, the positive repectively negative class, the elements of the confusion matrix can be denoted as the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative Rate (FNR).

3.10.3 AUROC

The AUROC is the area under the Receiver Operating Characteristics (ROC) curve. While the ROC curve shows the TPR against the FPR when varying the threshold value p_{th}, the area under it represents the classification performance of the model [38]. The higher the AUROC value, the higher the probability that the model suspects a random positive sample to be positive compared to a random negative sample.

Chapter 4

Implementation

In this chapter, we describe how the model was implemented, how memory usage was decreased and the model training process.

4.1 Training and Validation Set Split

The data was split into a training and a validation set. Due to the very limited amount of data, no extra test set was used to evaluate our models. The splitting was done patient-wise into a training set and a validation set. The patients were initially split into four patient groups: (1) those who experienced only LOS, (2) the ones that experienced only NEC, (3) the ones that experienced both LOS and NEC and (4) the patients that experienced none of the events. We refer to the latter group as the group of ”negative patients” in the rest of the text. Each of these groups was then separately and randomly split into 70% training and 30% validation patients. The training and validation sets of each group were then merged into one training and one validation set, as illustrated in Figure 4.1. This lead to a training set consisting in 19 patients, and a validation set consisting of eleven patients. The list of patients within training and validation set, respectively, was then shuffled.

4.2 Limit Memory Usage and Increase Computational Speed

To limit memory usage and increase the computational speed for the data processing and model training, besides limiting the number of patients, downsampling of the vital signs data and timelime limiting, other actions were

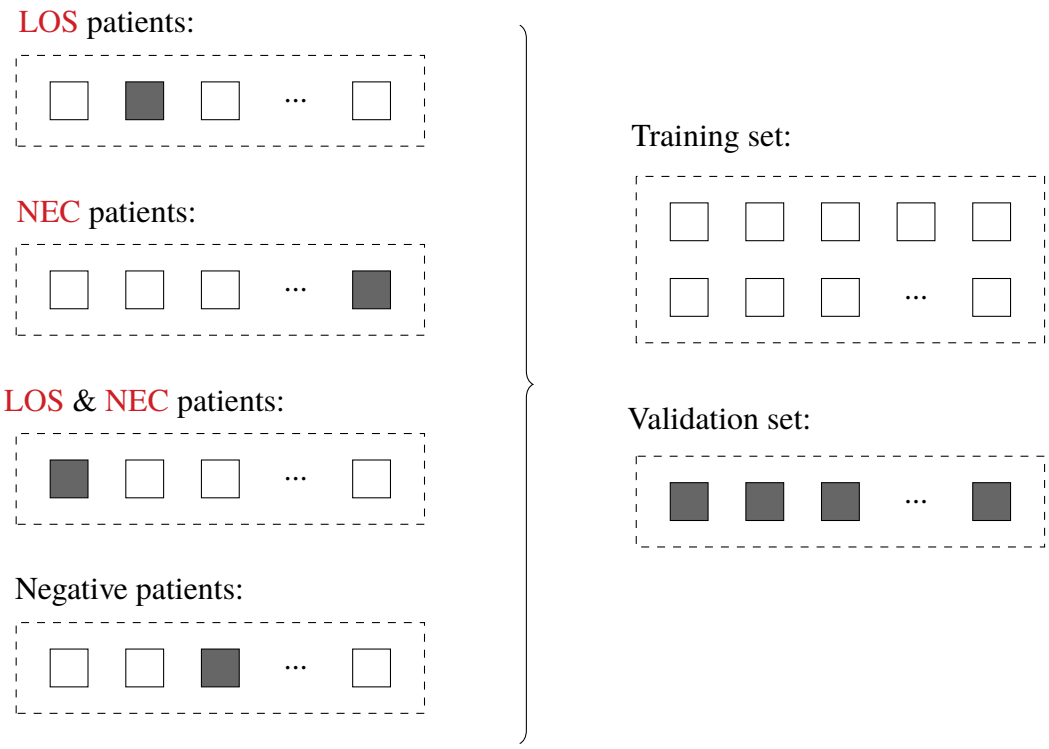


Figure 4.1: The data splitting process, where each of the groups of only LOS patients, only NEC patients, patients with both LOS and NEC and negative patients are randomly split 70/30 and then merged to the training and validation set. Each square correspond to one patient. The gray filled square correspond to patients attributed to the validation set in their respective group.

taken. The model was trained using buckets, described in Section 4.2.1, and the attention matrices were yielded by implementing log-steps, presented in Section 4.2.2.

4.2.1 Complexity reduction for model training

The cross-attention model complexity grows quadratically with the length of the query timeseries, as $\mathcal{O}(T \cdot T')$. The maximal length of the vital signs data is $T = T' = 250560$ (Table 3.2). Therefore, methods to decrease the model complexity have been developed, for instance in by Rabe and Staats [39], Liu *et al.,* [40], Dao *et al.,* [41] and Dao [42]. When designing the model for this project, the ideas from these sources together with code from Github inspired by these sources [43] have been used to design a faster transformer network. The reduction in complexity is obtained by performing the attention on sub-pieces of the queries and keys. Both components are divided into buckets of bucket size b, leading to $n_Q = \lceil \frac{T}{b} \rceil$ query buckets and $n_K = \lceil \frac{T'}{b} \rceil$ key buckets, as

$$Q = \left[\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{n_Q}\right]^T \tag{4.1}$$

$$K = \left[\mathbf{k}_1, \mathbf{k}_2, ..., \mathbf{k}_{n_K}\right]^T. \tag{4.2}$$

The attention mechanism is then performed on each combination of query bucket q_i, and key bucket k_j, for $\forall i \in [n_Q]$ and $\forall j \in [n_K]$. The values are fragmented into buckets of the same sizes as the key buckets, paired with the key buckets. The mask matrix for the bucket combination is computed as described in (2.6).

4.2.2 Complexity reduction for attention matrix computations

When generating the attention matrices, as described in (2.4), the bucket based attention computations described above cannot be utilised since they do not generate the complete matrix $A \in \mathbb{R}^{T \times T'}$. Thus, the model to generate the attention matrices had to be implemented using another type of complexity reduction. It was decided to use log-steps for the key sequences, meaning not all measurements for the modalities were used to train the model, but only the ones satisfying

$$t' = \left[2^k - 1\right]_{k=0}^{\lfloor log_2(T') \rfloor} \tag{4.3}$$

Consequently, only the measurements at timestamps $t' = 0, 1, 3, 7, 15, ..., 2^{\lfloor log_2(T') \rfloor} - 1$ of each modality were used to train the model. For instance, if $T' = 250560$, *i.e.,* the maximal modality length, then $2^{\lfloor log_2(T') \rfloor} - 1 = 131071$ and the number of timestamps t' are eighteen. The query sequence was not reduced. The final matrix A is thereby reduced, but timestamps from the whole key sequence are still used, with more timestamps available closer to the beginning of the time series.

4.3 Hyperparameters

The model hyperparameters that could be adjusted were:

- b - the bucket size
- lr - the learning rate
- n_l - the number of convolutional layers in the Q, K and V transforms
- h - the dimension of the hidden convolutional layers in the Q, K and V transforms
- k - the kernel size of the hidden convolutional layers in the Q, K and V transform
- p_{do} - the dropout probability
- p_{th} - the threshold value for a class to be predicted as positive

Table 4.1: Selected values for each model parameter

Parameter	**Selected value**
b	4096
lr	0.001
n_l	3, 5, 7
h	60
k	60
p_{do}	0, 0.3, 0.7
p_{th}	0.5

The selected values for each of these parameters are seen in Table 4.1. The motivation behind the batch size, b, choice was due to hardware limitations arising during the backpropagation step. After trying $b = 1024, 2048, 3072, 4096, 5120$, it was concluded that the computational speed decreased with the batch size while the maximal memory available limited batch sizes larger than 5120.

For the learning rate, the values that were evaluated were 0.001 and 0.0001, where the former generated faster and more steadily decreasing training and validation loss. Thus, $lr = 0.001$ was selected.

The number of layers, n_l, were varied, to decide on how model complexity affected the performance. As the dropout probability, p_{do}, was also varied, different variations on parameters increasing the model complexity and parameters decreasing the model complexity were investigated.

The size of the hidden dimension in the transforms, h, and the kernel size, k, were both set to 60, as it aligned well with the implementation of the memory efficient transformers and the sample rate of the modalities.

The probability threshold is relevant for classiying the model predictions. As the outputs of the model are values between 0 and 1, we used a fix threshold to perform the final detection of whether a class is predicted as positive or not for each time sample. The probability threshold is the value for which $(\hat{Y}[k] > p_{th}) = 1$ and $\hat{Y}[k] \leq p_{th} = 0$, *i.e.,* class k is estimated to be positive or negative. Since each sample is either positive or negative, the threshold was set to be $p_{th} = 0.5$, *i.e.,* the probability for the sample to be positive must be greater than the possibility for it to be negative.

4.3.1 Optimiser

The optimiser used during the project was the Adam optimiser. The Adam optimiser was used since it is usually suitable for deep learning models like the one implemented during this project, and since it was the optimiser used by Vaswani *et al.,* in the Transformer network [33], by Chauhan *et al.,* in the Perceiver model [32] and by Ge *et al.,* in Recurrent Neural Network (RNN)-based models with attention mechanisms [36]. More about the Adam optimiser can be read in Goodfellow *et al.,* [44].

4.3.2 Loss function

The loss function used to train the model is a multi-label version of the soft margin loss and is based on max-entropy. It was selected due to

its appropriateness for multi-label data. In Python, the function is called `MultiLabelSoftMarginLoss` in the PyTorch library.

4.3.3 Class weights

To adjust for the class imbalance, *i.e.*, the varying sample sizes for $class0$, LOS and NEC, class weights were utilised. If not using class weights, the risk of prediction bias increases. The weights were calculated as inverse proportion of the sample occurrences of each class across all patients in the in the training data set. An indication of the class occurrences in the training set can be given by the number of occurrences across all patients seen in Table 3.1. Thereby, the class with the lowest occurrence in the training dataset would receive the highest weight among the classes. Each sample in the labels were then multiplied by its respective class weight when calculating the loss during the model training iterations. The same weights were used when estimating the validation loss, to give comparable results with the training loss.

4.4 Model Training

The model was trained for varying number of layers, $n_l = [3, 5, 7]$, and dropout probabilities, $p_{do} = [0, 0.3, 0.5]$. After experimenting, the learning rate was set to $l_r = 0.001$. The models were all trained for 100 epochs. We found experimentally that this ensured sufficient convergence for all our combinations of hyperparameters. The same training/validation set split was used across all models for consistency and to ease comparison. The training was done through back-propagation. After training was completed, the softmax activation function was used on the model output to generate the final estimates $\hat{Y}$, at each time sample in the range 0 to 1. When the best model was found, the same hyperparameter values were used to perform 5-fold cross-validation, *i.e.*, the model was trained again but with another four distinct training/validation set splits.

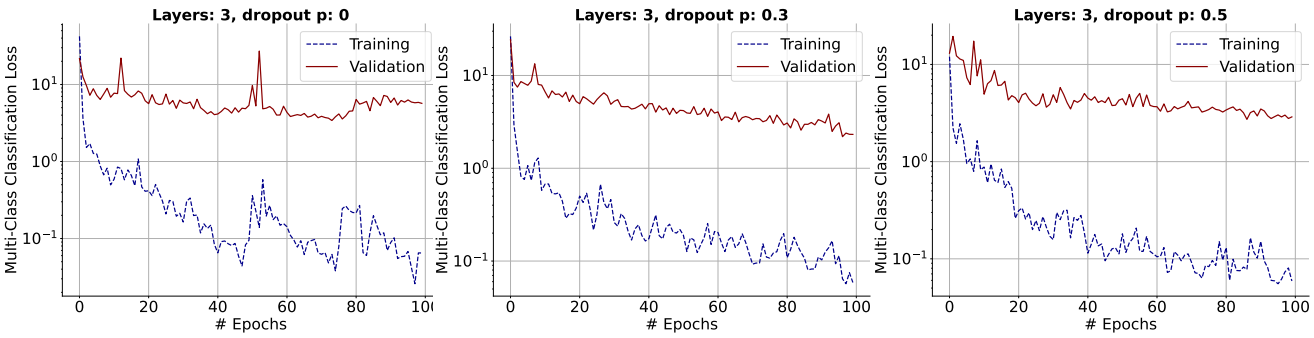
Chapter 5

Results

In this chapter, the major results are presented. These mainly consist of the evaluation metrics described in Chapter 3.10, along with some examples of attention matrices and plots of prediction by the final model.

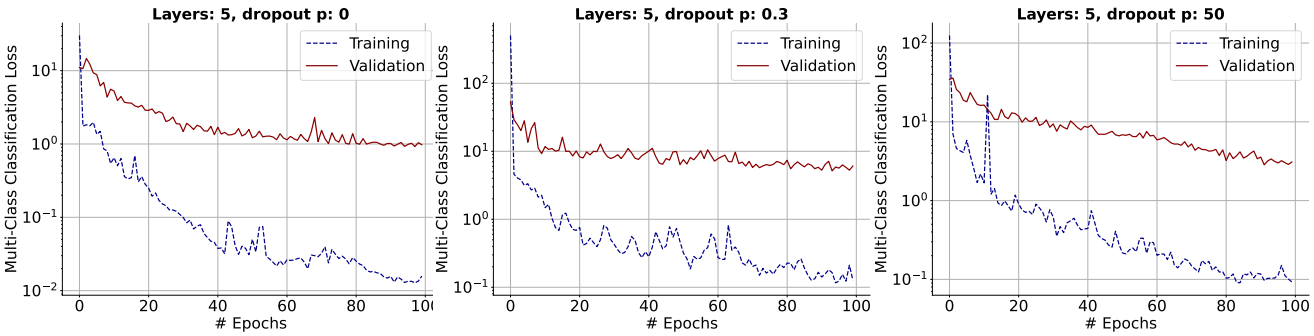
5.1 Loss Plots

The Figures 5.1, 5.2 and 5.3 show the loss plots for 100 epochs when the number of layers, n_l, in the Q, K and V transforms were 3, 5 and 7, respectively. For each number of layers, the dropout probabilities $p_{do} = 0, 0.3, 0, 5$ are evaluated. In each sub-figure, both the training and validation loss per epoch are presented.



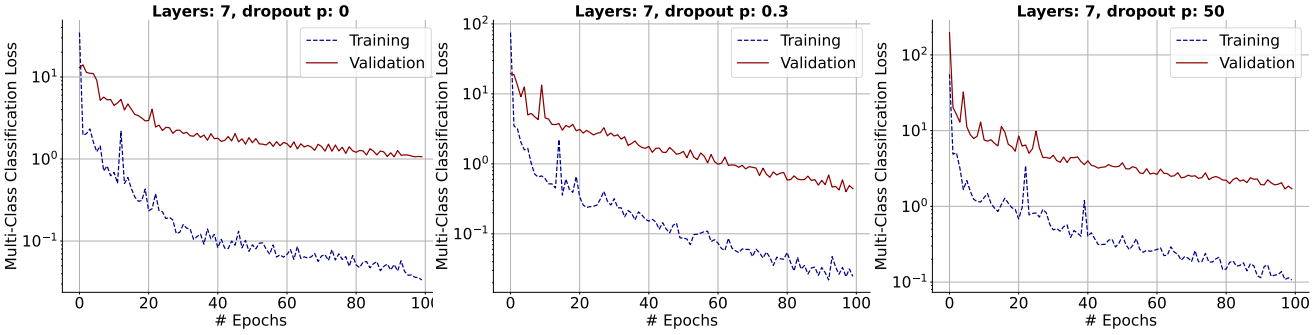
(a) (b) (c)

Figure 5.1: Losses for models with 3 layers and (a) $p_{do} = 0$, (b) $p_{do} = 0.3$ and (c) $p_{do} = 0.5$



(a) (b) (c)

Figure 5.2: Losses for models with 5 layers and (a) $p_{do} = 0$, (b) $p_{do} = 0.3$ and (c) $p_{do} = 0.5$



(a) (b) (c)

Figure 5.3: Losses for models with 7 layers and (a) $p_{do} = 0$, (b) $p_{do} = 0.3$ and (c) $p_{do} = 0.5$

5.2 AUROC

To create the AUROC, the two positive classes LOS and NEC were merged togehter after prediction to be able to define the TPR and the FPR. The best performing model in terms of highest AUROC in combination with low training and validation losses was the model with $n_l = 7$ and $p_{do} = 0$. For this model, the AUROC was 0.66. The ROC of the model is seen in Figure 5.4.

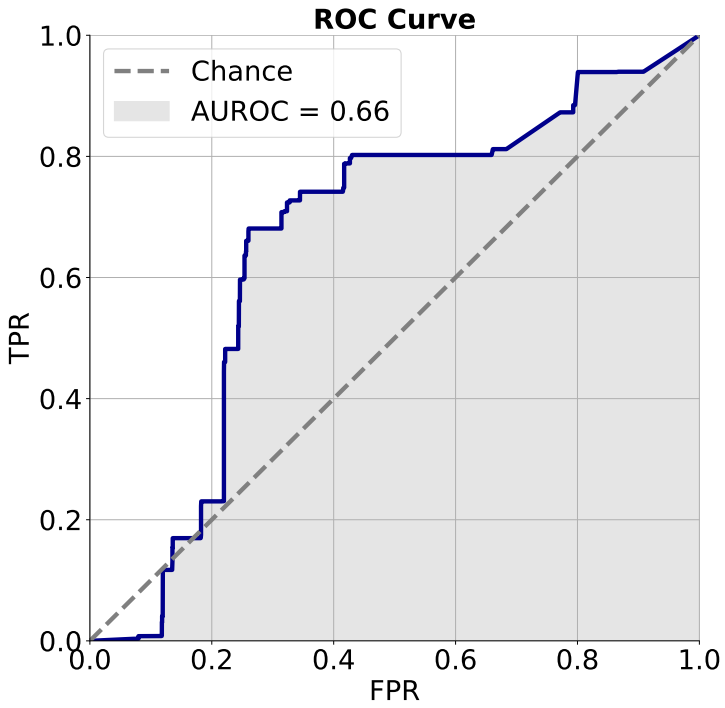


Figure 5.4: ROC curve and AUROC score of best performing model with merged positive classes

After performing 5-fold cross-validation, the average AUROC score across the five models was 0.586.

5.3 Confusion Matrices

For the best performing model, the confusion matrix can be seen in Figure 5.5.. The matrix entries have been normalised along the True class axis. Figure 5.6

shows the confusion matrix for the same model when the LOS and NEC classes have been merged into one positive class, after prediction. For both confusion matrices, the predictions had the threshold value $p_{th} = 0.5$, indicating a class must be more probable to be positive than negative according to the model in order to be predicted as positive.

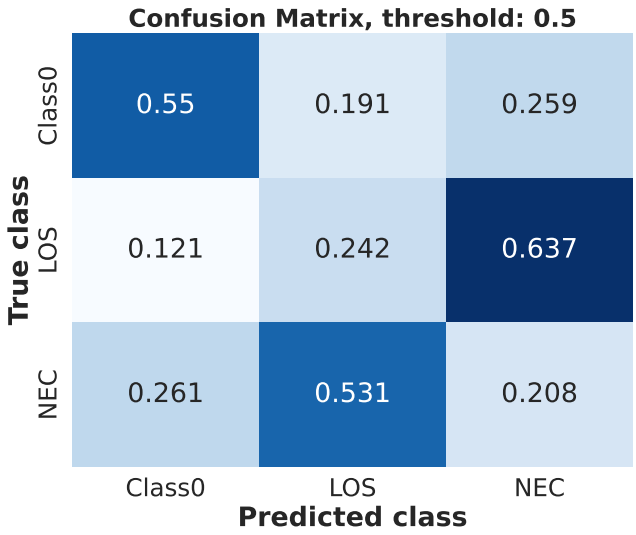


Figure 5.5: Confusion matrix of best performing model

The average confusion matrix, with merged positive classes, after 5-fold cross-validation is seen in Figure 5.7.

5.4 Attention Matrices

Examples of two attention matrices, yielded as described in Section 4.2.2, for the modalities vital signs and FiO_2 for patient 3 is shown in Figures 5.8a and 5.8b.

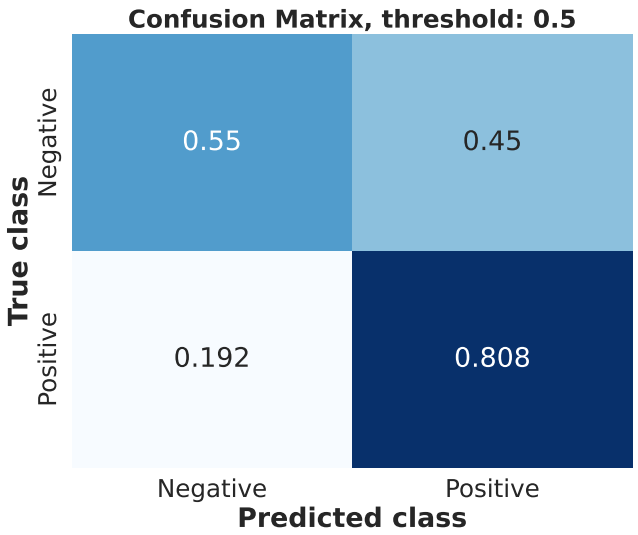


Figure 5.6: Confusion matrix of best performing model with merged positive classes

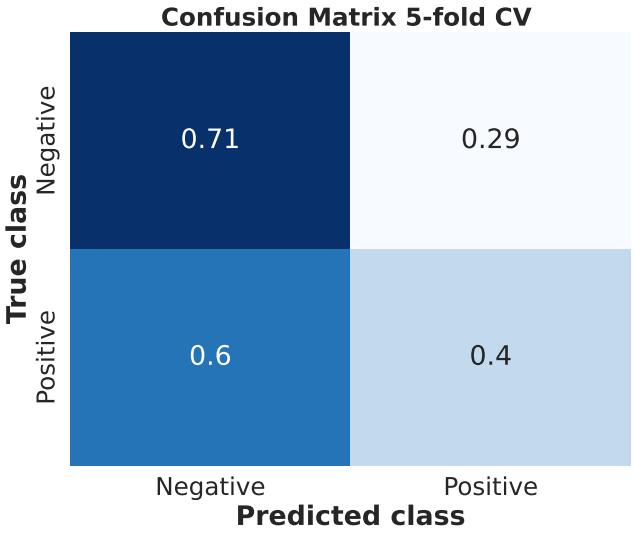
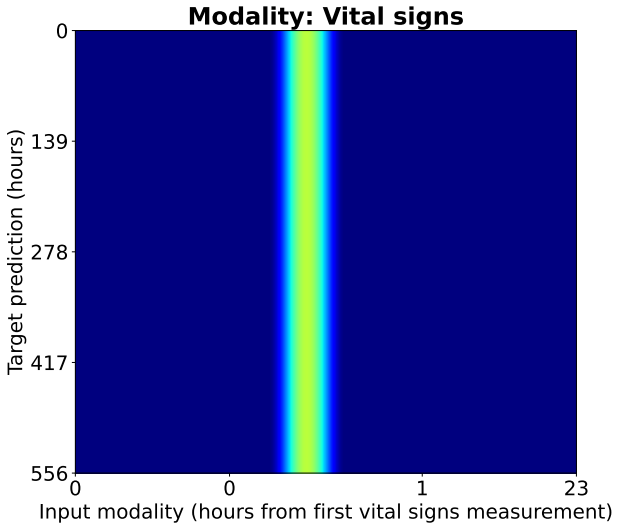
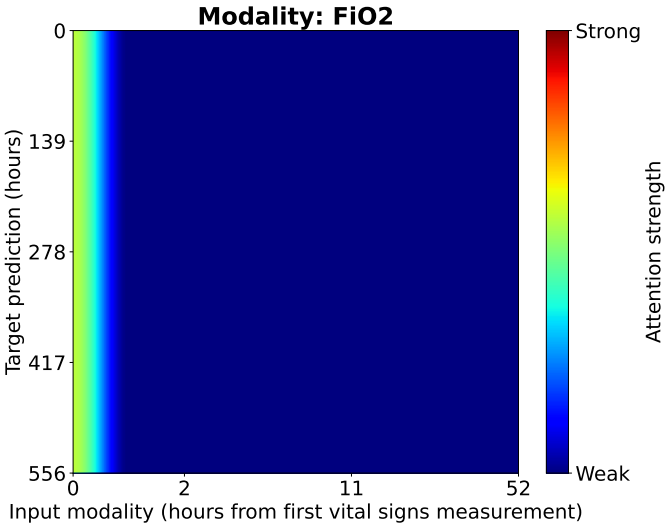


Figure 5.7: Confusion matrix after 5-fold cross-validation (CV) and merged positive classes



(a) Vital signs attention plot for patient 3



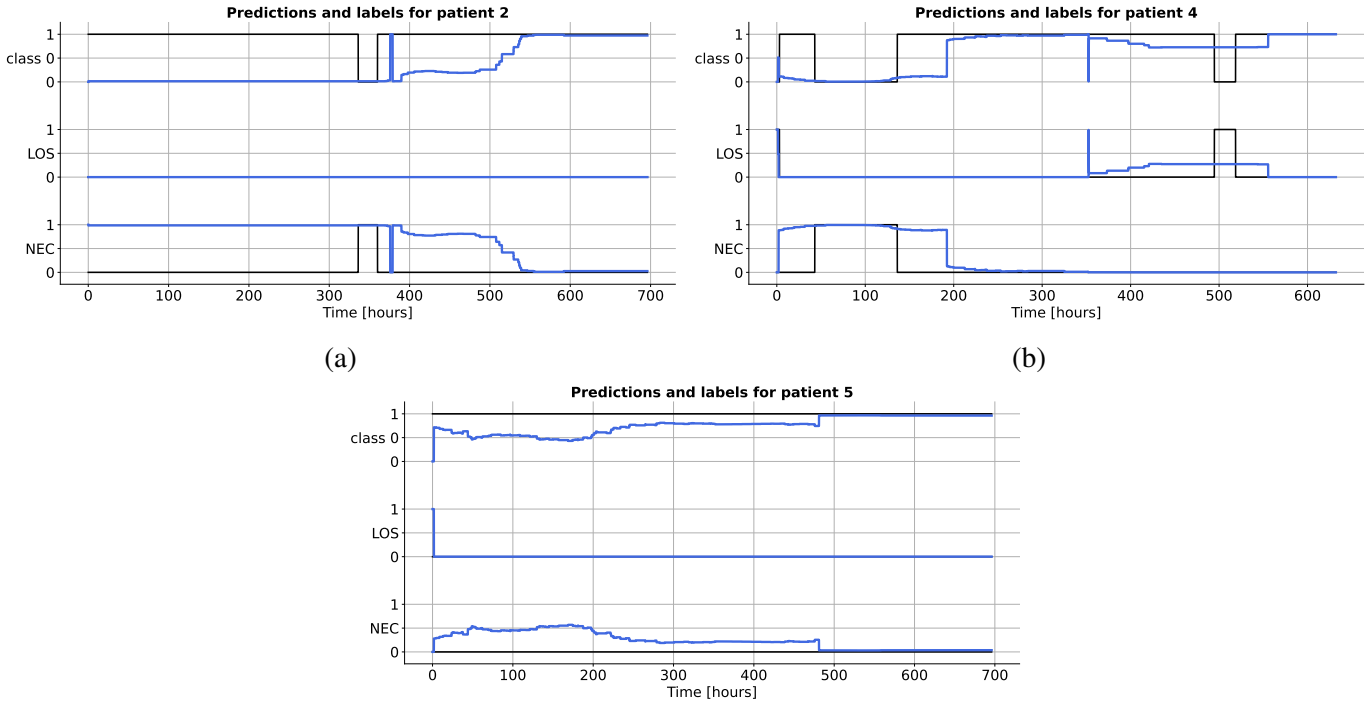
(b) $FiO2$ attention plot for patient 3

5.5 Examples of Predictions

Below follows model predictions on earlier presented patients.

5.5.1 Training set patients

Patient 2, patient 4 and patient 5 were randomly selected to be included in the training set when training the model. After being trained for 100 epochs, the best performing model did the target predictions seen in Figure 5.9.



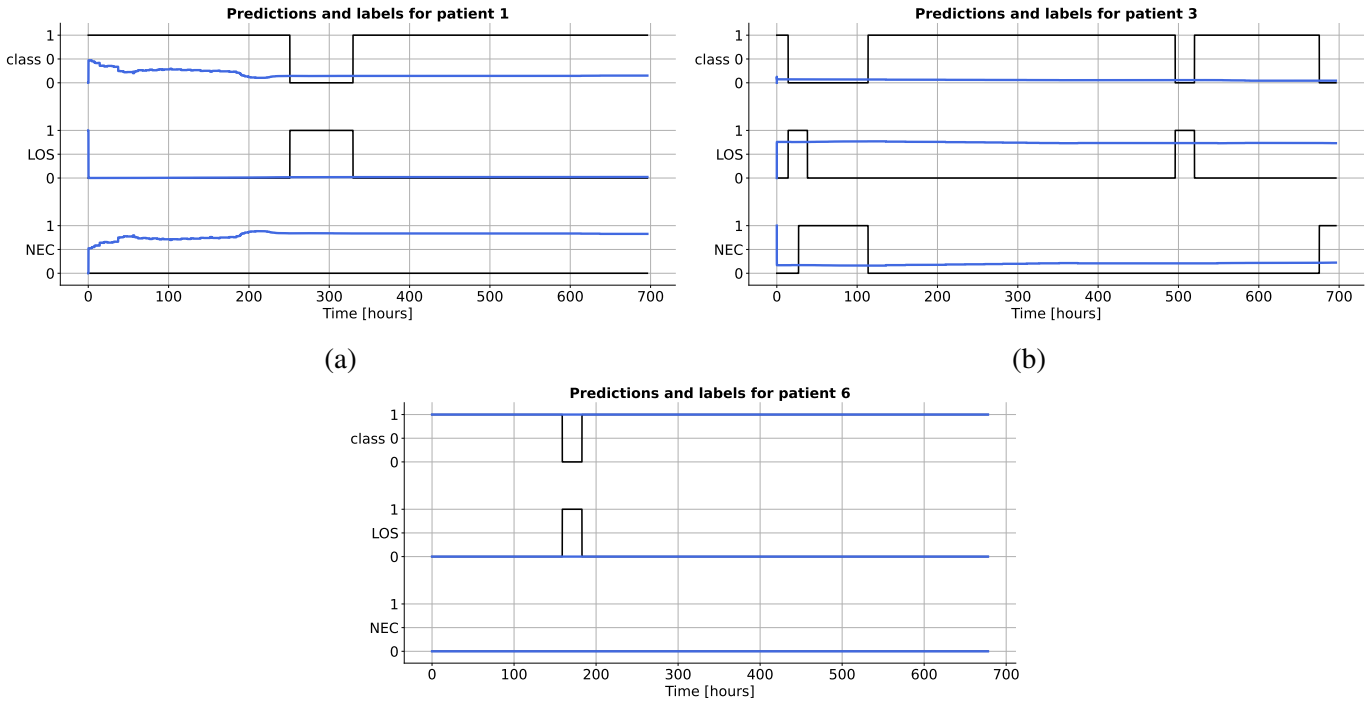
(a) (b)

(c)

Figure 5.9: Labels (black) and predictions (blue) of training patients (a) 2, (b) 4 and (c) 5

5.5.2 Validation set patients

Patient 1, patient 3 and patient 6 were randomly selected to be included in the validation set when training the model. After being trained for 100 epochs, the best performing model did the target predictions seen in Figure 5.10.



(a) (b)

(c)

Figure 5.10: Labels (black) and predictions (blue) of validation patients (a) 1, (b) 3 and (c) 6

5.6 Time per Epoch

The time in minutes it took to train one epoch, on average, for the models with different number of layers, are presented in Table 5.1.

Table 5.1: Average time per epoch

Number of layers:	3	5	7
Time [min/epoch]	5.6	5.8	6.0

Chapter 6

Discussion

6.1 Results Analysis

In this section, the results presented in Chapter 5 are analysed and discussed.

6.1.1 Loss plots

The loss plots show that when increasing the number of layers, both training and validation losses typically get lower after running for 100 epochs. Furthermore, adding dropout to the model does generally not improve the model performance in terms of minimising the loss. It turned out the model achieving the lowest validation loss was the one with $n_l = 7$ and $p_{do} = 0.3$, and the model achieving the lowest training loss was the one with $n_l = 5$ and $p_{do} = 0$. However, neither of these models were the best performing ones when also evaluating AUROCs and confusion matrices.

6.1.2 AUROC

As seen in Figure 5.4, the best performing model does not perform much better than chance when distinguishing between positive and negative samples. The AUROC score after performing 5-fold cross-attention is even lower, indicating that the models have a hard time distinguishing between the negative and positive class. This score can be compared to the ones presented in Section 2.3 for models implemented for similar purposes, which acheived significantly higher scores. However, it is always difficult to compare studies using different datasets, and furthermore taking into account that the model implemented in this project is a multi-label problem while as the other studies focused on only

LOS.

6.1.3 Confusion matrices

It becomes clear when analysing the confusion matrix in Figure 5.5 that LOS and NEC more often than not get confused, *i.e.,* the true NEC samples most often get confused for LOS and vice versa. This explains why the number of correctly predicted positive samples are high when merging the two positive classes, seen in Figure 5.6. This confusion matrix indicates the model often finds the positive samples, but it also falsely predicts almost half the negative samples as positive. The confusion matrix resulting from averaging the five model results from the 5-fold cross-validation, seen in Figure 5.7, suggests that the models are generally better at predicting the negative samples correctly than the positive ones.

6.1.4 Attention matrices

The two attention matrices in Figure 5.8a and 5.8b for patient 4 show that much of the focus of the model is put into the beginning of the input modalities when trying to predict the targets. This result is however very limited as it only concerns one patient, and the data is majorly reduced through both down-sampling and log-spaces, yielding more samples at the beginning of the modalities.

6.1.5 Examples of predictions

For each of the presented patients, the model predictions show some interesting features:

1. Patient 1: The model detects a severe event early, but confuses NEC with LOS
2. Patient 2: The model predicts NEC but starts predicting it too early
3. Patient 3: The model predicts positive labels, stronger for LOS than for NEC
4. Patient 4: The model predicts the first NEC event almost correctly, then predicts LOS correctly but a bit early and with low probability
5. Patient 5: The model predicts NEC after around 50 hours, although the patient is negative

6. Patient 6: The model fails to detect the LOS event around hour 150

As presented in Section 3.4, patient 4, patient 5 and patient 6 all have missing vital signs data for different periods of time.

Patient 6 has missing vital signs data almost from the beginning until directly after the diagnosis at around hour 290. Furthermore, patient 2 also has missing vital signs data between around hour 10 and hour 370, *i.e.,* until after the NEC event at around hour 320.

Thus, patient 1 shows an example of the model confusion between LOS and NEC, which aligns with the results seen in the confusion matrix in Figure 5.5. Patient 2 exemplifies how the model acts when vital signs data is missing during the positive event. For patient 3, experiencing several LOS and NEC events, the model identifies it early but also gives a stronger indication of LOS than NEC. It furthermore predicts the events as positive between the positive labels, which might not be medically incorrect but will be incorrect according to the label definitions in this project. For patient 4, the model shows an almost perfect prediction of the NEC event, but followed by an uncertain but not completely inaccurate prediction of the LOS event. The predictions for patient 5 shows the case of the model predicting NEC when the patient labels are actually $class0$ at all times. As seen in the confusion matrix in Figure 5.5, the model falsely predicts NEC almost 26% of the true $class0$ samples. For patient 6, the model fails to detect the sepsis event, which might be explained by the missing data at the time of the event and shows another behaviour of the model when data is missing than seen for patient 2. Generally, the model performs poorly when vital signs data is missing, which might not be a surprise as the model is built upon the vital signs data.

6.1.6 Time per epoch

The short presentation of the time in minutes per epoch of training for different number of layers in the models shows that the time slightly increases with the model complexity, as expected. Unfortunately, it was more difficult to analyse the memory consumption when training the different models, due to the computational servers being shared amongst all CMM workers.

6.2 Project Evaluation

The three sub-goals of the project were presented in Section 1.4. All three of them were accomplished: a functioning MHCA model was implemented, the

model performance was evaluated and the attention maps were presented and analysed. However, the conclusions that can be drawn from the attention maps are limited, as they are patient specific and due to the low performance of the final model.

The purpose of the project, as presented Section 1.3, was to create an insights in the adverse event's disease course by examining the attention model results. This is not completely accomplished, since the poor model performance hinders making any clear conclusions based on its results. However, this project presented an example of how cross-attention can be used for medical data with several modalities, which could be used as an example or base for further research on the topic.

6.3 MHCA-models for Severe Events Detection in Neonates

Overall, MHCA models theoretically seem appropriate for detecting adverse events for neonates, due to their handling of irregularly sampled time series and the possibility of complexity reduction methods such as the one utilised in this project. The AUROC score and confusion matrices show that the models in this project do not outperform other ML-models with similar purposes. However, further optimisation of the model should be done before making a fair comparison.

Chapter 7

Conclusions

7.1 Conclusions

As presented in the Discussion, Chapter 6, the final model is lacking some prediction quality. Combined with the small dataset and quantity of missing data, it cannot be confidently concluded whether MHCA models are appropriate for detecting the severe neonatal events LOS and NEC. The topic should be further investigated, using more data and with further fine-tuning of the hyperparameters described in Section 4.3 and Table 4.1. More computational power might also enable to avoid downsampling, and thus utilise the original sampling of 1 Hz, thus not removing information in the vital signs data. More ideas for future work are described in Section 7.3.

7.2 Limitations

The three main limitations of the project were:

1. The time scope - The limited period of time for conducting the project lead to the model not being appropriately optimised, and thus might lead to impaired model performance.

2. The computational resources - Although the CMM servers provide a significant amount of computational power, even more power would be necessary to fully use the available data and optimise hyper-parameters within a reasonable time.

3. The data available - As the database of the Karolinska NICU departments is continuously being updated with annotations, only data

from a limited number of former patients were available when starting this project. Including data from patients at other NICUs and more positive patients would improve the opportunities for the model to be well-trained. Furthermore, it would hopefully compensate for the significant amount of missing data for the current patients.

7.3 Future Work

Due to the limitations presented in the previous section, there are some remaining issues that should be adressed in future studies on the topic presented in this project. A selection of the possible future work's are presented below.

7.3.1 Increased dataset

As the number of patients in the final dataset were only 30, of whom 21 experienced any of the severe events, it would be of interest to extend the dataset with more patients. Increasing the amount of data available for training would enhance the model's ability to generalise effectively, learn important features and make it more robust against diversity among patients as well as outlier data.

7.3.2 Increase computational power

One major limitation during the project was the limited computational power. With increased computational power, more complex models could be trained. Optimally, it would also eliminate the need for downsampling of the vital signs data, which could be a major drawback as the learning was based on the vital signs data. Thus, downsampling reduces the data resolution which might impact the model performance. One interesting factor for future works would thus be to investigate the impact on the model training if the non-downsampled version of the vital signs data is used. It would also be of interest to use the entire modality datasets to compute the attention matrices. This could provide interesting insights in where the model has the highest attention for each of the modalities.

7.3.3 Further parameter exploration

As the attention-based model created during this project is complex and with a considerable amount of parameters that could be optimised, it would be of interest to further explore different values of these. Some parameters were set before training, such as the hidden dimension h and the kernel size k. Furthermore, evaluating extended ranges for the number of layers n_l and dropout probability p_{do} might also lead to an improved model.

7.3.4 Inclusion of more adverse events

As mentioned in the Section 2.1, the original dataset consisted of more severe events for neonates, besides LOS and NEC. It would be of scientific interest to investigate if vital sign-based cross-attention models could be utilised to detect additional neonatal adverse events.

7.3.5 Explore other label definitions

During this project, positive events were defined as the 24 hours before the first clinical suspicion or positive blood culture of one of the events. However, this definition is simplified and symptom onsets could present before or after those 24 hours. Another option would be to create some type of smooth labels, so that the labels would not only be 0 or 1, but gradually increase and decrease around the timestamp the event was detected. An alternative approach would be to introduce penalties factors, or regularisation, in the model. That way, the model could learn that it is more important to predict a positive event close before the detection timestamp, while it might not be as important after the clinicians have already detected it.

7.4 Reflections

Conclusively, this project has provided some initial insights in the use of attention-based models for detection of neonatal LOS and NEC. Further studies on the topic needs to be conducted to get an indication on the true potential of these types of models.

References

[1] C. Fleischmann, F. Reichert, A. Cassini, R. Horner, T. Harder, R. Markwart, M. Tröndle, Y. Savova, N. Kissoon, P. Schlattmann, K. Reinhart, B. Allegranzi, and T. Eckmanns, "Global incidence and mortality of neonatal sepsis: a systematic review and meta-analysis," *Archives of Disease in Childhood*, vol. 106, no. 8, pp. 745–752, 2021. doi: 10.1136/archdischild-2020-320217. [Online]. Available: https://adc.bmj.com/content/106/8/745 [Page 2.]

[2] N. Samuels, R. A. van de Graaf, R. C. de Jonge, I. K. Reiss, and M. J. Vermeulen, "Risk factors for necrotizing enterocolitis in neonates: a systematic review of prognostic studies," *BMC Pediatrics*, vol. 17, 04 2017. doi: 10.1186/s12887-017-0847-3. [Online]. Available: https://doi.org/10.1186/s12887-017-0847-3 [Pages 2 and 9.]

[3] R. H. Clark, P. Gordon, W. M. Walker, M. Laughon, P. B. Smith, and A. R. Spitzer, "Characteristics of patients who die of necrotizing enterocolitis," *Journal of Perinatology*, vol. 32, no. 3, pp. 199–204, 03 2012. doi: 10.1038/jp.2011.65. [Online]. Available: https://www.nature.com/articles/jp201165 [Pages 2, 13, and 14.]

[4] B. Sullivan and K. Fairchild, "Vital signs as physiomarkers of neonatal sepsis," *Pediatric Research*, vol. 91, pp. 1–10, 09 2021. doi: 10.1038/s41390-021-01709-x [Pages 2, 4, 8, 11, 12, 13, and 14.]

[5] L. B. Mithal, R. Yogev, H. L. Palac, D. Kaminsky, I. Gur, and K. K. Mestan, "Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis," *Early Human Development*, vol. 117, pp. 83–89, 2018. doi: https://doi.org/10.1016/j.earlhumdev.2018.01.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037837821730350X [Pages 2, 9, and 14.]

[6] J. H. Kim, V. Sampath, and J. Canvasser, "Challenges in diagnosing necrotizing enterocolitis," *Pediatric Research*, vol. 88, no. 1, pp. 16–20, August 2020. doi: 10.1038/s41390-020-1090-4. [Online]. Available: https://doi.org/10.1038/s41390-020-1090-4 [Pages 2 and 10.]

[7] W. H. Organization, *Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions*. World Health Organization, 2020. [Page 2.]

[8] Z. Peng, G. Varisco, R.-H. Liang, D. Kommers, W. Cottaar, P. Andriessen, C. Pul, and X. Long, "Deeplos: Deep learning for late-onset sepsis prediction in preterm infants using heart rate variability," *Smart Health*, vol. 26, p. 100335, 10 2022. doi: 10.1016/j.smhl.2022.100335 [Pages 2, 8, 9, and 14.]

[9] L. C. Downey, P. B. Smith, and D. K. Benjamin, "Risk factors and prevention of late-onset sepsis in premature infants," *Early Human Development*, vol. 86, no. 1, Supplement, pp. 7–12, 2010. doi: https://doi.org/10.1016/j.earlhumdev.2010.01.012 Proceedings from the 2nd International Conference on: Nutrition of the Preterm Infant: Current Issues. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378378210000149 [Pages 2, 9, 13, and 14.]

[10] S. N. Shukla and B. M. Marlin, "Multi-time attention networks for irregularly sampled time series," 2021. [Pages 3, 15, and 18.]

[11] K. Fairchild, D. Lake, J. Kattwinkel, J. Moorman, D. Bateman, P. Grieve, J. Isler, and R. Sahni, "Vital signs and their cross-correlation in sepsis and nec: A study of 1,065 very-low-birth-weight infants in two nicus," *Pediatric research*, vol. 81, 12 2016. doi: 10.1038/pr.2016.215 [Pages 4 and 11.]

[12] Preterm birth. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/preterm-birth [Page 7.]

[13] J. B. Linde. Prematur födsel (föda för tidigt). [Online]. Available: https://www.karolinska.se/for-patienter/graviditet-och-forlossning/dags-att-foda/prematur-fodsel-foda-for-tidigt/ [Page 7.]

[14] F. Reiterer, "Neonatal pneumonia," in *Neonatal Bacterial Infection*, B. Resch, Ed. Rijeka: IntechOpen, 2013, ch. 2. [Online]. Available: https://doi.org/10.5772/54310 [Page 7.]

[15] M. T. Melville JM, "The immune consequences of preterm birth," *Frontiers in Neuroscience*, vol. 7, 05 2013. doi: 10.3389/fnins.2013.00079. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3659282/ [Page 7.]

[16] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, and D. C. Angus, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 02 2016. doi: 10.1001/jama.2016.0287. [Online]. Available: https://doi.org/10.1001/jama.2016.0287 [Pages 8 and 9.]

[17] J. F. Schneider, "Neonatal brain infections," *Pediatric Radiology*, vol. 41, 05 2011. doi: 10.1007/s00247-011-2041-3. [Online]. Available: https://doi.org/10.1007/s00247-011-2041-3 [Page 9.]

[18] B. A. Shah and J. F. Padbury, "Neonatal sepsis: an old problem with new insights," *Virulence*, vol. 5, no. 1, p. 170—178, January 2014. doi: 10.4161/viru.26906. [Online]. Available: https://europepmc.org/articles/PMC3916371 [Page 9.]

[19] Y. Su, R.-H. Xu, L.-Y. Guo, X.-Q. Chen, W.-X. Han, J.-J. Ma, J.-J. Liang, L. Hao, and C.-J. Ren, "Risk factors for necrotizing enterocolitis in neonates: A meta-analysis," *Frontiers in Pediatrics*, vol. 17, 01 2023. doi: 10.3389/fped.2022.1079894. [Online]. Available: https://doi.org/10.1186/s12887-017-0847-3 [Pages 9, 12, 13, and 14.]

[20] K. B. Blennow. Necrotising enterocolitis (nec). [Online]. Available: https://www.karolinska.se/for-patienter/graviditet-och-forlossning/sjuka-nyfodda-barn/neonatal-care/diagnoses/necrotising-enterocolitis-nec/ [Pages 9 and 10.]

[21] W. Hsueh, M. S. Caplan, X.-W. Qu, X.-D. Tan, I. G. D. Plaen, and F. Gonzalez-Crussi, "Neonatal necrotizing enterocolitis: Clinical considerations and pathogenetic concepts," *Pediatric and Developmental Pathology*, vol. 6, 01 2003. doi: 10.1007/s10024-002-0602-z. [Online]. Available: https://doi.org/10.1007/s10024-002-0602-z [Page 9.]

[22] A. Honoré, D. Forsberg, K. Adolphson, S. Chatterjee, K. Jost, and E. Herlenius, "Vital sign-based detection of sepsis in neonates using

machine learning," *Acta Paediatrica*, vol. 112, no. 4, pp. 686–696, 2023. doi: https://doi.org/10.1111/apa.16660. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/apa.16660 [Pages 11, 14, and 18.]

[23] M. Nitzan, A. Romem, and R. Koppel, "Pulse oximetry: Fundamentals and technology update," *Medical devices (Auckland, N.Z.)*, vol. 7, pp. 231–9, 07 2014. doi: 10.2147/MDER.S47319 [Page 11.]

[24] S. Reuter, C. Moser, and M. Baack, "Respiratory distress in the newborn," *Pediatrics in review / American Academy of Pediatrics*, vol. 35, pp. 417–29, 10 2014. doi: 10.1542/pir.35-10-417 [Page 11.]

[25] V. Berggren. Respirator. [Online]. Available: https://www.karolinska.se/for-patienter/graviditet-och-forlossning/sjuka-nyfodda-barn/a-o/ventilator/ [Page 12.]

[26] P. Silva and P. Rocco, "The basics of respiratory mechanics: ventilator-derived parameters," *Annals of Translational Medicine*, vol. 6, pp. 376–376, 10 2018. doi: 10.21037/atm.2018.06.06 [Page 12.]

[27] J. Dekker, F. Stenning, L. Willms, T. Martherus, S. Hooper, and A. te Pas, "Time to achieve desired fraction of inspired oxygen using a t-piece ventilator during resuscitation of preterm infants at birth," *Resuscitation*, vol. 136, pp. 100–104, 2019. doi: https://doi.org/10.1016/j.resuscitation.2019.01.024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0300957218308773 [Page 13.]

[28] N. N, J. B, S. S, T. N, K. FA, and H. R., "Mortality in sepsis and its relationship with gender," *Pak J Med Sci*, vol. 5, pp. 1201–6, 09 2015. doi: 10.12669/pjms.315.6925 [Page 13.]

[29] N. S. Boghossian, M. Geraci, and E. M. E. andJeffrey D Horbar, "Sex differences in mortality and morbidity of infants born at less than 30 weeks," *Pediatrics*, vol. 142, November 2018. doi: 10.1542/peds.2018-2352. [Online]. Available: https://doi.org/10.1542/peds.2018-2352 [Page 13.]

[30] K. Fairchild and J. Aschner, "Hero monitoring to reduce mortality in nicu patients," *Res Rep Neonatol*, vol. 2, pp. 65–76, 08 2012. doi: 10.2147/RRN.S32570 [Page 14.]

[31] K. Jenny R. Fox, Leroy R. Thacker, "Early detection tool of intestinal dysfunction: Impact on necrotizing enterocolitis severity," *American Journal of Perinatology*, no. 10, pp. 927–932, 08 2015. doi: 10.1055/s-0034-1543984 [Page 14.]

[32] V. K. Chauhan, A. Thakur, O. O'Donoghue, O. Rohanian, and D. A. Clifton, "Continuous patient state attention models," Health Informatics, preprint, 12 2022. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2022.12.23.22283908 [Pages 15, 18, 30, and 41.]

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Pages 15, 16, 33, and 41.]

[34] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A Survey of Transformers," 06 2021, arXiv:2106.04554 [cs]. [Online]. Available: http://arxiv.org/abs/2106.04554 [Pages 16, 17, and 18.]

[35] F. Fleure, "Lecture notes from deep learning 14x050, lecture 13: Attention models," 2018, accessed: 2023-10-15. [Online]. Available: https://fleuret.org/dlc/ [Pages xi and 17.]

[36] W. Ge, J.-W. Huh, Y. R. Park, J.-H. Lee, Y.-H. Kim, G. Zhou, and A. Turchin, "Using deep learning with attention mechanism for identification of novel temporal data patterns for prediction of icu mortality," *Informatics in Medicine Unlocked*, vol. 29, p. 100875, 2022. doi: https://doi.org/10.1016/j.imu.2022.100875. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914822000296 [Pages 18 and 41.]

[37] "Introduction to singularity," accessed: 2023-10-09. [Online]. Available: https://docs.sylabs.io/guides/3.5/user-guide/introduction.html [Page 35.]

[38] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. doi: https://doi.org/10.1016/S0031-3203(96)00142-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422 [Page 36.]

[39] M. N. Rabe and C. Staats, "Self-attention does not need $o(n^2)$ memory," 2021. [Page 39.]

[40] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," 2021. [Page 39.]

[41] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. R'e, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *ArXiv*, vol. abs/2205.14135, 2022. [Page 39.]

[42] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," 2023. [Page 39.]

[43] Lucidrains. (2023) memory-efficient-attention-pytorch. [Online]. Available: https://github.com/lucidrains/memory-efficient-attention-pytorch/tree/main [Page 39.]

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org. [Page 41.]

Appendix A

Singularity definition

```
BootStrap: library
From: debian:11

%post
    apt-get -y update
    apt-get install -y --no-install-recommends python pip
git
    python3 -m pip --no-cache-dir install -r /requirements.txt

%files
    requirements.txt /requirements.txt
    /opt/lib /opt/lib
    /opt/psql /opt/psql

%environment
    export LC_ALL=C

%runscript

%labels
    Author Sarah
```

Appendix B

Singularity requirements

```
git+https://gitlab.com/antoinehonore/utils_db
git+https://gitlab.com/antoinehonore/utils_features
git+https://gitlab.com/antoinehonore/utils_results_analysis
pandas==1.5.3
sqlalchemy
psycopg2-binary
matplotlib
torch
parse
openpyxl
scikit-learn
git+https://github.com/barketplace/memory-efficient-attention-pytorch
```

TRITA-EECS-EX- 2023:0000